Discrete Inference and Learning
Lecture 4

MVA

2020 – 2021

http://thoth.inrialpes.fr/~alahari/disinflearn

Slides based on material from Nikos Komodakis, M. Pawan Kumar

Outline

- Previous classes
  - Graph cuts, Belief propagation and variants
  - (Inference)

- Today
  - Quick recap of the course
  - Learning parameters

Before moving on…

Project suggestions (also sent by email)

- Implement BP on trees, then graph, extend to TRW, compare
- Implement graph cut + extension (Ishikawa, other multi-label) or variation of implementation + small application
- Complex application of graph cut, requiring modelling (e.g., sequence of images)
- Geometric scene labelling with graph cuts
- Joint modelling of two labelling problems (e.g., segmentation + detection)
- Implement fast primal-dual algorithm + evaluate
- Implement deformable parts model for object detection
- …

- Or your own (but check with us first)
- **Select projects before 25th January and email us (karteek.alahari@inria.fr, guillaume.charpiat@inria.fr)**

Projects

- **Choose projects before 25/1** (Monday!)

- Presentations on 31/3
  - In English or French
  - 15min, including questions

- Report due on 30/3

Recap

- What inference algorithm would you use for
  - a graph with only chains
    - 2-label problem ?
    - Multi-label problem ?
  - Tree structured graph
    - 2 label problem ?
    - Multi-label problem ?

Recap

- Basics: problem formulation
  - Energy Function
  - MAP Estimation
  - Computing min-marginals
  - Reparameterization

- Solutions
  - Belief Propagation and related methods [Lecture 3]
  - Graph cuts [Lecture 2]

Outline

- Recap of the course
- Learning parameters

Conditional Random Fields (CRFs)

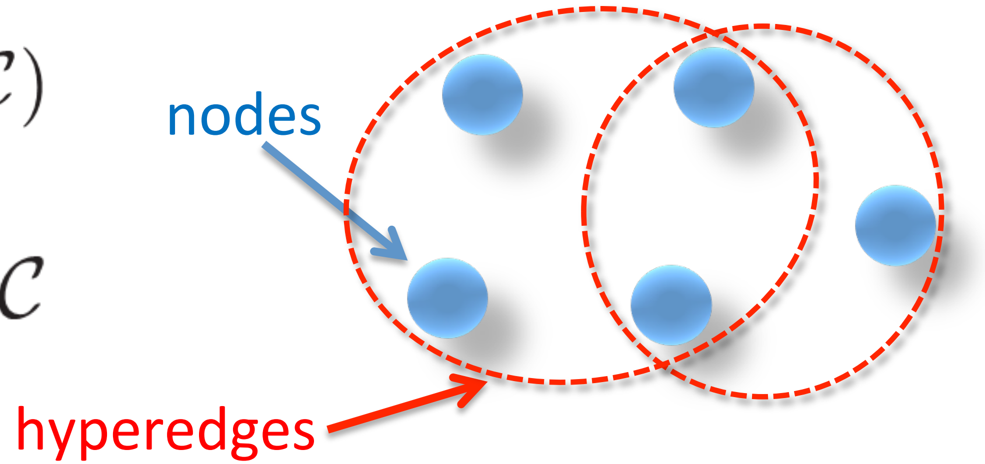
- Ubiquitous in computer vision
  - segmentation / stereo matching
    optical flow / image restoration
    image completion / object detection/localization
    ...
- and beyond
  - medical imaging, computer graphics, digital communications, physics…
- Really powerful formulation

Conditional Random Fields (CRFs)

- Key task: inference/optimization for CRFs/MRFs
- Extensive research for more than 20 years
- Lots of progress
- Many state-of-the-art methods:
  - Graph-cut based algorithms
  - Message-passing methods
  - LP relaxations
  - Dual Decomposition
  - ….

MAP inference for CRFs/MRFs

- Hypergraph $G = (\mathcal{V}, \mathcal{C})$
  - Nodes $\mathcal{V}$
  - Hyperedges/cliques $\mathcal{C}$



- High-order MRF energy minimization problem

$$\mathrm{MRF}_G(\mathbf{U}, \mathbf{H}) \equiv \min_{\mathbf{x}} \sum_{q \in \mathcal{V}} \underbrace{U_q(x_q)}_{\substack{\text{unary potential} \\ \text{(one per node)}}} + \sum_{c \in \mathcal{C}} \underbrace{H_c(\mathbf{x}_c)}_{\substack{\text{high-order potential} \\ \text{(one per clique)}}}$$

CRF training

- But how do we choose the CRF potentials?

- Through training
  - Parameterize potentials by **w**
  - Use training data to <u>learn</u> correct **w**

- Characteristic example of structured output learning [Taskar], [Tsochantaridis, Joachims]

- Equally, if not more, important than MAP inference
  - Better optimize correct energy (even approximately)
  - Than optimize wrong energy exactly

Outline

- Supervised Learning
- Probabilistic Methods
- Loss-based Methods
- Results

Image Classification

Is this an urban or rural area?

Input: **d**

Output: $\mathbf{x} \in \{-1,+1\}$

Image Classification

Is this scan healthy or unhealthy?

