# Efficient Inference and Learning for Undirected Probabilistic Graphical Models


Xu (Shell) Hu

École des Ponts ParisTech


Joint work with Guillaume Obozinski

# Outline


1. Variational Methods for Undirected Graphical Models
2. Learning of Conditional Random Fields
3. IDAL Algorithm
4. Experiments

# Undirected Graphical Models

Examples: Hidden MRF/CRF (generative/discriminative pair).

### A discriminative model: conditional random field

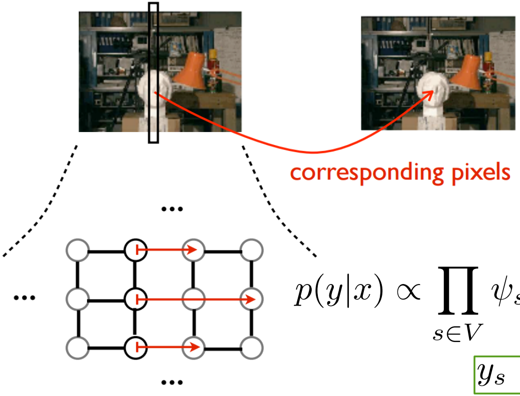
- CRF defines a joint distribution over the random variables $Y := [Y_1, \ldots, Y_S]$ given the observation $X$:

$$p(y \mid x; w) := \frac{1}{Z(x, w)} \prod_{c \in \mathcal{C}} \psi_c(x, y_c).$$

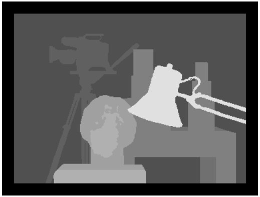
- $\psi_c(x, y_c)$ is a local function (a.k.a. factor) with respect to clique $c$. Usually, $\psi_c(x, y_c) = \langle w_c, \phi_c(x, y_c) \rangle$.

# Applications of CRF

- Computer vision (e.g. depth estimation):




neighbor constrained displacements



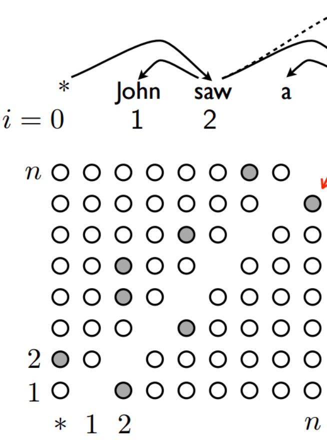
$$p(y|x) \propto \prod_{s\in V} \psi_s(\boxed{y_s},\boxed{x}) \cdot \prod_{st\in E} \psi_{st}(y_s, y_t)$$

$y_s$ : The displacement of pixel s

$x$ : Images of 2 views

# Applications of CRF

- Natural language processing (e.g. dependency parsing):




$y(i,j) = 1$ if arc $i \to j$ is selected and zero otherwise

$$p(y|x) \propto \prod_{ij\in E} \varphi_{ij}(\boxed{y(i,j)}, \boxed{x}) \cdot \varphi_T(y)$$

$x$ : The sentence

*y has to be a tree*

# Undirected Graphical Models

- Factorized form: $p(y) = \frac{1}{Z} \prod_c \psi_c(y_c)$.
- Exponential family form: $p(y|\theta) = \exp(\theta(y) - F(\theta))$
  - Natural parameter: $\theta(y) = \sum_c \theta_c(y_c)$.
  - Log-partition function: $F(\theta) = \log \sum_y \exp(\theta(y)) = \log Z$.

# Inference in Undirected Graphical Models

*Task*: estimate marginal probabilities given $\theta$.

