Structured output models for image segmentation

Aurelien Lucchi

Machine Learning Workshop (MLWS) IDIAP - EPFL
Monday November 19th, 2012

Collaborators: Yunpeng Li, Kevin Smith, Raphael Sznitman, Bohumil Maco, Graham Knott, Pascal Fua.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Outline

1. Review Conditional Random Fields (CRF)
2. Maximum likelihood training for CRFs
3. Maximum Margin Training for CRFs
    1. Cutting plane (Structured SVM)
    2. Online subgradient descent

1. Review CRF

Structured prediction

- Non structured output

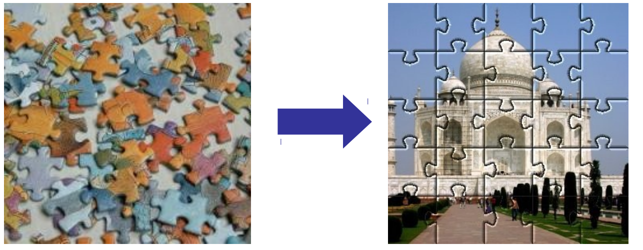
  $$f : X \to R$$

  - inputs X can be any kind of objects
  - output y is a real number

    $$y = \{-1, +1\}$$

- Prediction of complex outputs

  $$f : X \to Y$$



  - Structured output y is complex (images, text, audio...)
  - Ad hoc definition of structured data: data that consists of several parts, and not only the parts themselves contain information, but also the way in which the parts belong together

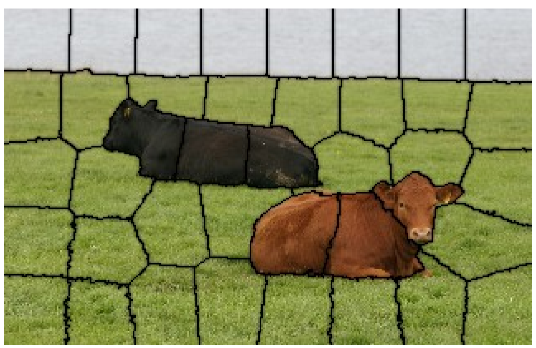
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Slide courtesy: Christoph Lampert

Structured prediction for image segmentation

$$X$$

$$(x_1, \cdots, x_i, \cdots, x_n)$$

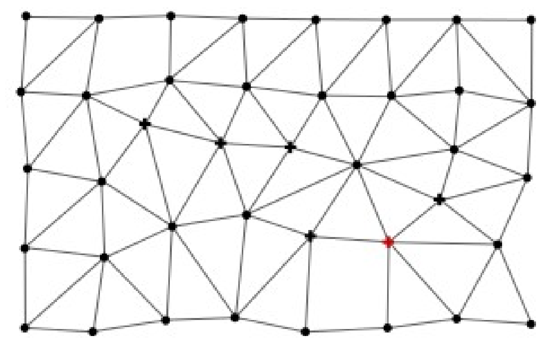


$$x_i \in \mathbb{R}^F$$

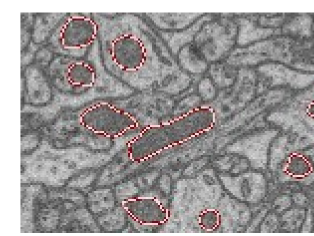
Histograms, Filter responses, ...

$$Y$$

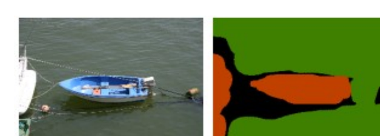
$$(y_1, \cdots, y_i, \cdots, y_n)$$



$$y_i = \{-1, +1\}$$



$$y_i = \{1, ...21\}$$



Original Ground truth

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

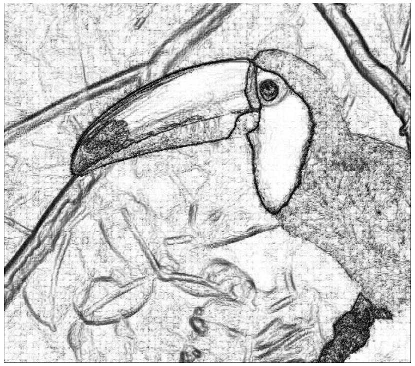
CRF for image segmentation

$$E_{\mathbf{w}}(X,Y) = \sum_{i \in \mathcal{V}} D(y_i; x_i) + \sum_{(i,j) \in \mathcal{E}} V(y_i, y_j)$$

Maximum-a-posteriori (MAP) solution : $$Y^* = \arg\min_{Y \in \mathcal{Y}} E_w(X,Y)$$

Data (**D**)

Unary likelihood



Pair-wise Terms

MAP Solution

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Boykov and Jolly [ICCV 2001], Blake et al. [ECCV 2004]
Slide courtesy : Pushmeet Kohli

CRF for image segmentation

$$E_{\mathbf{w}}(X,Y) = \sum_{i \in \mathcal{V}} D(y_i; x_i) + \sum_{(i,j) \in \mathcal{E}} V(y_i, y_j)$$

Maximum-a-posteriori (MAP) solution : $Y^* = \arg\min_{Y \in \mathcal{Y}} E_w(X,Y)$

Data (**D**)

Unary likelihood

Pair-wise Terms

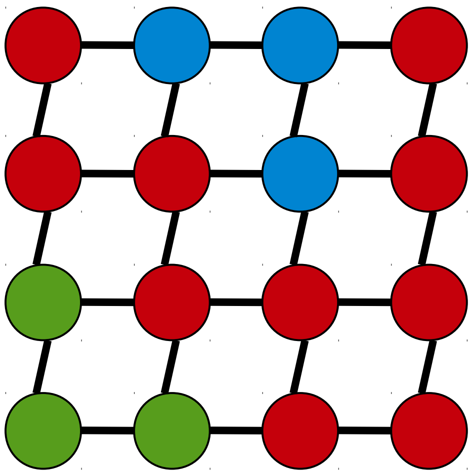
MAP Solution

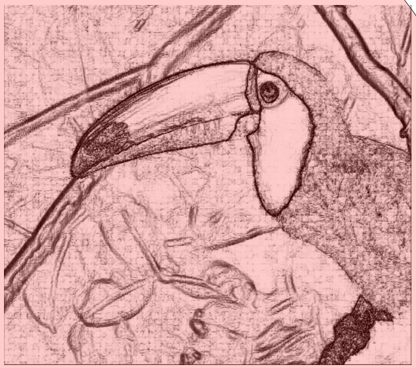
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Boykov and Jolly [ICCV 2001], Blake et al. [ECCV 2004]
Slide courtesy : Pushmeet Kohli

CRF for image segmentation

$$E_{\mathbf{w}}(X,Y) = \sum_{i \in \mathcal{V}} D(y_i; x_i) + \sum_{(i,j) \in \mathcal{E}} V(y_i, y_j)$$





Pair-wise Terms

Favors the same label for neighboring nodes.