Input: **d** Output: $\mathbf{x} \in \{-1,+1\}$

Image Classification

Labeling **X** = **x**

Label set **L** = {-1,+1}

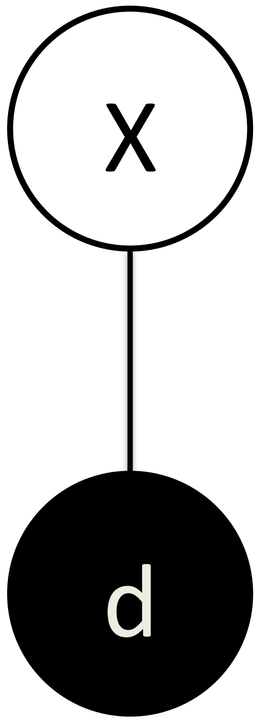


Image Classification

Which city is this?

Input: **d** Output: $\mathbf{x} \in \{1,2,\ldots,h\}$

Image Classification

What type of tumor does this scan contain?

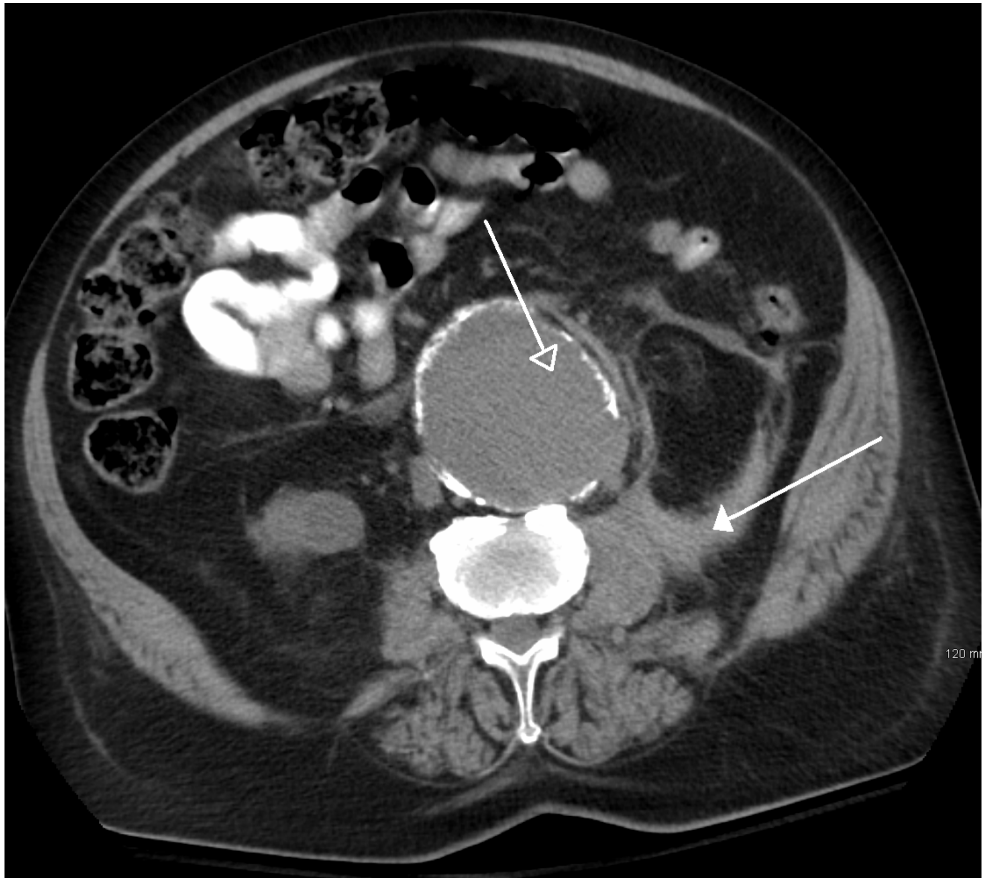
Input: **d** Output: $\mathbf{x} \in \{1,2,\ldots,h\}$

Object Detection

Where is the object in the image?

Input: **d** Output: $\mathbf{x} \in$ {Pixels}

Object Detection



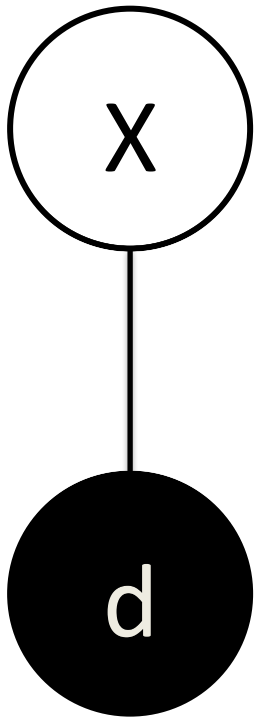
Where is the rupture in the scan?