## Example: inference on a chain by dynamic programming

$$p(x_j) = \frac{1}{Z} \sum_{x_{V\setminus\{j,n\}}} \prod_{i=1}^{n-1} \psi_i(x_i) \prod_{i=2}^{n-1} \psi_{i-1,i}(x_{i-1}, x_i) \underbrace{\sum_{x_n} \psi_n(x_n) \psi_{n-1,n}(x_{n-1}, x_n)}_{\mu_{n\to n-1}(x_{n-1})}$$

$$= \frac{1}{Z} \sum_{x_{V\setminus\{j,n,n-1\}}} \prod_{i=1}^{n-2} \psi_i(x_i) \prod_{i=2}^{n-2} \psi_{i-1,i}(x_{i-1}, x_i) \times$$

$$\times \underbrace{\sum_{x_{n-1}} \psi_{n-1}(x_{n-1}) \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \mu_{n\to n-1}(x_{n-1})}_{\mu_{n-1\to n-2}(x_{n-2})}$$

$$= \frac{1}{Z} \sum_{x_{V\setminus\{1,j,n,n-1\}}} \mu_{1\to 2}(x_2) \ \ldots \ \mu_{n-1\to n-2}(x_{n-2})$$

The key quantity: $Z = \sum_{x_i} \mu_{i-1\to i}(x_i) \psi_i(x_i) \mu_{i+1\to i}(x_i)$.

# Variational View of Inference

- The key problem is computing $F$.
- Variational inference $\min_{q \in \mathcal{P}} D_{\mathrm{KL}}(q \| p)$:

$$F(\theta) = \log \sum_y \exp(\theta(y)) \geq \sum_y q(y) \log \frac{\exp \theta(y)}{q(y)}$$
$$= \mathbb{E}_q[\theta(y)] + H_{\mathrm{Shannon}}(y; q)$$

- Fenchel's duality: $F(\theta) = \sup_{q \in \mathcal{P}} \left[\mathbb{E}_q[\theta(y)] + H_{\mathrm{Shannon}}(y; q)\right]$.
- The maximum $q$ is obtained at $q^*(y) = p(y) = \exp(\theta(y) - F(\theta))$, which is also known as the *maximum entropy principle*.

# Variational View of Inference

- Thanks to the Factorization: $\mathbb{E}_q[\theta(y)] = \sum_c \sum_{y_c} q(y_c)\theta_c(y_c)$.
- Equivalent Fenchel conjugate with only marginals:

$$F(\theta) = \sup_{\mu \in \mathcal{M}} \left[\langle \mu, \theta \rangle + H_{\mathrm{Shannon}}(y; \mu)\right]$$

- Marginal polytope:

$$\mathcal{M} = \{\mu \colon \mu_c(y_c) \text{ is a valid marginal probability for some } q \in \mathcal{P}\}$$

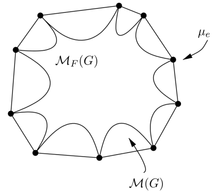
# Variational View of Inference

- Variational view doesn't reduce the complexity of inference.
- Intractable terms $\mathcal{M}$ and $H_{\text{Shannon}}(y;\mu)$:

## Mean-field inference

$$\mathcal{M} \to \{\mu \in \mathcal{M} \colon \mu(y) = \prod_i \mu_i(y_i)\}$$

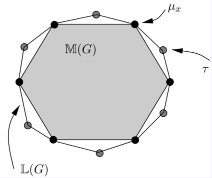
$$H_{\text{Shannon}}(y;\mu) \to \sum_i H(y_i;\mu_i)$$




## Loopy belief propagation

$$\mathcal{M} \to \{\mu \colon \mu_c \in \Delta_c, \sum_{y_j} \mu_{ij}(y_i, y_j) = \mu_i(y_i)\}$$

$$H_{\text{Shannon}}(y;\mu) \to \sum_i H(y_i;\mu_i) - \sum_{ij} I(y_i, y_j;\mu_{ij})$$



# Abstract CRF model

Let $\mathcal{C} = \mathcal{V} \cup \mathcal{E}$, $\quad \log p_w(y^o|x^o) = \sum_{c \in \mathcal{C}} \langle w_{\tau_c}, \phi_c(x^o, y^o_c) \rangle - \log Z(x^o, w)$,

with $y_{\{s,t\}} = y_s y_t^\intercal$ and $\quad Z(x^o, w) = \sum_{y_1} \dots \sum_{y_S} \exp\Big( \sum_{c \in \mathcal{C}} \langle\, w_{\tau_c}, \phi_c(x^o, y_c)\, \rangle \Big)$