CRF for image segmentation

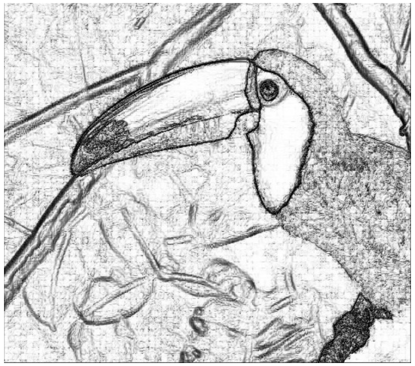
$$E_{\mathbf{w}}(X,Y) = \sum_{i \in \mathcal{V}} D(y_i; x_i) + \sum_{(i,j) \in \mathcal{E}} V(y_i, y_j)$$

Maximum-a-posteriori (MAP) solution :

$$Y^* = \arg\min_{Y \in \mathcal{Y}} E_w(X,Y)$$

Data (**D**)

Unary likelihood



Pair-wise Terms

MAP Solution

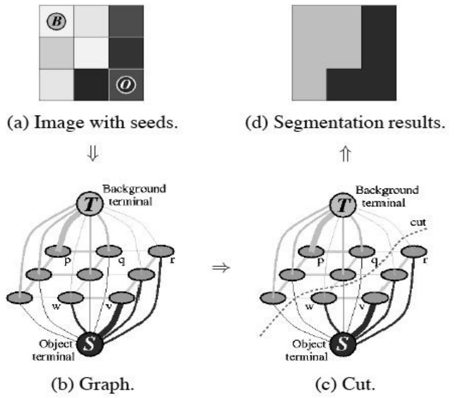
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Boykov and Jolly [ICCV 2001], Blake et al. [ECCV 2004]
Slide courtesy : Pushmeet Kohli

Energy minimization

- MAP inference for discrete graphical models:

$$Y^* = \arg\min_{Y \in \mathcal{Y}} E_w(X, Y)$$



  - Dynamic programming
    - Exact on non loopy graphs
  - Graph-cuts (Boykov, 2001)
    - Optimal solution if energy function is submodular
  - Belief propagation (Pearl, 1982)
    - No theoretical guarantees on loopy graphs but seems to work well in practice.
  - Mean field (root in statistical physics)
  - ...

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Training a structured model ?

- First rewrite the energy function as:

$$E_w(X,Y) = \sum_{i \in \mathcal{V}} D(y_i) + \sum_{i,j \in \mathcal{E}} V(y_i, y_j)$$

$$= w^T \psi(X,Y)$$

Log-linear model

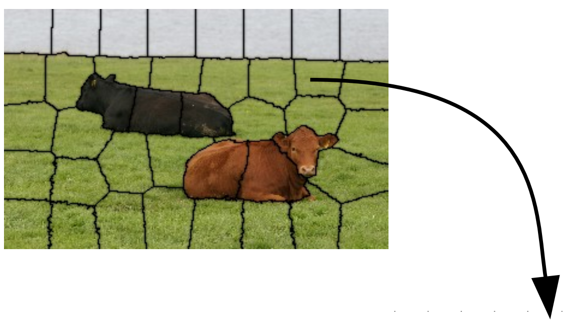
- Efficient Learning/Training – need to efficiently learn parameters **w** from training data ?

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Training a structured model ?

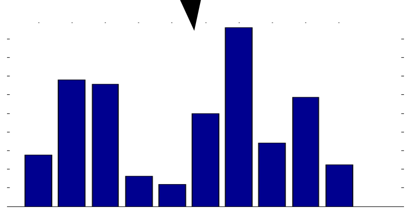
- Energy function is parametrized by vector **w**

$$E_{\mathbf{w}}(X,Y) = \sum_{i \in \mathcal{V}} D(y_i; x_i) + \sum_{i,j \in \mathcal{E}} V(y_i, y_j) = \mathbf{w}^T \psi(X,Y)$$



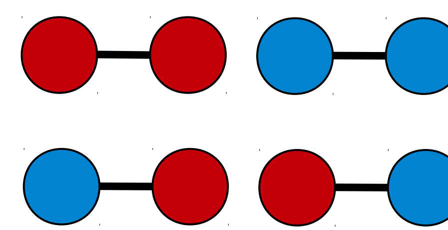
$$\mathbf{w} = ((\mathbf{w}^D)^T, (\mathbf{w}^V)^T)^T$$

$(\mathbf{w}^D)^T$



+

$\mathbf{w}^P$	-1	1
-1	?	?
1	?	?



$$D(y_i) = (\mathbf{w}^D)^T x_i$$

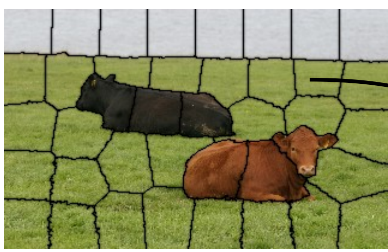
$$V(y_i, y_j) = \mathbf{w}^V(y_i, y_j)$$

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Training a structured model ?

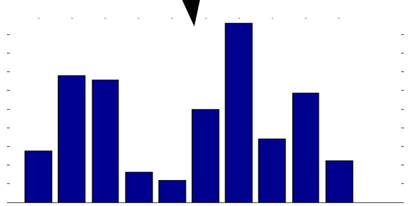
- Energy function is parametrized by vector **w**

$$E_{\mathbf{w}}(X,Y) = \sum_{i \in \mathcal{V}} D(y_i; x_i) + \sum_{i,j \in \mathcal{E}} V(y_i, y_j) = \mathbf{w}^T \psi(X,Y)$$



$$\mathbf{w} = ((\mathbf{w}^D)^T, (\mathbf{w}^V)^T)^T$$

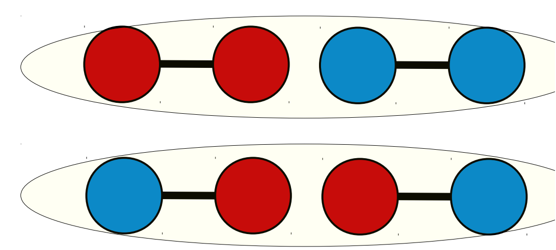
$$(\mathbf{w}^D)^T$$



+

$\mathbf{w}^P$	-1	1
-1	0	1
1	1	0

Low energy



High energy

$$D(y_i) = (\mathbf{w}^D)^T x_i$$

$$V(y_i, y_j) = \mathbf{w}^V(y_i, y_j)$$

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

2. Maximum likelihood training

Maximum likelihood

$$
\begin{aligned}
w^* &= \arg\max_w L(w) = \arg\max_w \log p(Y|X,w) \\
&= \arg\max_w \prod_n \log p(Y^n|X^n,w) \\
&= \arg\max_w \sum_n \log p(Y^n|X^n,w) \\
&= \arg\max_w \sum_n \log \frac{1}{Z(w)} \exp^{-E(Y^n;X^n)} \\
&= \arg\max_w \sum_n -E(Y^n;X^n) - \log Z(w)
\end{aligned}
$$

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Note: We assumed that p is a Gibbs distribution

Maximum likelihood

$$L(w) = \sum_i w^T \psi(x_i, y_i) - \log \sum_y exp^{w^T \psi(x_i, y)}$$

- L(w) is differentiable and convex (it has a positive definite Hessian) so gradient descent can find the global optimum.