Input: **d**

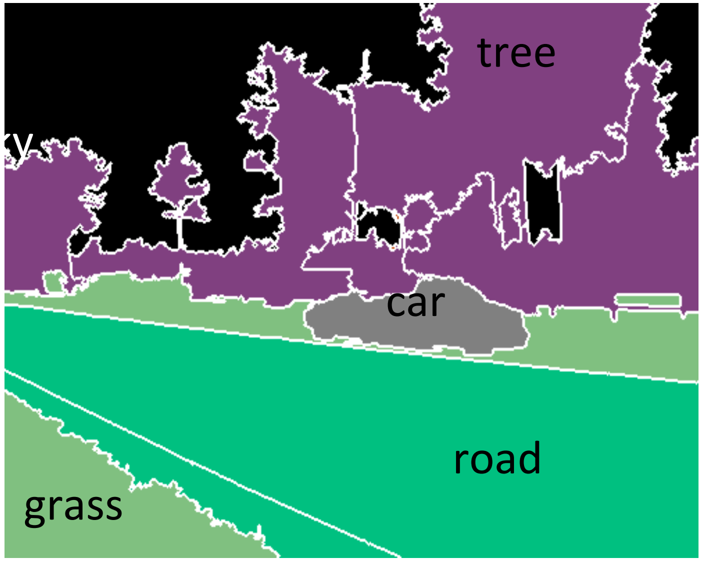
Output: $\mathbf{x} \in \{\text{Pixels}\}$

Object Detection

Labeling **X** = **x** Label set **L** = {1, 2, ..., h}



Segmentation

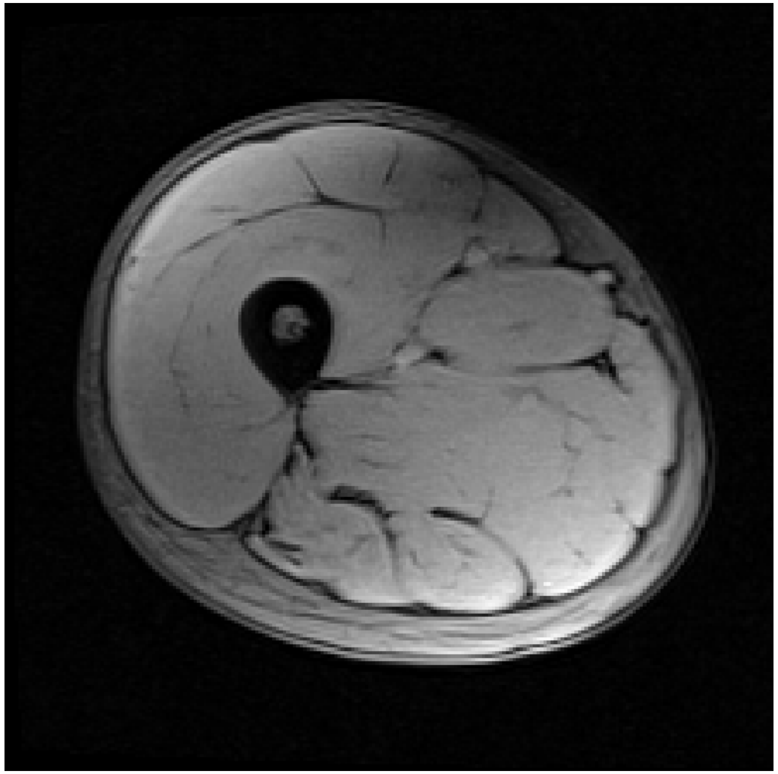


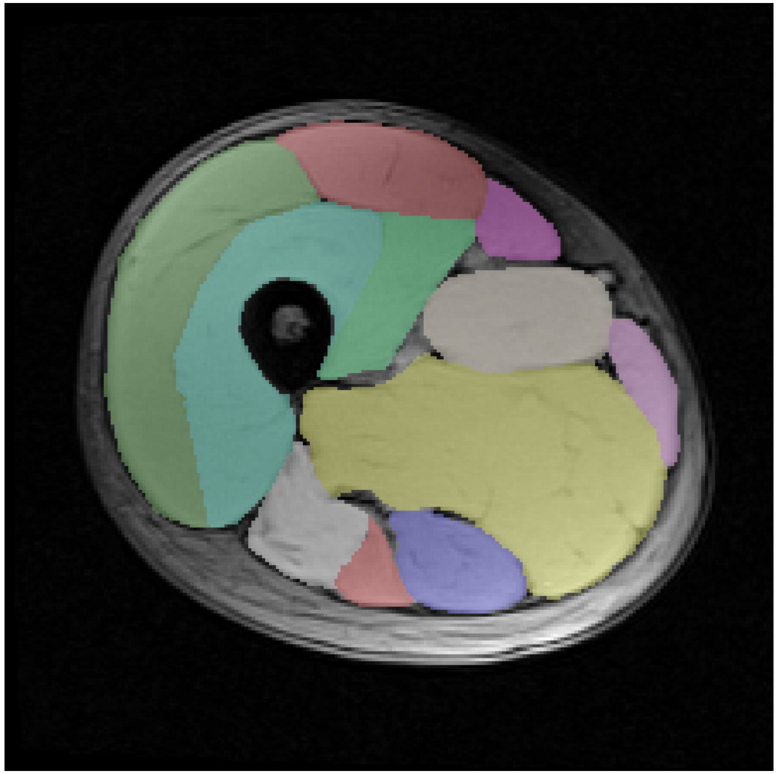
What is the semantic class of each pixel?