In fact
$$
\begin{aligned}
-\log p_w(y^o|x^o) &= \log \sum_y \exp\Big( \sum_{c \in \mathcal{C}} \langle\, w_{\tau_c}, \phi_c(x^o, y_c) - \phi_c(x^o, y^o_c) \rangle \Big) \\
&= \log \sum_y \exp \sum_{c \in \mathcal{C}} \langle \Psi_{(c)}^\intercal w, y_c \rangle \\
&=: F\big(\Psi^\intercal w\big) \quad \text{with} \quad F(\theta) = \log \sum_y \exp \sum_{c \in \mathcal{C}} \langle \theta_{(c)}, y_c \rangle.
\end{aligned}
$$

# Regularized maximum likelihood estimation

The regularized maximum likelihood estimation problem

$$\min_w -\log p_w(y^o|x^o) + \frac{\lambda}{2}\|w\|_2^2$$

is reformulated as

$$\min_w F(\Psi^\intercal w) + \frac{\lambda}{2}\|w\|_2^2 \qquad \text{with} \qquad F(\theta) = \log \sum_y \exp \sum_{c \in \mathcal{C}} \langle \theta_{(c)}, y_c \rangle,$$

$F$ is essentially another way of writing the log-partition function $Z$.

## Big issue: NP-hardness of inference in graphical models

- $F$ and its gradient are NP-difficult to compute.

$\Rightarrow$ the maximum likelihood estimator is intractable.

- $F$ or $\nabla F$ can be estimated using MCMC methods to perform *approximate inference*.
- *Approximate inference* can also be solved as an optimization problem with *variational methods*.

# Compare with the "disconnected graph" case

$$\min_w \sum_{s=1}^{S} \log p_w(y_s^o|x^o) + \frac{\lambda}{2}\|w\|_2^2$$

$$\min_w \sum_{s=1}^{S} F_s(\psi_s^\intercal w) + \frac{\lambda}{2}\|w\|_2^2 \quad \text{with} \quad F_s(w) := \log \sum_{y_s} \exp\langle \theta_{(s)}, y_s\rangle.$$

- $F_s$ is easy to compute: the sum of $K$ terms
- The objective is a sum of a large number of terms

$\Rightarrow$ Very fast randomized algorithms can be used to solve this problem

- SAG Roux et al. (2012)
- SVRG Johnson and Zhang (2013)
- SAGA Defazio et al. (2014), etc
- SDCA Shalev-Shwartz and Zhang (2016)

$$\max_{\alpha_1,\dots,\alpha_S} \sum_{s=1}^{S} F_s^*(\alpha_s) + \frac{1}{2\lambda}\|\sum_{s=1}^{S} \psi_s \alpha_s\|_2^2$$

**Could we do the same for CRFs? With SDCA?**

# Fenchel conjugate of the log-partition function

$$F(\theta) = \max_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle + H_{\text{Shannon}}(\mu),$$

- The marginal polytope $\mathcal{M}$ is the set of all realizable moments vectors

$$\mathcal{M} := \Big\{ \mu = (\mu_c)_{c \in \mathcal{C}} \mid \exists Y \quad \text{s.t.} \quad \forall c \in \mathcal{C},\ \mu_c = \mathbb{E}[Y_c] \Big\}.$$

- $H_{\text{Shannon}}$ is the Shannon entropy of the maximum entropy distribution with moments $\mu$.

$$P^{\#}(w) := F\big(\Psi^\intercal w\big) + \frac{\lambda}{2}\|w\|_2^2$$

$$D^{\#}(\mu) := H_{\text{Shannon}}(\mu) - \iota_{\mathcal{M}}(\mu) - \frac{1}{2\lambda}\|\Psi\mu\|_2^2$$

$$\min_w P^{\#}(w) \qquad \text{and} \qquad \max_\mu D^{\#}(\mu)$$

form a pair of primal and dual optimization problems.