Maximum likelihood

$$\nabla_w L(w) = \sum_i [\psi(x_i, y_i) - \frac{\sum_y exp^{w^T \psi(x_i,y)} \psi(x_i, y)}{\sum_y exp^{w^T \psi(x_i,y)}}]$$

$$= \sum_i [\psi(x_i, y_i) - \sum_y p(x_i, y|w) \psi(x_i, y)]$$

- For general CRFs, there is still a problem with the computation of the derivative because the number of possible configurations for y is typically (exponentially) large.

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

$$y_i = \{-1, +1\} \rightarrow |\sum_y| = 2^n$$

Training a structured model ?

- Other solutions exist:
  - Pseudo-likelihood
  - Variational approximation
  - Contrastive divergence
  - **Maximum-margin framework (e.g. Structured SVM)**

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

3.1. Maximum Margin Training of Structured Models: cutting plane (structured SVM)

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Structured SVM

$$E_{\mathbf{w}}(X,Y) = \sum_{i \in \mathcal{V}} D(y_i; x_i) + \sum_{(i,j) \in \mathcal{E}} V(y_i, y_j)$$

- Given a set of N training examples with ground truth labels $(Y^{(1)}, \ldots, Y^{(N)})$, we can write

$$Y^* = \arg\min_{Y \in \mathcal{Y}} E_w(X,Y)$$

$\equiv$

$$\forall n, Y \in \mathcal{Y}_n \backslash Y^{(n)}$$

$$E_{\mathbf{w}}(Y^{(n)}) \leq E_{\mathbf{w}}(Y)$$

Energy for the correct labeling at least as low as energy of any incorrect labeling..

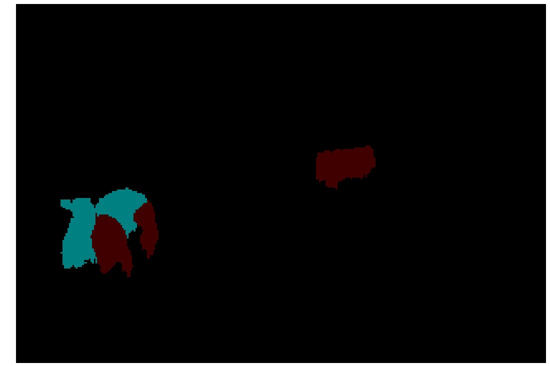
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Structured SVM

E()

ground-truth

<

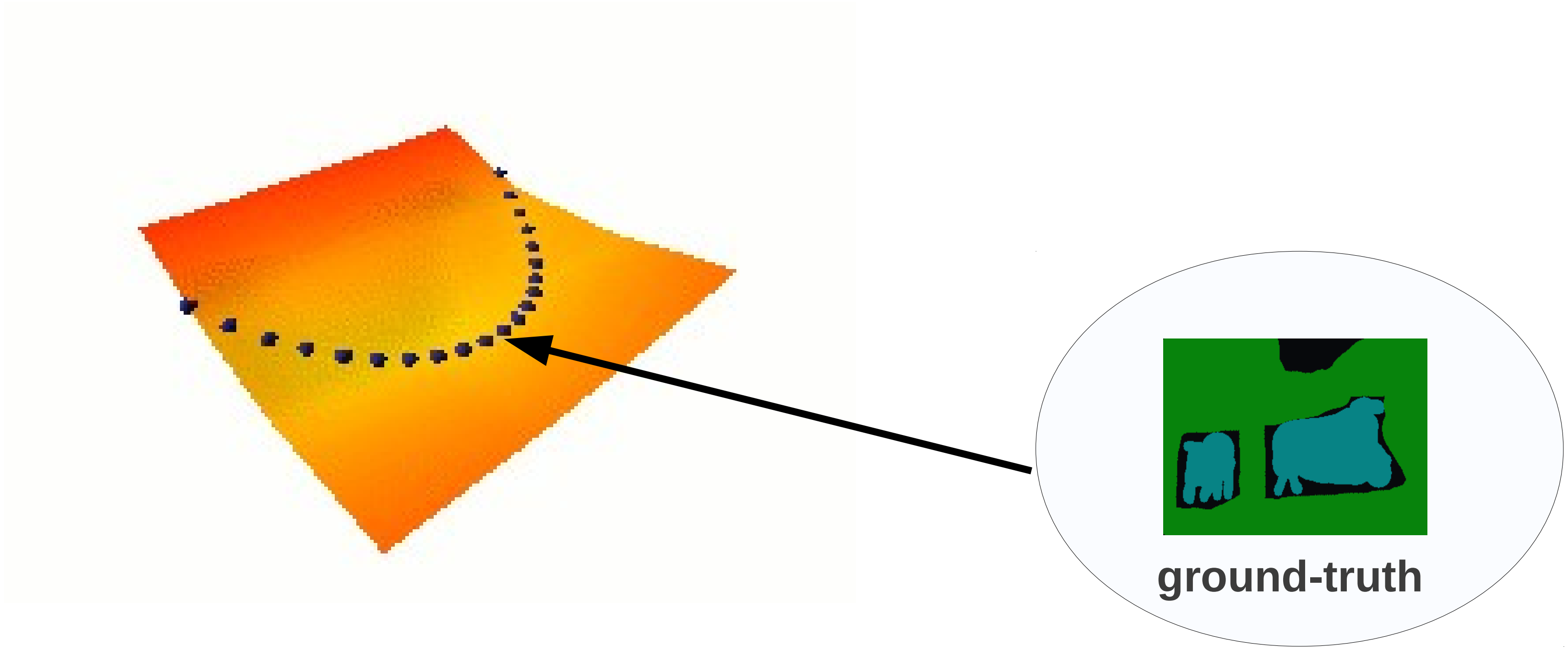
E()

E()

E()

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Energy based landscape



See http://www.cs.nyu.edu/~yann/research/ebm/

Structured SVM

- Given a set of N training examples with ground truth labellings $(Y^{(1)}, \ldots, Y^{(N)})$ we optimize :

$$\min_{\mathbf{w},\xi \geq \mathbf{0}} \frac{1}{2}||\mathbf{w}||_2^2 + \frac{C}{N}\sum_{n=1}^{N} \xi_n$$

$$\text{s.t. } \forall n, Y \in \mathcal{Y}_n \backslash Y^{(n)} : \delta E_{\mathbf{w}}(Y) \geq \Delta(Y^{(n)}, Y) - \xi_n$$

$$\delta E_{\mathbf{w}}(Y) = E_{\mathbf{w}}(Y) - E_{\mathbf{w}}(Y^{(n)}) \qquad \Delta(Y^{(n)}, Y) = \sum_{i \in \mathcal{V}} I(y_i \neq y_i^{(n)})$$

