SDCA-Powered Inexact Dual Augmented Lagrangian Method for Fast CRF Learning

Guillaume Obozinski

Swiss Data Science Center

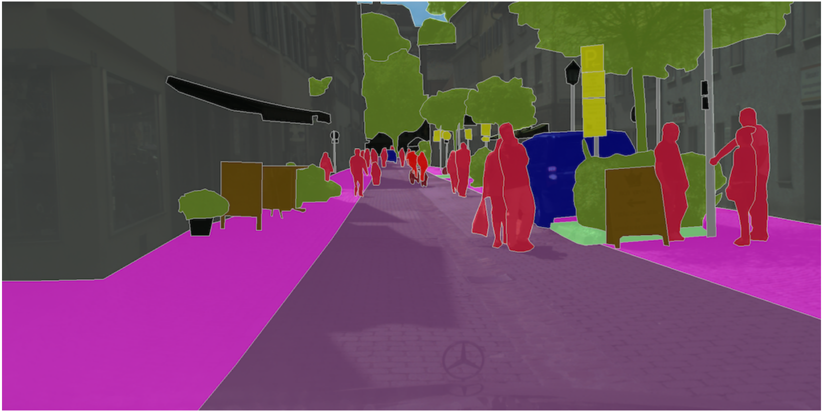
SDSC

Joint work with Shell Xu Hu

Imaging and Machine Learning workshop, IHP, April 2nd 2019

Outline

A motivating example: semantic segmentation



Cityscapes dataset (Cordts et al., 2016)

Recent fast algorithms for large sums of functions

$$\min_w F(w) + \frac{\lambda}{2}\|w\|_2^2 \qquad \text{with} \quad F(w) = \sum_{s=1}^{n} F_s(w)$$

and typically $F_s(w) = f_s(w^\intercal \varphi(x_s)) = \ell(w^\intercal \varphi(x_s), y_s)$

Stochastic gradient methods with variance reduction

Iterate: pick s at random and update $w^{t+1} = w^t - \eta g^t$ with

$$\begin{array}{lcllll}
\text{(SVRG)} & g^t = & \nabla F_s(w^t) - \nabla F_s(\widetilde{w}) & + \frac{1}{n} \nabla F(\widetilde{w}) & \text{and} & \widetilde{w} = \widetilde{w}_{\text{epoch}} \\
\text{(SAG)} & g^t = & \nabla F_s(w^t) - g_s^{t-1} & + g^{t-1} & \text{and} & g_s^t = \nabla F_s(w^t) \\
\text{(SAGA)} & g^t = & \nabla F_s(w^t) - g_s^{t-1} & + \frac{1}{n} g^{t-1} & \text{and} & g_s^t = \nabla F_s(w^t)
\end{array}$$

Stochastic Dual Coordinate Ascent (Implicit Variance reduction)

$$\max_{\alpha_1,\ldots,\alpha_n} \sum_{s=1}^{n} f_s^*(\alpha_s) + \frac{1}{2\lambda} \Big\| \sum_{s=1}^{n} \varphi(x_s)\alpha_s \Big\|_2^2$$

Iterate $\quad \alpha_s^{t+1} \leftarrow \text{Prox}_{\frac{\lambda}{L_s} f_s^*}\left(\alpha_s^t - \frac{1}{L_s}\varphi(x_s)^\intercal w^t\right), \qquad \alpha_i^{t+1} \leftarrow \alpha_i^t, \ \forall i \neq s.$

Variance reduction techniques yield improved rates

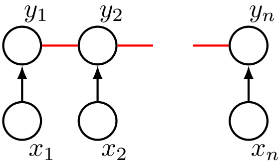
- κ : condition number
- d : ambient dimension

Running times to have $\mathrm{Obj}(w) - \mathrm{Obj}(w^*) \leq \varepsilon$

Stochastic GD	$d\,\kappa\,\frac{1}{\varepsilon}$
GD	$d\,n\kappa\,\log\frac{1}{\varepsilon}$
Accelerated GD	$d\,n\sqrt{\kappa}\,\log\frac{1}{\varepsilon}$
SAG(A), SVRG, SDCA, MISO	$d\,(n+\kappa)\,\log\frac{1}{\varepsilon}$
Accelerated variants	$d\,(n+\sqrt{n\kappa})\,\log\frac{1}{\varepsilon}$

- Exploiting sum structure yields faster algorithms...