**Both $H_{\text{Shannon}}$ and $\mathcal{M}$ are intractable $\to$ NP-hard problem in general**

# Relaxing the marginal into the local polytope.

## A classical relaxation for $\mathcal{M}$: the local polytope $\mathcal{L}$

For $\mathcal{C} = \mathcal{E} \cup \mathcal{V}$

**Node and edge simplex constraints:**

$$\forall s \in \mathcal{V}, \quad \boxed{\triangle_s := \left\{\mu_s \in \mathbb{R}_+^k \;\middle|\; \mu_s^\intercal 1 = 1\right\}}$$

$$\forall \{s,t\} \in \mathcal{E}, \quad \boxed{\triangle_{\{s,t\}} := \left\{\mu_{st} \in \mathbb{R}_+^{k \times k} \;\middle|\; 1^\intercal \mu_{st}^\intercal 1 = 1\right\}}.$$

$$\boxed{\mathcal{I} := \left\{\mu = (\mu_c)_{c \in \mathcal{C}} \;\middle|\; \forall c \in \mathcal{C}, \quad \mu_c \in \triangle_c\right\}}$$

$$\mathcal{L} := \left\{\mu \in \boxed{\mathcal{I}} \;\middle|\; \boxed{\forall \{s,t\} \in \mathcal{E}, \quad \mu_{st}\,\mathbf{1} = \mu_s, \quad \mu_{st}^\intercal\,\mathbf{1} = \mu_t}\right\}$$

$$\mathcal{L} = \boxed{\mathcal{I}} \cap \boxed{\{\mu \mid A\mu = 0\}}$$

for an appropriate definition of $A$...

# Surrogates for the entropy

Various entropy surrogates exist, e.g.:

- Bethe entropy (nonconvex),
- Tree-reweighted entropy (TRW) (convex on $\mathcal{L}$ but not on $\mathcal{I}$)

## **Separable** surrogates $H_{\text{approx}}$

We consider surrogates of the form $\boxed{H_{\text{approx}}(\mu) = \sum_{c \in \mathcal{C}} h_c(\mu_c)}$, such that

- each function $h_c$ is **smooth**[a] and **convex on** $\triangle_c$ and
- $H_{\text{approx}}$ is **strongly convex on** $\mathcal{L}$

In particular we propose to use

- the Gini entropy: $h_c(\mu_c) = 1 - \|\mu_c\|_F^2$
- a quadratic counterpart of the *oriented tree-reweighted entropy*:

---

[a] i.e. has Lipschitz gradients

# Relaxed dual problem

$$\mathcal{M} \xrightarrow{\text{relax to}} \mathcal{L} = \mathcal{I} \cap \{\mu \mid A\mu = 0\}$$

$$H_{\text{Shannon}} \xrightarrow{\text{relax to}} H_{\text{approx}}(\mu) := \sum_{c \in \mathcal{C}} h_c(\mu_c) .$$

## Problem relaxation

$$D^{\#}(\mu) := H_{\text{Shannon}}(\mu) - \iota_{\mathcal{M}}(\mu) - \frac{1}{2\lambda}\|\Psi\mu\|_2^2$$

relax to $\downarrow$

$$D(\mu) := H_{\text{approx}}(\mu) - \iota_{\mathcal{I}}(\mu) - \iota_{\{A\mu=0\}} - \frac{1}{2\lambda}\|\Psi\mu\|_2^2$$

so that with

$$f_c^*(\mu_c) : h_c(\mu_c) - \iota_{\triangle_c}(\mu_c) \quad \text{and} \quad g^*(\mu) = -\frac{1}{2\lambda}\|\Psi\mu\|_2^2$$

we have $\quad D(\mu) = \sum_{c \in \mathcal{C}} f_c^*(\mu_c) + g^*(\mu) - \iota_{\{A\mu=0\}}.$

# A dual augmented Lagrangian formulation

$$D(\mu) = \sum_{c \in \mathcal{C}} f_c^*(\mu_c) + g^*(\mu) - \iota_{\{A\mu = 0\}}$$

**Idea:** without the linear constraint, we could exploit the form of the objective to use a fast algorithm such as *stochastic dual coordinate ascent*.

$$D_\rho(\mu, \xi) = \sum_{c \in \mathcal{C}} f_c^*(\mu_c) + g^*(\mu) + \langle \xi, A\mu \rangle - \frac{1}{2\rho}\|A\mu\|_2^2$$