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Structured SVM

- Since the SSVM operates by solving a quadratic program (QP), all the constraints must be linear.
  - Energy function must be expressible as an inner product between the parameter vector w and a feature map.

$$E_{\mathbf{w}}(X,Y) = \sum_{i \in \mathcal{V}} D(y_i; x_i) + \sum_{(i,j) \in \mathcal{E}} V(y_i, y_j)$$

$$E_{\mathbf{w}}(Y) = \langle \mathbf{w}, \Psi(Y) \rangle \,.$$

$$\mathbf{w} = ((\mathbf{w}^D)^T, (\mathbf{w}^V)^T)^T$$

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Structured SVM

$$\min_{\mathbf{w},\xi\geq\mathbf{0}} \frac{1}{2}||\mathbf{w}||_2^2 + \frac{C}{N}\sum_{n=1}^{N}\xi_n$$

$$\text{s.t. } \forall n, Y \in \mathcal{Y}_n \backslash Y^{(n)} : \delta E_{\mathbf{w}}(Y) \geq \Delta(Y^{(n)}, Y) - \xi_n$$

Exponential number of constraints: $N \times 2^{|\mathcal{V}|}$

Electron microscopy (EM) dataset
See http://cvlab.epfl.ch/research/medical/em/mitochondria/

$$|\mathcal{V}| = 133555 \text{ nodes}$$

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Structured SVM

- In order to deal with the exponential number of constraints in the QP, Tsochantaridis proposed a **cutting plane algorithm**.
  - Iteratively finds the most violated constraint and adds it to the working set of constraints.

    $$\hat{Y} = argmin_{Y \in \mathcal{Y}_n} E_{\mathbf{w}}(Y) - \Delta(Y^{(n)}, Y)$$

  - See Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: ICML. (2004)

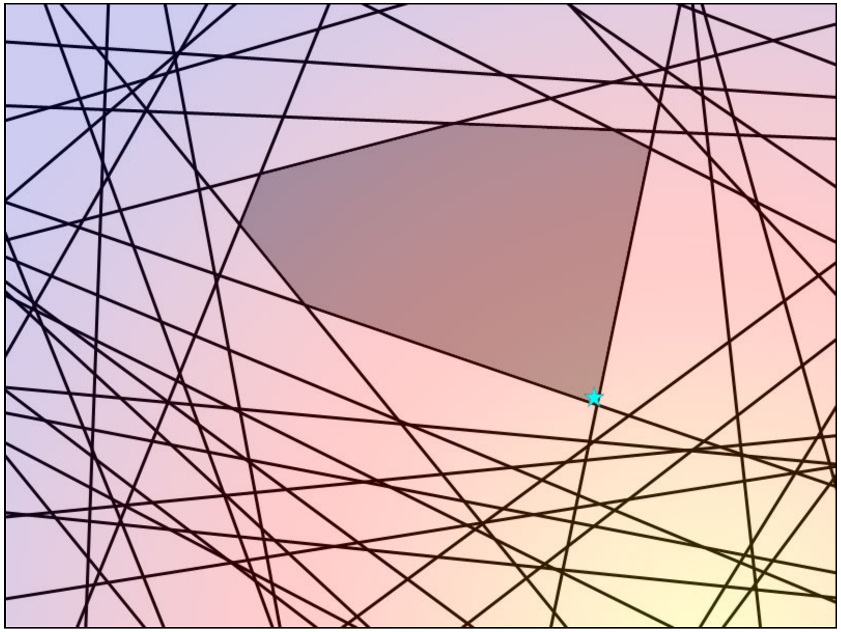
Structured SVM

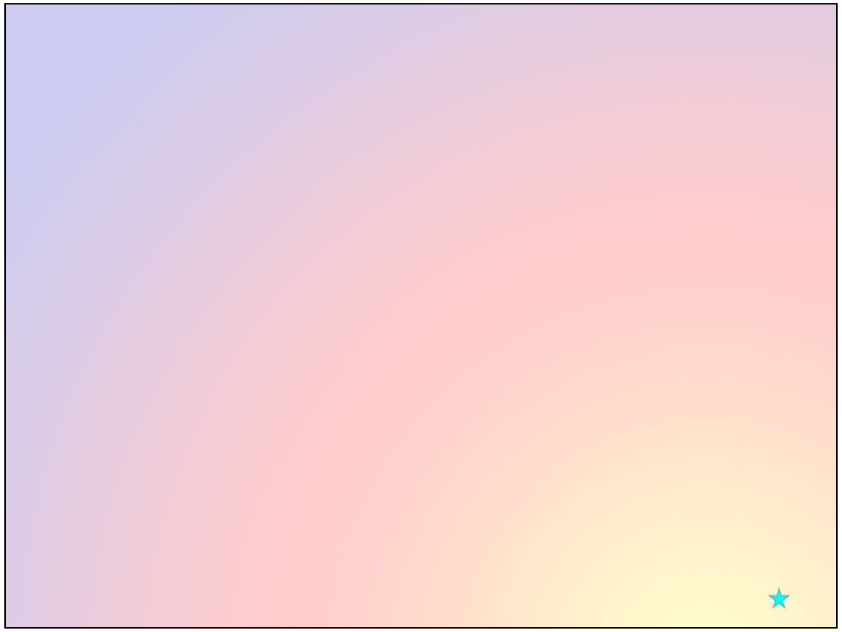
- In order to deal with the exponential number of constraints in the QP, Tsochantaridis proposed a **cutting plane algorithm**.
  - Iteratively finds the most violated constraint and adds it to the working set of constraints.

$$\hat{Y} = argmin_{Y \in \mathcal{Y}_n} E_{\mathbf{w}}(Y) - \Delta(Y^{(n)}, Y)$$

  - See Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: ICML. (2004)