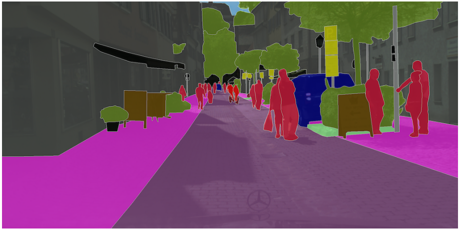
$$\min_w \sum_{s=1}^{n} \ell(w^\intercal \phi(x_s), y_s) + \frac{\lambda}{2}\|w\|_2^2$$



✓

Conditional Random Fields

- Input image x
- Features at pixel s: $\varphi_s(x)$
- Encoding of class at pixel s: $y_s = (y_{s1}, \dots, y_{sK})$ with
  - $y_{sk} = 1$ if in class k
  - $y_{sk} = 0$ else.



Options:

1. predict each pixel class individually: ***multiclass logistic regression***

$$p(y_s|x) \propto \exp\left(\sum_{k=1}^{K} y_{sk}\, w_k{}^\mathsf{T}\, \varphi_s(x)\right)$$

2. View image as a grid graph with vertices $\mathcal{V}$ and edges $\mathcal{E}$, and predict all pixels classes jointly while accounting for dependencies: ***CRF***

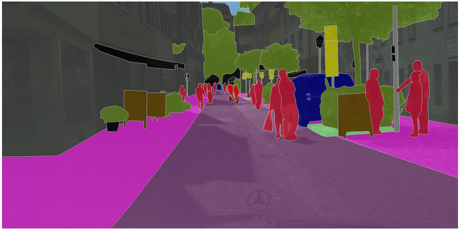
$$p(y_1, \dots, y_S|x) \propto \exp\left(\sum_{s\in\mathcal{V}} \sum_{k=1}^{K} y_{sk} w_{\tau_1,k}{}^\mathsf{T} \varphi_s(x) + \sum_{\{s,t\}\in\mathcal{E}} \sum_{k,l=1}^{K} w_{\tau_2,kl}\, y_{sk}\, y_{tl}\right)$$

Trick: log-likelihood as log-partition

$$
\begin{aligned}
-\log p(y^o|x^o) &= -\langle w, \phi(y^o, x^o)\rangle + A_{x^o}(w) \\
&= -\langle w, \phi(y^o, x^o)\rangle + \log \sum_{y} \exp\langle w, \phi(y, x^o)\rangle \\
&= \log \sum_{y} \exp\langle w, \phi(y, x^o) - \phi(y^o, x^o)\rangle \\
&= \log \sum_{y} \exp\Big(\sum_{c\in\mathcal{C}} \langle w_{\tau_c}, \phi_c(y_c, x^o) - \phi(y^o_c, x^o)\rangle\Big) \\
&= \log \sum_{y} \exp\Big(\sum_{c\in\mathcal{C}} \langle y_c, \theta_{(c)}\rangle\Big) \\
&\text{with} \quad \theta_{(c)} = \big(\langle w_{\tau_c}, \underbrace{\phi_c(y'_c, x^o) - \phi(y^o_c, x^o)}_{\psi_c(y'_c)}\rangle\big)_{y'_c\in\mathcal{Y}_c}.
\end{aligned}
$$

Conditional Random Fields

- Input image x
- Features at pixel s: $\varphi_s(x)$
- Encoding of class at pixel s: $y_s = (y_{s1}, \ldots, y_{sK})$ with
  - $y_{sk} = 1$ if in class k
  - $y_{sk} = 0$ else.