Input: **d**

Output: $\mathbf{x} \in \{1,2,\ldots,h\}^{|\mathrm{Pixels}|}$

Segmentation





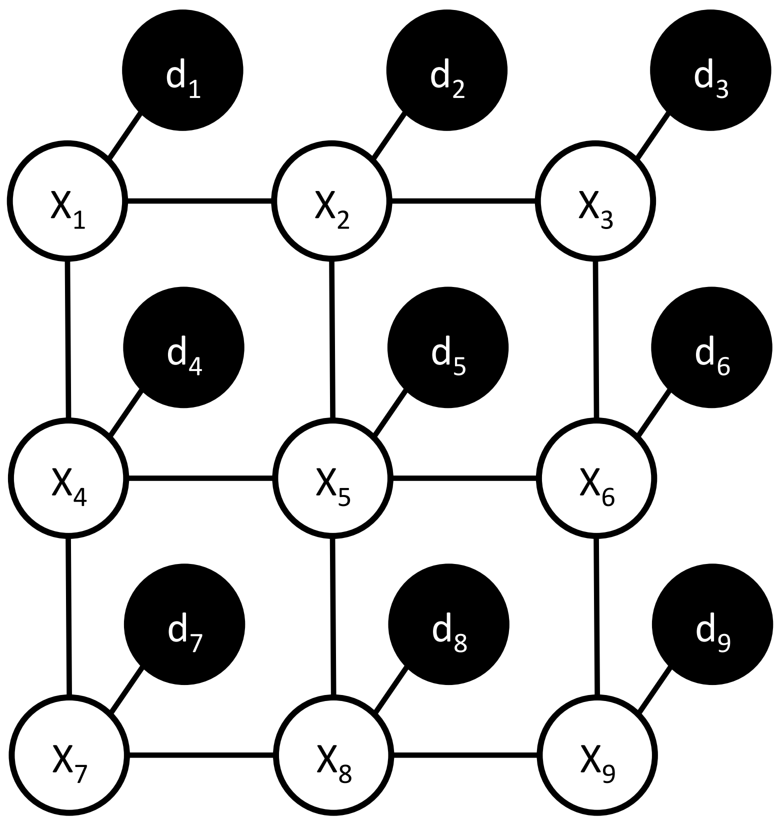
What is the muscle group of each pixel?

Input: $\mathbf{d}$

Output: $\mathbf{x} \in \{1,2,\ldots,h\}^{|\mathrm{Pixels}|}$

Segmentation

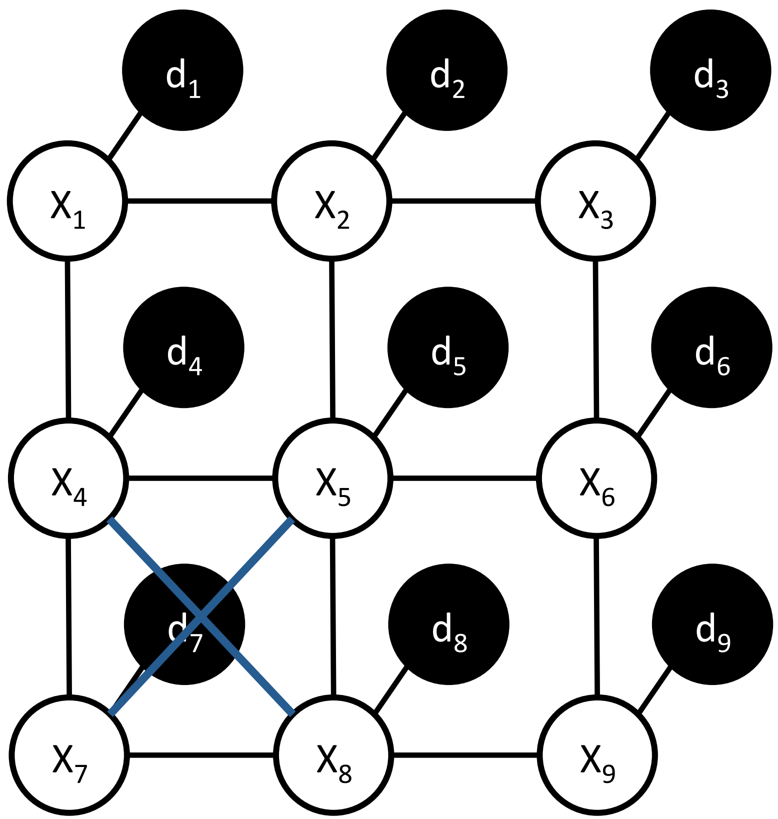
Labeling $\mathbf{X} = \mathbf{x}$ Label set $\mathbf{L} = \{1, 2, ..., h\}$



Segmentation

Labeling **X** = **x**

Label set **L** = {1, 2, ..., h}

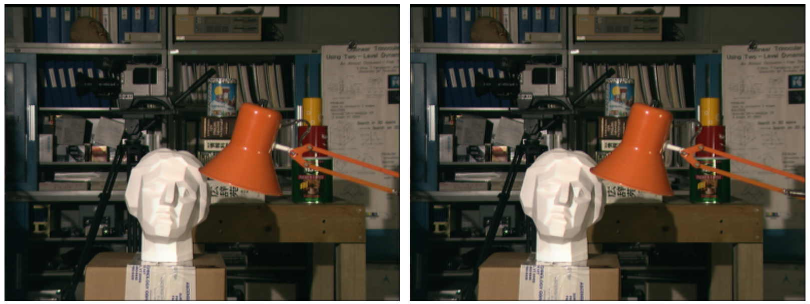


CRF training

- Stereo matching:
  - Z: left, right image
  - X: disparity map

Goal of training:
estimate proper **w**

$f:$



Z

X

$$f = \arg\min_{\mathbf{x}} \mathrm{MRF}_G(\mathbf{x}; \mathbf{u}, \mathbf{h})$$

parameterized by **w**

CRF training

- Denoising:
  - Z: noisy input image
  - X: denoised output image

Goal of training:
estimate proper **w**

f :