Illustrative Example





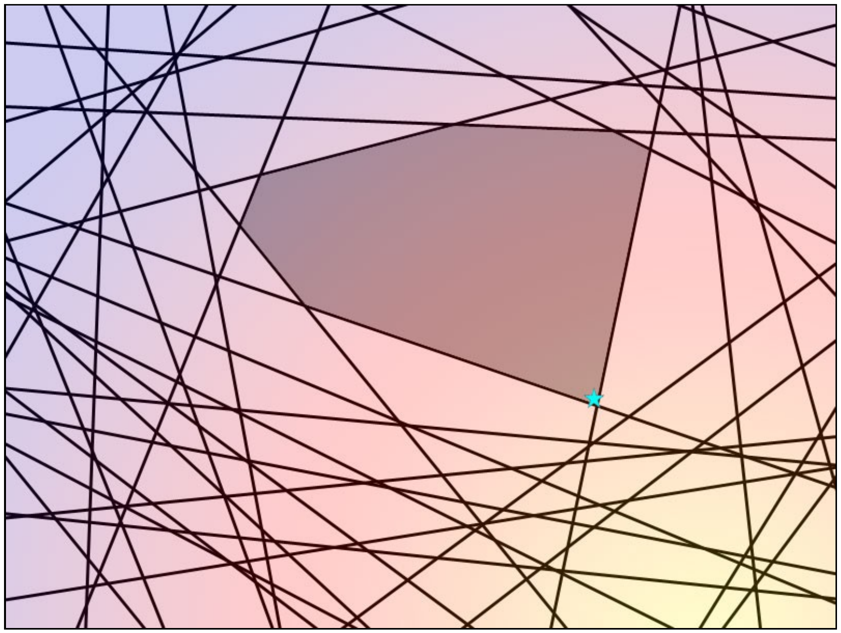
SSVM Problem

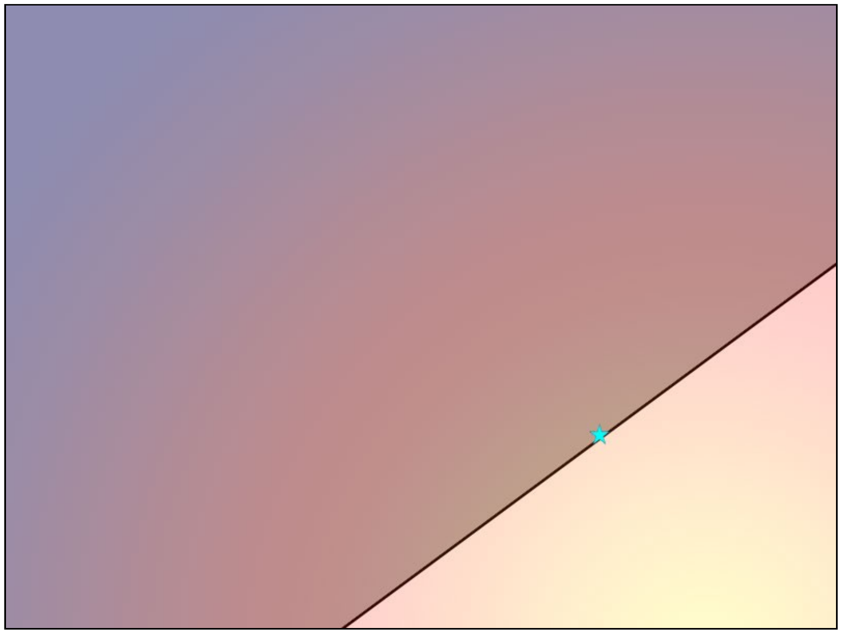
- Exponential constraints
- Most are dominated by a small set of "important" constraints

Cutting plane approach

- Repeatedly finds the next most violated constraint…
- …until set of constraints is a good approximation.

Illustrative Example





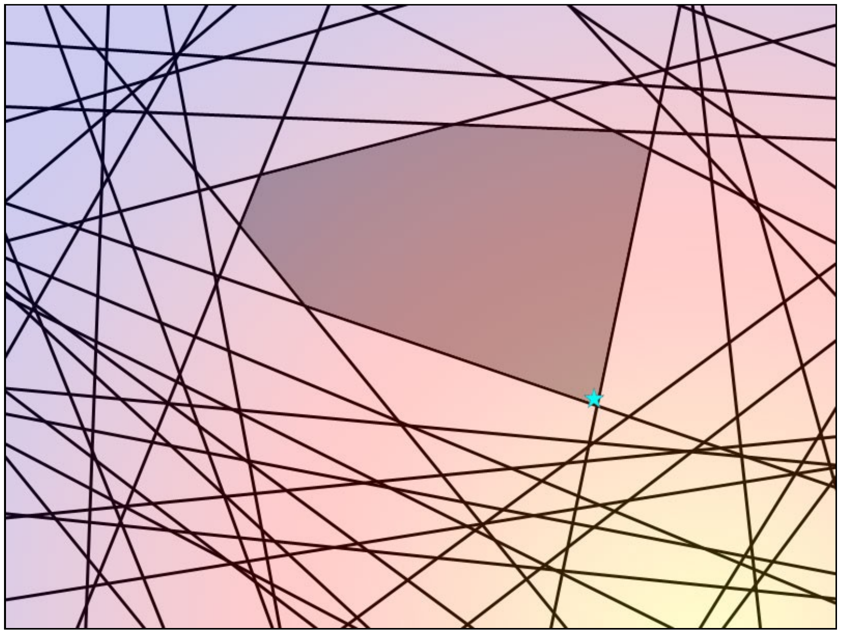
SSVM Problem

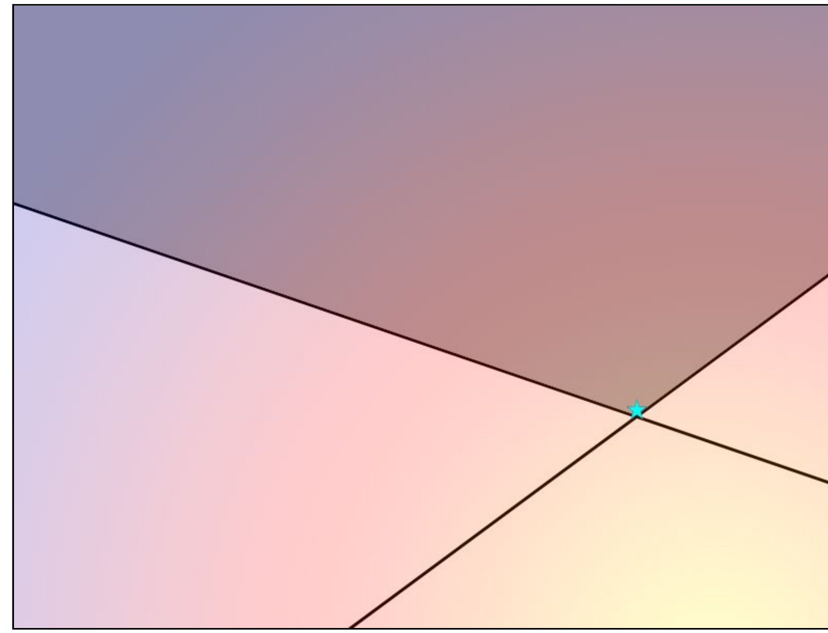
- Exponential constraints
- Most are dominated by a small set of "important" constraints

Cutting plane approach

- Repeatedly finds the next most violated constraint…
- …until set of constraints is a good approximation.