By strong duality, we need to solve

$$\min_\xi d(\xi) \qquad \text{with} \qquad d(\xi) := \max_\mu D_\rho(\mu, \xi).$$

# The algorithm

Need to solve

$$\min_{\xi} d(\xi) \qquad \text{with} \qquad d(\xi) := \max_{\mu} D_\rho(\mu, \xi).$$

with

$$D_\rho(\mu, \xi) = \sum_{c \in \mathcal{C}} f_c^*(\mu_c) + g^*(\mu) + \langle \xi, A\mu \rangle - \frac{1}{2\rho} \|A\mu\|_2^2.$$

Note that we have $\nabla d(\xi) = A\mu_\xi$ with $\mu_\xi = \arg\min_{\xi} D_\rho(\mu, \xi)$.

### Combining an *inexact dual Lagrangian method* with a subsolver $\mathcal{A}$

At epoch $t$:

- Maximize $D_\rho$ partially w.r.t. $\mu$ using a fixed number of steps of a (stochastic) linearly convergent algorithm $\mathcal{A}$ to get $\hat{\mu}^t$ from the $\hat{\mu}^{t-1}$.
- Take an inexact gradient step on $d$ with $\xi^{t+1} = \xi^t - \frac{1}{L} A\hat{\mu}^t$

# Main technical lemma

- Let $\xi^t$ (resp. $\hat{\mu}^t$) the value of $\xi$ (resp. $\mu$) at the end of epoch $t$
- Let $\hat{\Delta}_t := \max_\mu D_\rho(\mu, \xi^t) - D_\rho(\hat{\mu}^t, \xi^t)$ and $\Gamma_t := d(\xi^t) - d(\xi^*)$.
- Let $\Delta_t^0 := \max_\mu D_\rho(\mu, \xi^t) - D_\rho(\mu_0^t, \xi^t)$

If algorithm $\mathcal{A}$ used at epoch $t$ to maximize $D_\rho(\mu, \xi)$ w.r.t. $\mu$ is such that

$$\exists \beta \in (0,1), \qquad \mathbb{E}[\hat{\Delta}_t] \le \beta\, \mathbb{E}[\Delta_t^0] \quad ,$$

then $\exists\, \kappa \in (0,1)$ characterizing $d$ and $\exists\, C > 0$ such that, if $\mu_0^t = \hat{\mu}^{t-1}$,

$$\left\| \begin{matrix} \mathbb{E}[\hat{\Delta}_{T_{\text{ex}}}] \\ \mathbb{E}[\Gamma_{T_{\text{ex}}}] \end{matrix} \right\| \le C\, \lambda_{\max}(\beta)^{T_{\text{ex}}} \left\| \begin{matrix} \mathbb{E}[\hat{\Delta}_0] \\ \mathbb{E}[\Gamma_0] \end{matrix} \right\| \quad ,$$

where $\lambda_{\max}(\beta)$ is the largest eigenvalue of the matrix $M(\beta) = \begin{bmatrix} 6\beta & 3\beta \\ 1 & 1-\kappa \end{bmatrix}$

# Main theoretical result: linear convergence in the dual

Let $\mathcal{A}$ be an *iterative* algorithm used to solve partially $\max_{\mu} D_\rho(\mu, \xi)$.

- Let $\xi^t$ (resp. $\hat{\mu}^t$) the value of $\xi$ (resp. $\mu$) at the end of epoch $t$
- Let $\hat{\Delta}_t := \max_{\mu} D_\rho(\mu, \xi^t) - D_\rho(\hat{\mu}^t, \xi^t)$ and $\Gamma_t := d(\xi^t) - d(\xi^*)$.