Z

X

$$f = \arg\min_{\mathbf{x}} \mathrm{MRF}_G(\mathbf{x}; \mathbf{u}, \mathbf{h})$$

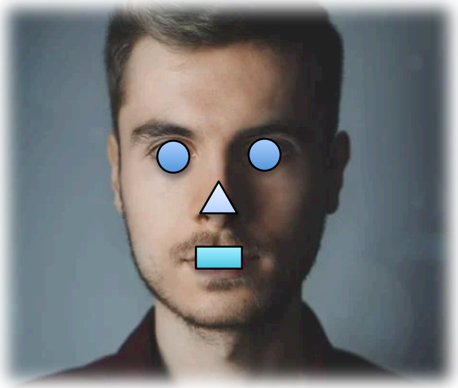
parameterized by **w**

CRF training

- Object detection:
  - Z: input image
  - X: position of object parts

Goal of training:
estimate proper **w**

$f:$ 



Z

X

$$f = \arg\min_{\mathbf{x}} \mathrm{MRF}_G(\mathbf{x}; \mathbf{u}, \mathbf{h})$$

parameterized by **w**

CRF training (some further notation)

$$\mathrm{MRF}_G(\mathbf{x};\mathbf{u}^k,\mathbf{h}^k) = \sum_p u_p^k(x_p) + \sum_c h_c^k(\mathbf{x}_c)$$

$$u_p^k(x_p) = \mathbf{w}^T g_p(x_p,\mathbf{z}^k), \;\; h_c^k(\mathbf{x}_c) = \mathbf{w}^T g_c(\mathbf{x}_c,\mathbf{z}^k)$$

vector valued feature functions

$$\mathrm{MRF}_G(\mathbf{x};\mathbf{w},\mathbf{z}^k) = \mathbf{w}^T \left(\sum_p g_p(x_p,\mathbf{z}^k) + \sum_c g_c(\mathbf{x}_c,\mathbf{z}^k) \right) = \mathbf{w}^T g(\mathbf{x},\mathbf{z}^k)$$

Learning formulations

Risk minimization

$$\hat{\mathbf{x}}^k = \arg\min_{\mathbf{x}} \mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k)$$

$$\min_{\mathbf{w}} \sum_{k=1}^{K} \Delta\left(\mathbf{x}^k, \hat{\mathbf{x}}^k\right)$$

K training samples $\left\{(\mathbf{x}^k, \mathbf{z}^k)\right\}_{k=1}^{K}$

Regularized Risk minimization

$$\hat{\mathbf{x}}^k = \arg\min_{\mathbf{x}} \mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k)$$

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} \Delta\left(\mathbf{x}^k, \hat{\mathbf{x}}^k\right)$$

$$R(\mathbf{w}) = ||\mathbf{w}||^2, \;\; ||\mathbf{w}||_1, \;\; \text{etc.}$$

Regularized Risk minimization

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

Replace Δ(.) with easier to handle upper bound L_G
(e.g., convex w.r.t. **w**)

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} \Delta\left(\mathbf{x}^k, \hat{\mathbf{x}}^k\right)$$

Choice 1: Hinge loss

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

$$\boxed{L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) - \min_{\mathbf{x}}\left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)}$$

- Upper bounds Δ(.)
- Leads to **max-margin learning**

Max-margin learning

$$\mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) \leq \mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k) + \xi_k$$

energy of ground truth

any other energy

desired margin

slack

Max-margin learning

$$\min_{\mathbf{w}} \qquad \sum_k \xi_k$$

subject to the constraints:

$$\mathrm{MRF}_G(\mathbf{x}^k;\mathbf{w},\mathbf{z}^k) \leq \mathrm{MRF}_G(\mathbf{x};\mathbf{w},\mathbf{z}^k) - \Delta(\mathbf{x},\mathbf{x}^k) + \xi_k$$

energy of ground truth | any other energy | desired margin | slack

Max-margin learning

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_k \xi_k$$

subject to the constraints:

$$\underbrace{\mathrm{MRF}_G(\mathbf{x}^k;\mathbf{w},\mathbf{z}^k)}_{\text{energy of ground truth}} \leq \underbrace{\mathrm{MRF}_G(\mathbf{x};\mathbf{w},\mathbf{z}^k)}_{\text{any other energy}} - \underbrace{\Delta(\mathbf{x},\mathbf{x}^k)}_{\text{desired margin}} + \underbrace{\xi_k}_{\text{slack}}$$

Max-margin learning

CONSTRAINED

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_k \xi_k$$

subject to the constraints:

$$\mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) \leq \mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k) + \xi_k$$

or equivalently

UNCONSTRAINED

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_k \xi_k$$

$$\xi_k = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) - \min_{\mathbf{x}} \left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)$$

Choice 2: logistic loss

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) + \underbrace{\log \sum_{\mathbf{x}} e^{-\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k)}}_{\text{partition function}}$$