Illustrative Example





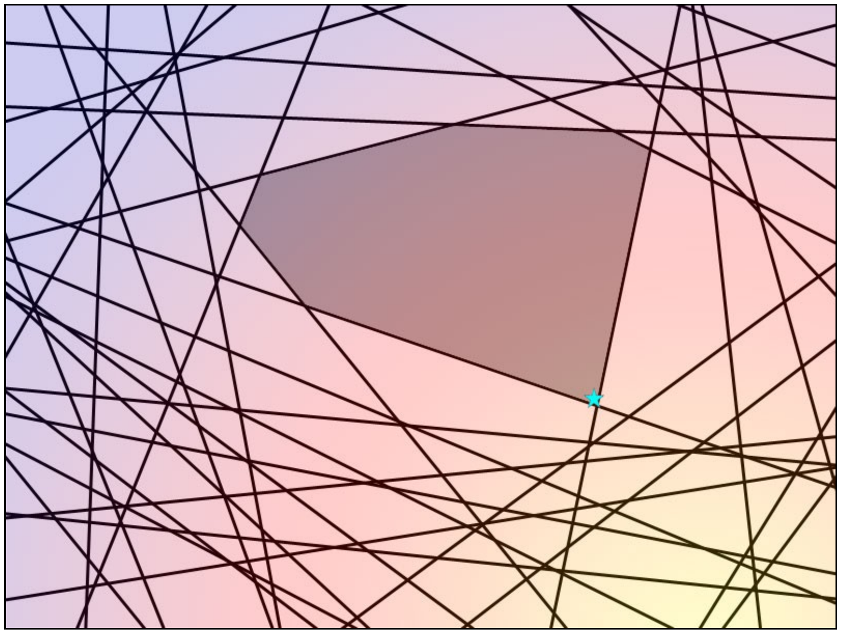
SSVM Problem

- Exponential constraints
- Most are dominated by a small set of "important" constraints

Cutting plane approach

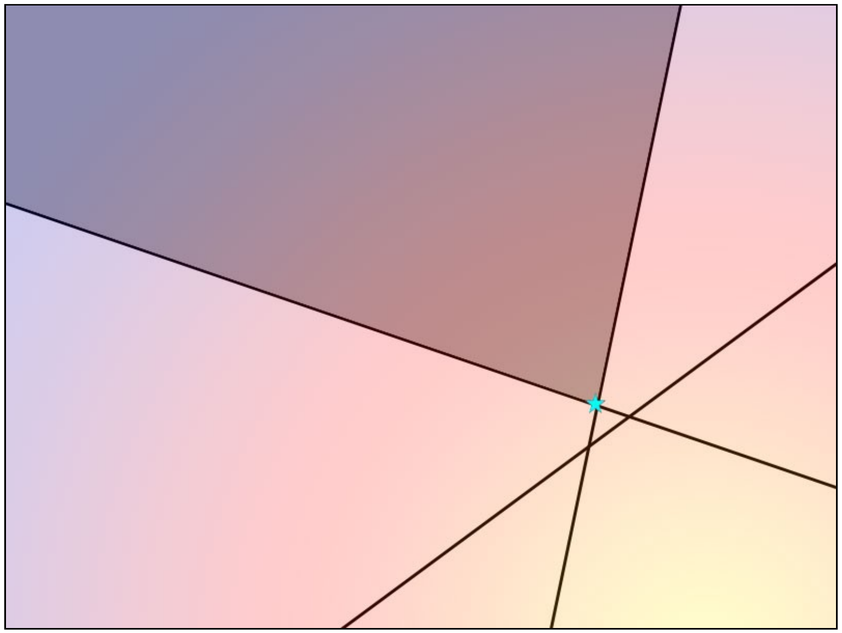
- Repeatedly finds the next most violated constraint…
- …until set of constraints is a good approximation.

Illustrative Example



SSVM Problem

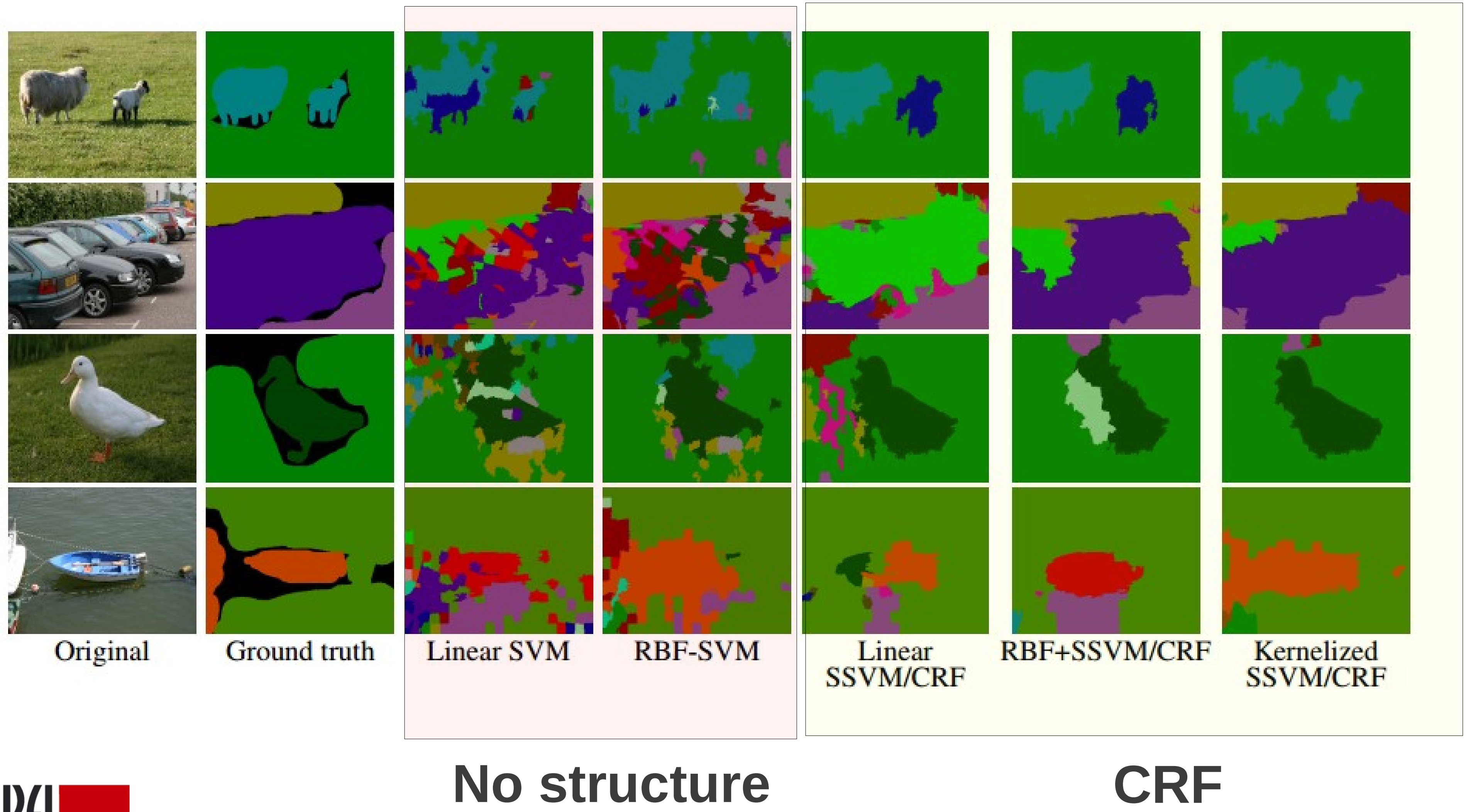
- Exponential constraints
- Most are dominated by a small set of "important" constraints



Cutting plane approach

- Repeatedly finds the next most violated constraint…
- …until set of constraints is a good approximation.