**Proposition:** If

- $\mathcal{A}$ is a linearly convergent algorithm
- at epoch $t$, $\mathcal{A}$ is initialized with $\hat{\mu}^{t-1}$ ($\rightarrow$ use of warm-starts)
- $\mathcal{A}$ is run for a fixed ahead $T_{\text{in}}$ number of iteration at each epoch

then we have

- $\hat{\Delta}_t$, $\Gamma_t \xrightarrow{a.s.} 0$ linearly
- the residuals $\|A\hat{\mu}^t\|_2^2 \xrightarrow{a.s.} 0$ linearly
- the smooth part of the objective a.s. converges linearly

# Global linear convergence in the primal

Let $P$ be the relaxed primal objective

$$P(w) := F_{\mathcal{L}}\big(\Psi^{\intercal} w\big) + \frac{\lambda}{2}\|w\|_2^2, \quad \text{with} \quad F_{\mathcal{L}}(\theta) := \max_{\mu \in \mathcal{L}} \langle \theta, \mu \rangle + H_{\text{approx}}(\mu).$$

## Corollary

Let $\quad \hat{w}^t = -\frac{1}{\lambda}\Psi\hat{\mu}^t$.

If

- $\mathcal{A}$ is a *linearly convergent* algorithm and
- the function $\mu \mapsto -H_{\text{approx}}(\mu) + \frac{1}{2\rho}\|A\mu\|_2^2$ is *strongly convex*,

then $\qquad P(\hat{w}^t) - P(w^\star)$ converges to $0$ linearly a.s.

Since a fixed nb of inner iterations are done at each epoch, the linear convergence is as a function of the total number of clique updates.

# Related work

A lot of work on approximate inference for CRFs:

- Komodakis et al. (2007); Sontag et al. (2008); Savchynskyy et al. (2011)

Learning method going beyond saddle formulations:

- Meshi et al. (2010); Hazan and Urtasun (2010); Lacoste-Julien et al. (2013)

Learning in the dual for structured SVMs **with only clique-wise updates**:

- With relaxation + smoothing of the linear constraints Meshi et al. (2015) and using block coordinate Frank-Wolfe (BCFW) or block coordinate ascent.
- With multiplier and a greedy primal dual algorithm, Yen et al. (2016) show a global linear convergence result in the dual.

Convergence rates for approximate gradient descent

- Schmidt et al. (2011); Devolder et al. (2014); Lan and Monteiro (2016); Lin et al. (2017)

However,

- the connexion with SDCA was not made,
- there was no linear convergence guarantee in the primal

# Experiments: Algorithms

| | |
|---|---|
| SoftBCFW | Stochastic block coordinate Frank-Wolfe + penalty method (Meshi et al., 2015) |
| SoftSDCA | Stochastic block coordinate prox ascent + penalty method |
| GDMM | Dual decomposed learning with factorwise oracle (Yen et al., 2016) |
| IDAL | Our algorithm |

# Datasets

**Gaussian mixture Potts model**

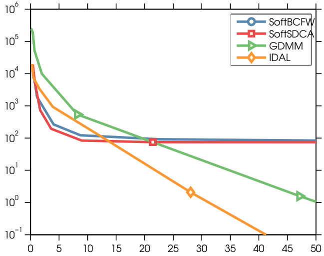
- $10 \times 10$ grid graph with 5 classes
- gaussian features in $\mathbb{R}^{10}$
- $(w_{\tau_1} \in \mathbb{R}^{10 \times 5}, w_{\tau_2} \in \mathbb{R}^{5 \times 5})$
- 50 training grids

**Semantic segmentation of images**

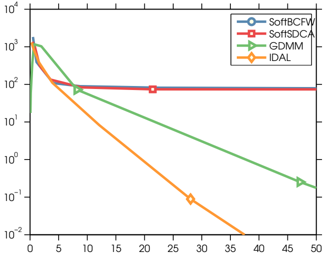
- MSRC-21 dataset (Shotton et al., 2006)
- 21 classes
- 50 features $(w_{\tau_1} \in \mathbb{R}^{50 \times 21}, w_{\tau_2} \in \mathbb{R}^{21 \times 21})$
- 335 training images

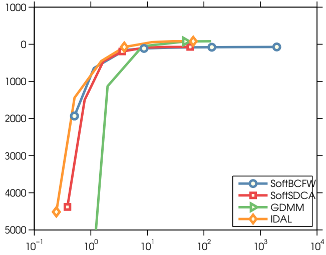
# Results for Gaussian mixture Potts model $(\lambda = 10, \rho = 1)$