- Can be shown to lead to **maximum likelihood learning**

Max-margin vs Maximum-likelihood

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \overbrace{\boxed{\mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k)} \boxed{- \min_{\mathbf{x}}} \left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)}^{\text{max-margin}}$$

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \underbrace{\boxed{\mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k)} + \boxed{\log \sum_{\mathbf{x}} e}^{-\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k)}}_{\text{maximum likelihood}}$$

Max-margin vs Maximum-likelihood

$$L_G\left(\mathbf{x}^k,\mathbf{z}^k;\mathbf{w}\right) = \overbrace{\boxed{\mathrm{MRF}_G(\mathbf{x}^k;\mathbf{w},\mathbf{z}^k)} + \boxed{\max_{\mathbf{x}}}\left(-\mathrm{MRF}_G(\mathbf{x};\mathbf{w},\mathbf{z}^k) + \Delta(\mathbf{x},\mathbf{x}^k)\right)}^{\text{max-margin}}$$

soft-max

$$L_G\left(\mathbf{x}^k,\mathbf{z}^k;\mathbf{w}\right) = \underbrace{\boxed{\mathrm{MRF}_G(\mathbf{x}^k;\mathbf{w},\mathbf{z}^k)} + \boxed{\log\sum_{\mathbf{x}} e}^{-\mathrm{MRF}_G(\mathbf{x};\mathbf{w},\mathbf{z}^k)}}_{\text{maximum likelihood}}$$

Solving the learning formulations

Maximum-likelihood learning

$$\min_{\mathbf{w}} \frac{\mu}{2}||\mathbf{w}||^2 + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) + \underbrace{\log \sum_{\mathbf{x}} e^{-\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k)}}_{\text{partition function}}$$

- Differentiable & convex
- Global optimum via gradient descent, for example

Maximum-likelihood learning

$$\min_{\mathbf{w}} \frac{\mu}{2}||\mathbf{w}||^2 + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) + \log \sum_{\mathbf{x}} e^{-\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k)}$$

gradient $\longrightarrow \nabla_{\mathbf{w}} = \mathbf{w} + \sum_k \left(g(\mathbf{x}^k, \mathbf{z}^k) - \sum_{\mathbf{x}} p(\mathbf{x}|w, \mathbf{z}^k) g(\mathbf{x}, \mathbf{z}^k) \right)$

Recall that: $\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) = \mathbf{w}^T g(\mathbf{x}, \mathbf{z}^k)$

Maximum-likelihood learning

$$\min_{\mathbf{w}} \frac{\mu}{2}||\mathbf{w}||^2 + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) + \log \sum_{\mathbf{x}} e^{-\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k)}$$

gradient $\longrightarrow \nabla_{\mathbf{w}} = \mathbf{w} + \sum_k \left(g(\mathbf{x}^k, \mathbf{z}^k) - \underbrace{\sum_{\mathbf{x}} p(\mathbf{x}|w, \mathbf{z}^k) g(\mathbf{x}, \mathbf{z}^k)} \right)$

- Requires MRF probabilistic inference
- **NP-hard** (exponentially many $\mathbf{x}$): approximation via loopy-BP ?

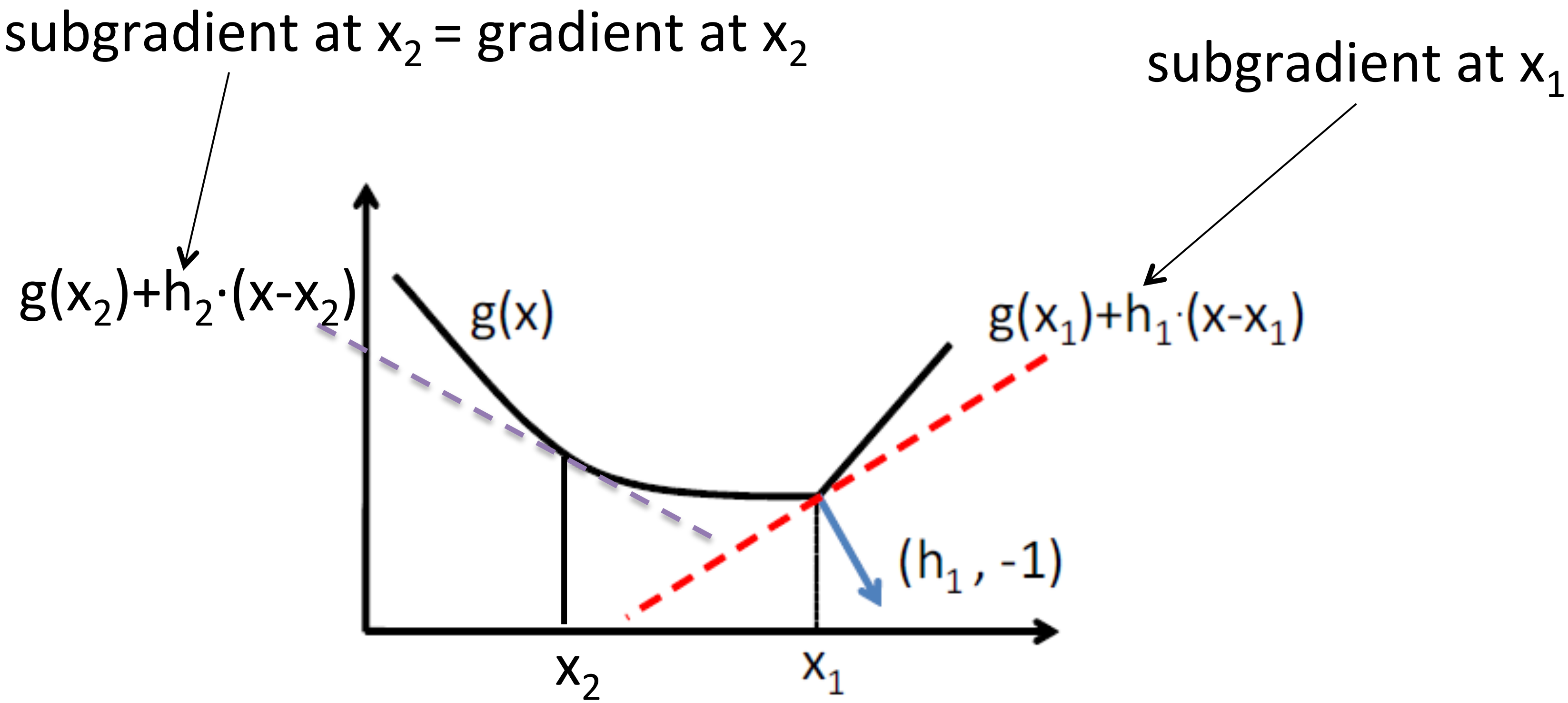
Max-margin learning (UNCONSTRAINED)