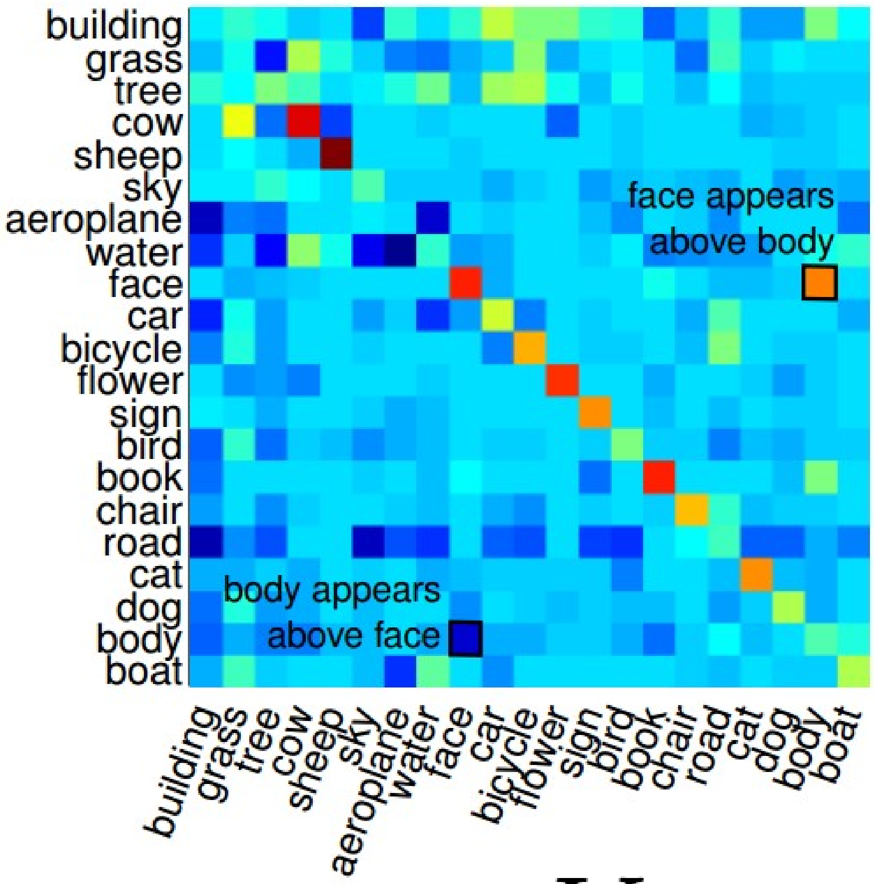
Results



http://infoscience.epfl.ch/record/180190/files/lucchi12.pdf

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Learned pairwise term



$$\mathbf{w}^V$$

- The pairwise term in matrix form where columns indicate classes $\mathbf{y}_i$ belonging to superpixel i and rows indicate classes $\mathbf{y}_j$ belonging to neighboring j for the MSRC-21 dataset.

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Drawbacks

- Finding the most violated constraint at each iteration of the cutting plane is intractable in loopy graphical models.
- Approximations can sometimes be imprecise enough to have a major impact on learning
  - An unsatisfactory constraint can cause the cutting plane algorithm to prematurely terminate.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

3.2. Maximum Margin Training of Structured Models: online subgradient descent (SGD)

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

SGD approach

- Can reformulate the problem as an unconstrained optimization by plugging the constraints in the objective function.

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{2} |\mathbf{w}|_2^2 + \frac{C}{N} \sum_{n=1}^{N} [E_{\mathbf{w}}(Y^n) + \Delta(Y^n, Y^*) - E_{\mathbf{w}}(Y^*)]_+$$

- SGD approach : compute and step in the negative direction of a sub-gradient of $\mathcal{L}_w$
- See N. Ratliff, J. A. Bagnell, and M. Zinkevich. (Online) Subgradient Methods for Structured Prediction. In AISTATS, 2007.

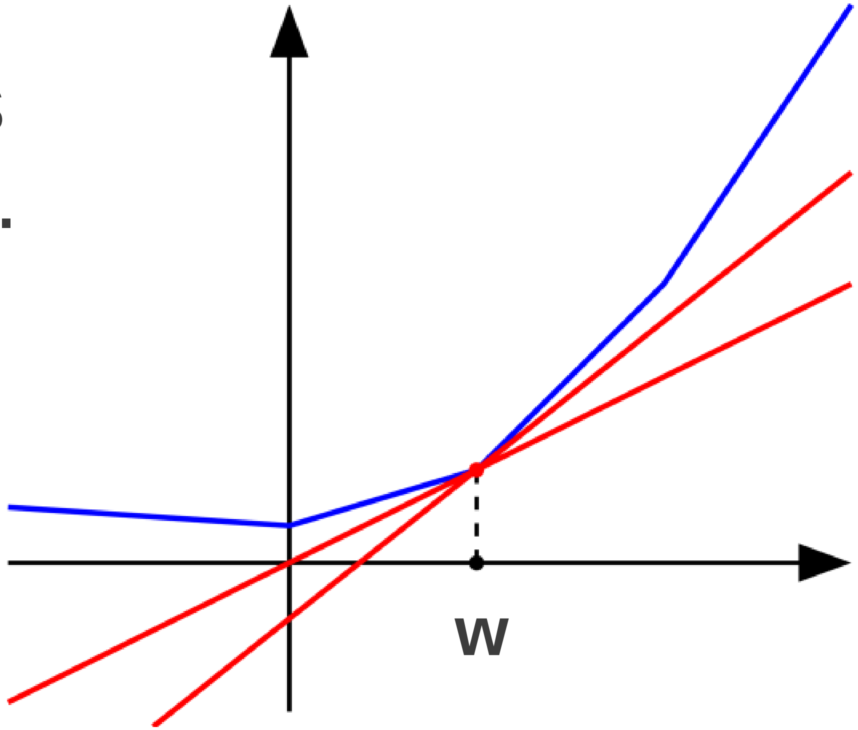
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Subgradient

- A subgradient for the convex loss function **L** at **w** is defined as a vector **g**, such that:

$$\forall \mathbf{w}' \in \mathcal{W}, \mathbf{g}^T(\mathbf{w}' - \mathbf{w}) \leq l(\mathbf{w}') - l(\mathbf{w})$$

- Set of all subgradients is called the subdifferential.
- If **L** is differentiable at **w**, then **g** is the gradient .



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

SGD approach

Algorithm 1

INPUTS :

$\mathcal{D}$: training set.

λ : multiplicative factor for step-size.

$\mathbf{w_0}$: arbitrary initial values, e.g. 0.