Options:

1. predict each pixel class individually: ***multiclass logistic regression***

$$p(y_s|x) \propto \exp\left(\sum_{k=1}^{K} y_{sk}\, w_k{}^{\mathsf{T}} \varphi_s(x)\right)$$

2. View image as a grid graph with vertices $\mathcal{V}$ and edges $\mathcal{E}$, and predict all pixels classes jointly while accounting for dependencies: ***CRF***

$$p(y_1, \ldots, y_S|x) \propto \exp\left(\sum_{s\in\mathcal{V}} \sum_{k=1}^{K} y_{sk} w_{\tau_1,k}{}^{\mathsf{T}} \varphi_s(x) + \sum_{\{s,t\}\in\mathcal{E}} \sum_{k,l=1}^{K} w_{\tau_2,kl}\, y_{sk}\, y_{tl}\right)$$

Abstract CRF model

$$p(y^o|x^o) \propto \exp\left(\sum_{s\in\mathcal{V}}\sum_{k=1}^{K} y^o_{sk}\, w_{\tau_1,k}{}^{\intercal}\, \varphi_s(x^o) + \sum_{\{s,t\}\in\mathcal{E}}\sum_{k,l=1}^{K} w_{\tau_2,kl}\, y^o_{sk}\, y^o_{tl}\right)$$

$$p(y^o|x^o) \propto \exp\left(\sum_{s\in\mathcal{V}}\langle w_{\tau_1}, \phi_s(y^o_s, x^o)\rangle + \sum_{\{s,t\}\in\mathcal{E}}\langle w_{\tau_2}, \phi_{st}(y^o_s, y^o_t, x^o)\rangle\right)$$

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 $Z(x^o, w) = \sum_{y_1}\dots\sum_{y_S}\exp\left(\sum_{c\in\mathcal{C}}\langle w_{\tau_c}, \phi_c(x^o, y_c)\rangle\right)$

In fact $-\log p_w(y^o|x^o) = \log\sum_{y}\exp\left(\sum_{c\in\mathcal{C}}\langle w_{\tau_c}, \phi_c(x^o, y_c) - \phi_c(x^o, y^o_c)\rangle\right)$

$$= \log\sum_{y}\exp\sum_{c\in\mathcal{C}}\langle \Psi_{(c)}^{\intercal} w, y_c\rangle$$

$$=: f(\Psi^{\intercal} w) \quad \text{with} \quad f(\theta) = \log\sum_{y}\exp\sum_{c\in\mathcal{C}}\langle\theta_{(c)}, y_c\rangle.$$

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 A.

Major 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 ∇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

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

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

- 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)

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

Could we do the same for CRFs? With SDCA?

Fenchel conjugate of the log-partition function

$$f(\theta) := \log \sum_{y} \exp \sum_{c \in \mathcal{C}} \langle \theta_{(c)}, y_c \rangle = \max_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle + H_{\text{Shannon}}(\mu),$$

- 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_{Shannon} is the Shannon entropy of the maximum entropy distribution with moments μ.

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

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

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

form a pair of primal and dual optimization problems.

Both H_{Shannon} and $\mathcal{M}$ are intractable $\rightarrow$ 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:

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

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

$$\boxed{\mathcal{I} := \Big\{\mu = (\mu_c)_{c \in \mathcal{C}} \mid \forall c \in \mathcal{C}, \quad \mu_c \in \triangle_c\Big\}}$$

$$\mathcal{L} := \Big\{\mu \in \boxed{\mathcal{I}} \mid \boxed{\forall \{s,t\} \in \mathcal{E}, \quad \mu_{st}\,\mathbf{1} = \mu_s, \quad \mu_{st}^\intercal\,\mathbf{1} = \mu_t}\Big\}$$

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

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_{approx}

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

- each function h_c is **smooth**[a] and **convex on** $\triangle_c$ and
- H_{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 $$D(\mu) = -\sum_{c \in \mathcal{C}} f_c^*(\mu_c) - g^*(\mu) - \iota_{\{A\mu=0\}}.$$

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_ρ partially w.r.t. μ 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 ξ^t (resp. $\hat{\mu}^t$) the value of ξ (resp. μ) 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. μ is such that