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) - \min_{\mathbf{x}}\left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)$$

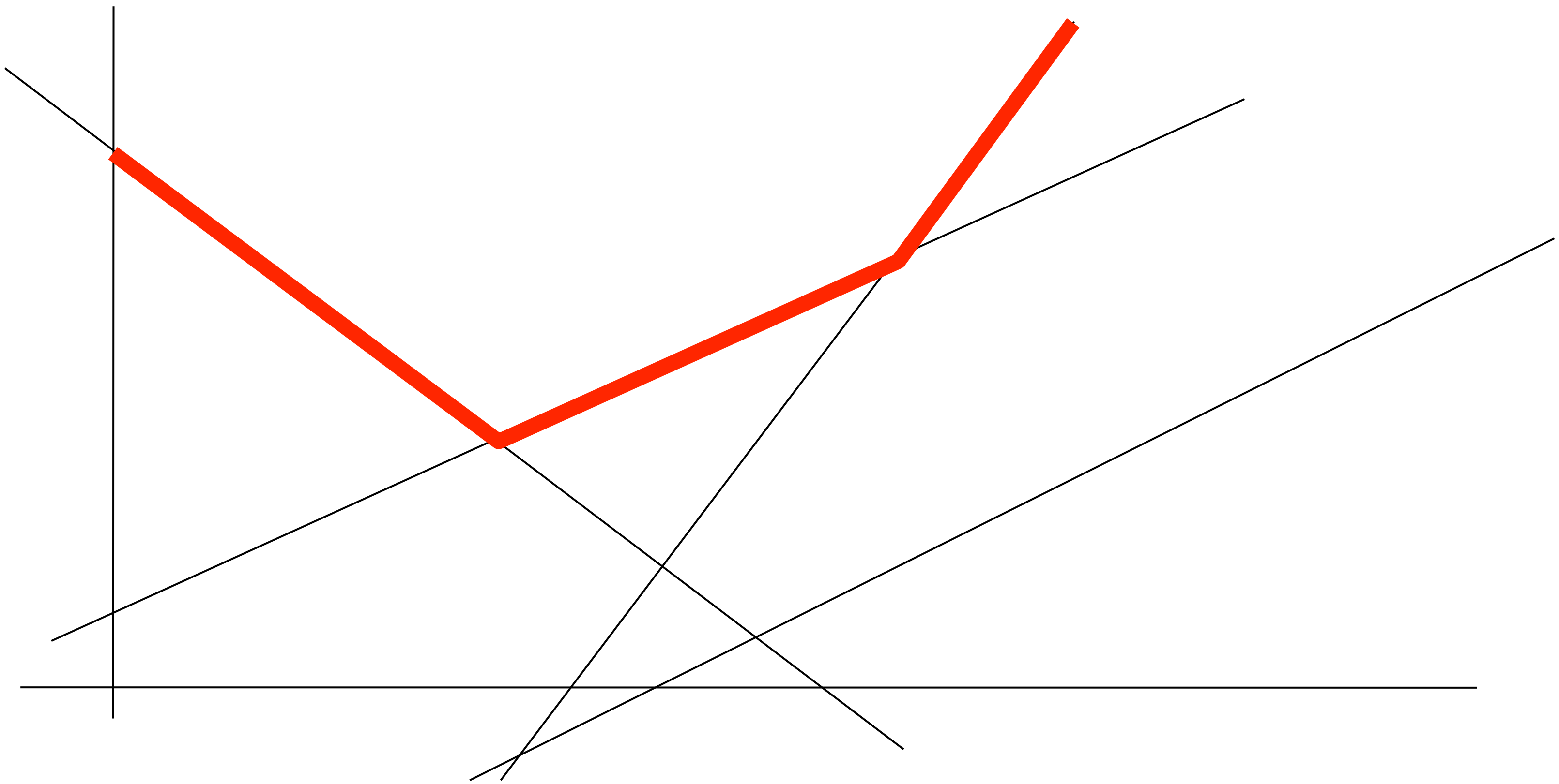
- Convex but non-differentiable
- Global optimum via **subgradient method**