OUTPUT : $\mathbf{w_T}$

for $t = 1 \ldots T$ **do**

  for all examples (X^n, Y^n) in $\mathcal{D}$ **do**

    $Y^* = \arg\min_{Y \in \mathcal{Y}_n} (E_{\mathbf{w}}(Y) - \Delta(Y^n, Y))$

    $\eta_t \leftarrow \frac{\lambda}{t}$

    $g_t \leftarrow \frac{\partial l(Y^n, Y, \mathbf{w_{t-1}})}{\partial \mathbf{w_{t-1}}}$

    $w_t \leftarrow w_{t-1} - \eta_t g_t$

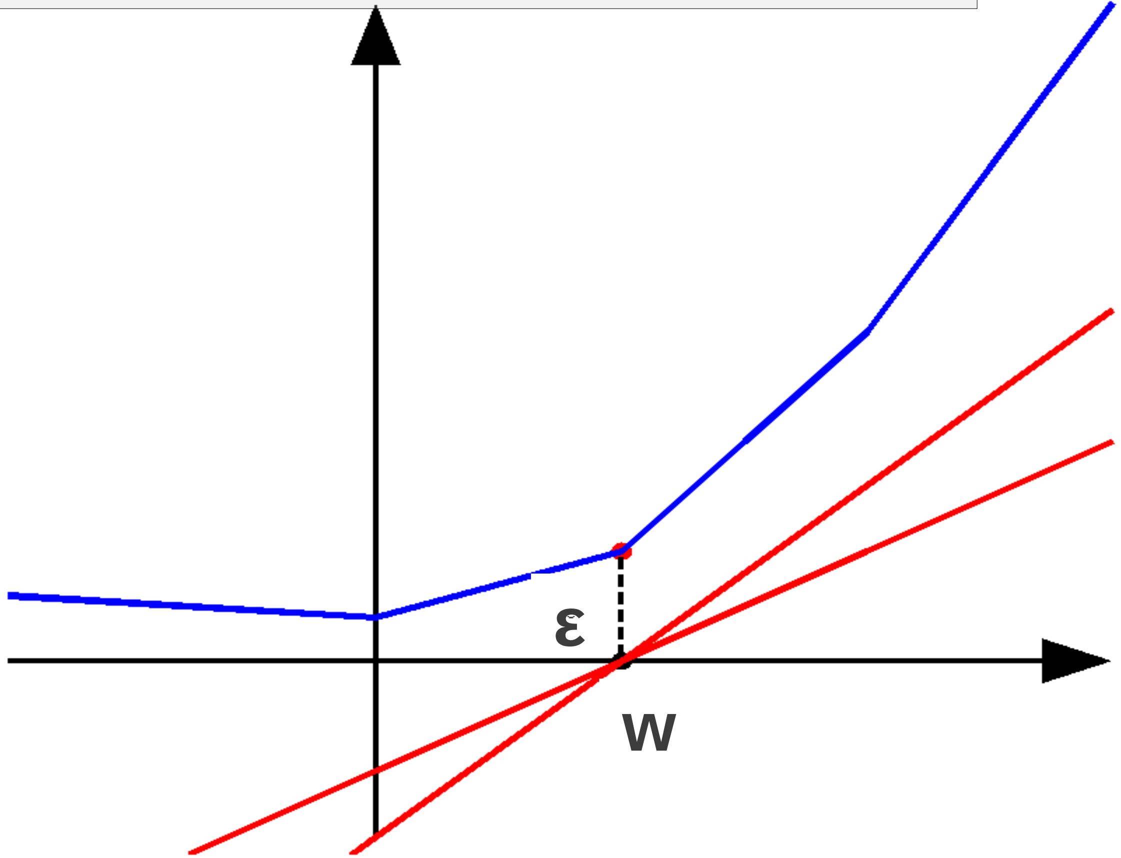
  end for

end for

What can we say about approximate subgradients ?

- Epsilon subgradients:

$$\forall \mathbf{w}' : \mathbf{g}^T(\mathbf{w} - \mathbf{w}') \geq l(\mathbf{w}) - l(\mathbf{w}') - \epsilon$$



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

What can we say about approximate subgradients ?

- Convergence guarentees (see S. M. Robinson. Linear convergence of epsilon-subgradient descent methods for a class of convex functions. Mathematical Programming, 86:41–50, 1999):

$$\mathcal{L}(\mathbf{w}^{best}) - \mathcal{L}(\mathbf{w}^*) \leq \frac{R^2 + \sum_i (\eta^{(i)})^2 G^2}{2 \sum_i \eta^{(i)}} + \epsilon$$

$$\mathbf{w}^{best} = argmin_{\mathbf{w}^{(t)}} \mathcal{L}(\mathbf{w}^{(}t))$$

$$\left|\left|\mathbf{w}^{(1)} - \mathbf{w}^*\right|\right|^2 \leq R^2 \qquad ||\mathbf{g}||^2 \leq G^2$$

Goes to 0 with appropriate step size

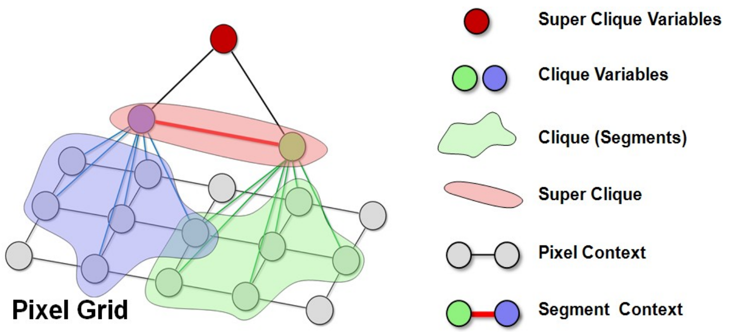
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

What can we say about approximate subgradients ?

- Models in computer vision are getting more complex.

- Need better approximation algorithms.

Ladický, Russell, Kohli, Torr ICCV09



$$\mathcal{L}(\mathbf{w}^{best}) - \mathcal{L}(\mathbf{w}^*) \leq \frac{R^2 + \sum_i (\eta^{(i)})^2 G^2}{2\sum_i \eta^{(i)}} + \epsilon$$

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Experimental results

	QP formulation	SGD + argmax	SGD + sampling	SGD + argmax CVPR13
Time for 1000 iterations	19628s	5315s	2481s	5842s
VOC score on EM dataset	80.5 %	79.9 %	77.5 %	84.5 %

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Future challenges

- Better energy functions, fully connected CRFs...
- Higher order potentials.
- Better and faster training algorithms.

Increasing connectivity

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Questions

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Resources

- http://cvlab.epfl.ch/research/medical/em/mitochond
- http://cvlab.epfl.ch/research/medical/em/synapses/
- http://cvlab.epfl.ch/~lucchi/

Credits

- Slides courtesy :
  - Christoph Lampert (Learning with Structured Inputs and Outputs)
  - Pushmeet Kohli (Efficiently Solving Dynamic Markov Random Fields using Graph Cuts)
  - Ben Taskar (Structured Prediction: A Large Margin Approach, NIPS tutorial)
  - Yisong Yue, Thorsten Joachims (An Introduction to Structured Output Learning Using Support Vector Machines)