Bound on duality gap

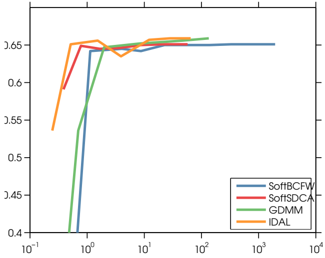



Gap on marginalization constraints



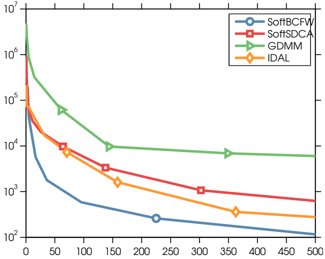

Dual objective



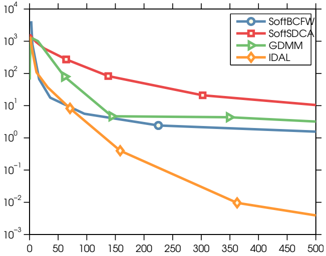

Accuracy on test data

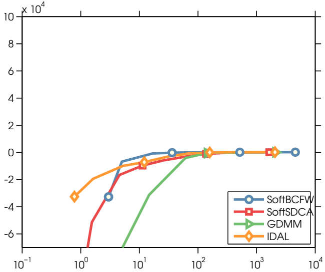
# Result on segmentation dataset, max margin variant $(\lambda = 1, \rho = 0.1)$

Bound on duality gap

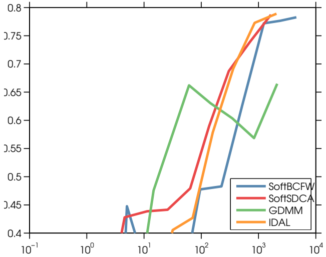



Gap on marginalization constraints




Dual objective




Accuracy on test data

# Future work

- How do we get rid of these relaxations?
- Do we need higher order marginals?
- Is there a better divergence for $D(p_0(y|x) \| p_\theta(y|x))$ than KL?

# References I


Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654.

Devolder, O., Glineur, F., and Nesterov, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75.

Gygli, M., Norouzi, M., and Angelova, A. (2017). Deep value networks learn to evaluate and iteratively refine structured outputs. In *International Conference on Machine Learning*, pages 1341–1351.

Hazan, T. and Urtasun, R. (2010). A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, pages 838–846.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Komodakis, N., Paragios, N., and Tziritas, G. (2007). MRF optimization via dual decomposition: Message-passing revisited. In *ICCV*, pages 1–8.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. (2013). Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, pages 53–61.

Lan, G. and Monteiro, R. D. (2016). Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547.

Lin, H., Mairal, J., and Harchaoui, Z. (2017). QuickeNing: A generic quasi-Newton algorithm for faster gradient-based optimization. *arXiv preprint arXiv:1610.00960*.

Meshi, O., Sontag, D., Globerson, A., and Jaakkola, T. S. (2010). Learning efficiently with approximate inference via dual losses. In *ICML*, pages 783–790.

# References II


Meshi, O., Srebro, N., and Hazan, T. (2015). Efficient training of structured SVMs via soft constraints. In *AISTATS*, pages 699–707.

Roux, N. L., Schmidt, M., and Bach, F. R. (2012). A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671.

Savchynskyy, B., Kappes, J., Schmidt, S., and Schnörr, C. (2011). A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling. In *CVPR*, pages 1817–1823.

Schmidt, M., Le Roux, N., and Bach, F. R. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*, pages 1458–1466.

Shalev-Shwartz, S. and Zhang, T. (2016). Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2):105–145.

Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2006). Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*. Springer.

Sontag, D., Meltzer, T., Globerson, A., Jaakkola, T., and Weiss, Y. (2008). Tightening LP relaxations for MAP using message passing. In *UAI*, pages 503–510.

Yen, I. E.-H., Huang, X., Zhong, K., Zhang, R., Ravikumar, P. K., and Dhillon, I. S. (2016). Dual decomposed learning with factorwise oracle for structural SVM of large output domain. In *NIPS*, pages 5024–5032.