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

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

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

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 ξ^t (resp. $\hat{\mu}^t$) the value of ξ (resp. μ) 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_{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 $\quad P(\hat{w}^t) - P(w^\star)$ 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)

Related work on BCFW with linear constraints: Gidel et al. (2018)

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

- 10×10 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

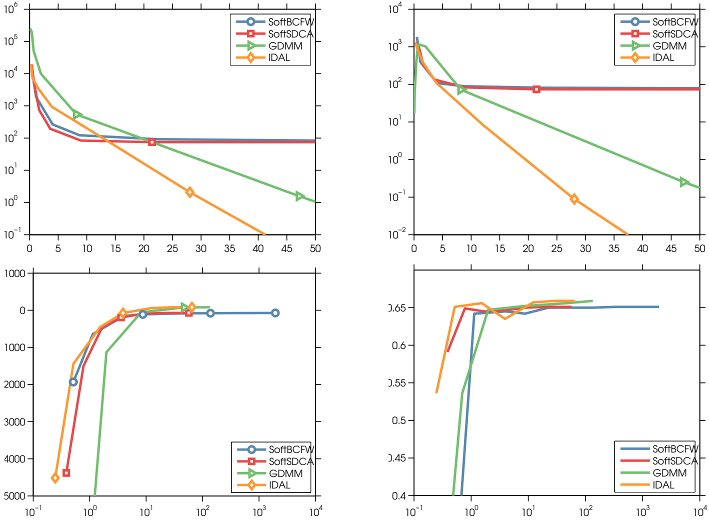
Semantic segmentation of images

- 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

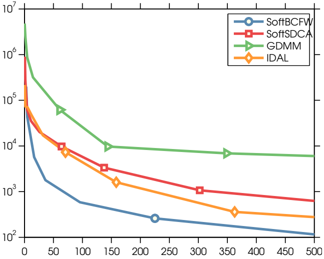


Dual objective

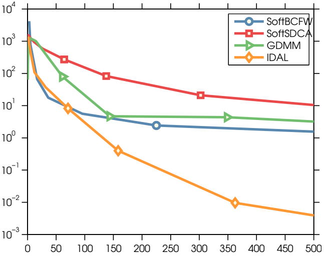
Accuracy on test data

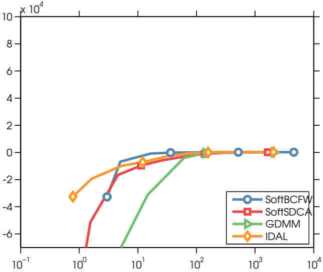
Result on segmentation dataset, max margin variant $(\lambda = 1, \rho = 0.1)$

Bound on duality gap

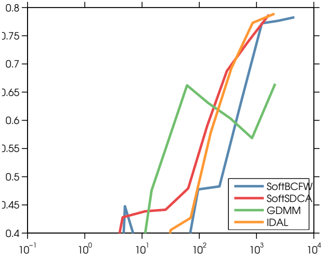


Gap on marginalization constraints





Dual objective



Accuracy on test data

Summary and conclusions

We proposed an algorithm combining SDCA and an inexact dual Lagrangian method that obtains

- Global linear convergences for the relaxed objective
  - In the primal and for the dual augmented Lagrangian formulation
  - Obtains good practical performance

Other contributions:

- Computable duality gaps to track convergence in the primal
- Representer theorem in the structured learning case "inside the graph"
- Unified derivation connecting formulations of previous work
- SDCA can accommodate linear constraints on the dual parameter.

Open questions:

- Use a better approx. for the entropy like OTRW ($\rightarrow$ non Lipschitz gradients)?
- Be stochastic on ξ as well?

Paper: SDCA-Powered Inexact Dual Augmented Lagrangian Method for Fast CRF Learning, X. Hu, G. Obozinski, AIStats, 2018.

References I

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Gidel, G., Pedregosa, F., and Lacoste-Julien, S. (2018). Frank-wolfe splitting via augmented lagrangian method. *AIStats*.

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*.

References II

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

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.