Subgradient



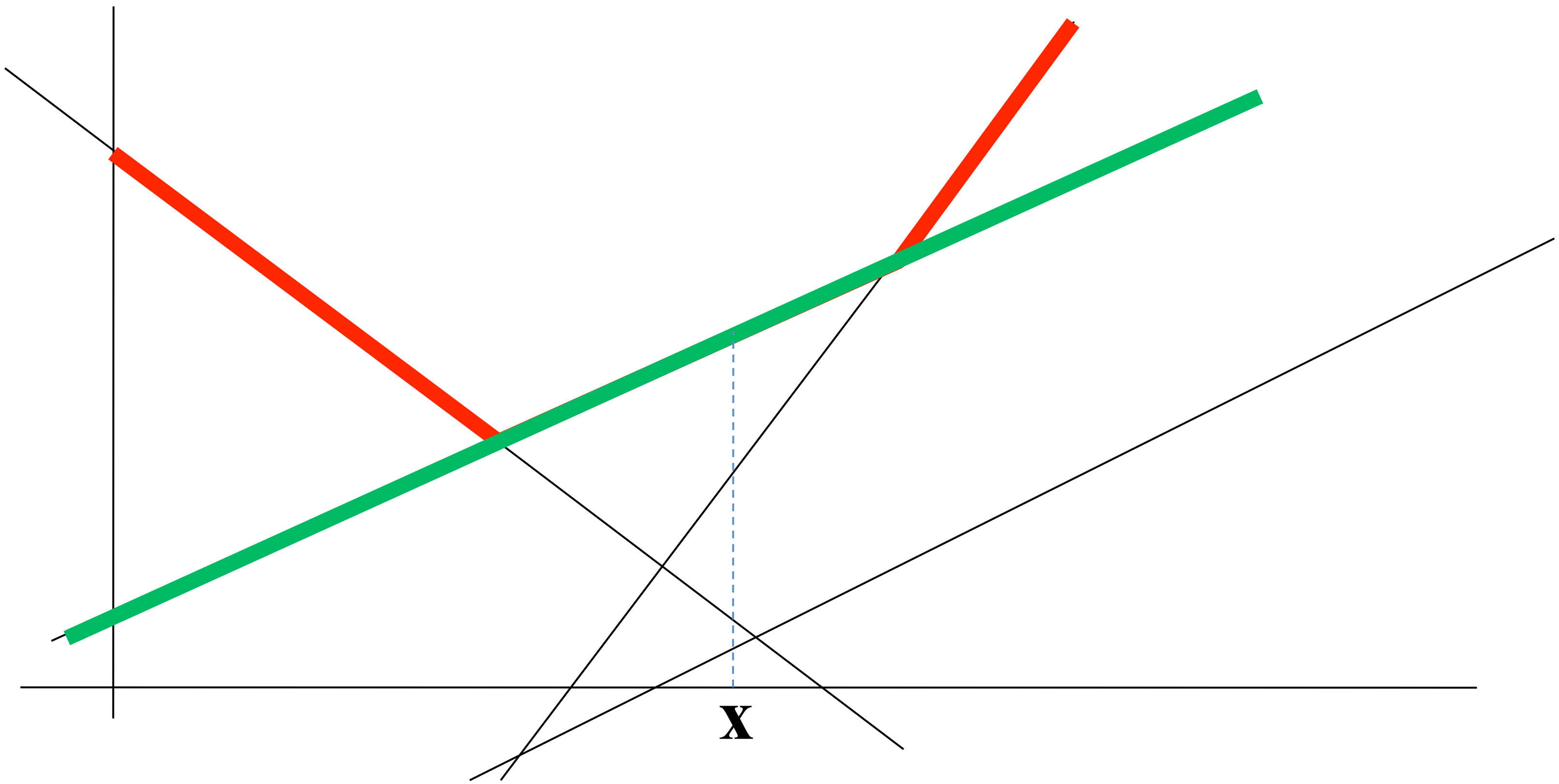
Subgradient

Lemma. *Let* $f(\cdot) = \max_{m=1,\ldots,M} f_m(\cdot)$, *with* $f_m(\cdot)$ *convex and differentiable. A subgradient of* f *at* $\mathbf{y}$ *is given by* $\nabla f_{\hat{m}}(\mathbf{y})$, *where* $\hat{m}$ *is any index for which* $f(\mathbf{y}) = f_{\hat{m}}(\mathbf{y})$.



Subgradient

Lemma. *Let* $f(\cdot) = \max_{m=1,\ldots,M} f_m(\cdot)$, *with* $f_m(\cdot)$ *convex and differentiable. A subgradient of* f *at* $\mathbf{y}$ *is given by* $\nabla f_{\hat{m}}(\mathbf{y})$, *where* $\hat{m}$ *is any index for which* $f(\mathbf{y}) = f_{\hat{m}}(\mathbf{y})$.



Subgradient

Lemma. *Let* $f(\cdot) = \max_{m=1,\ldots,M} f_m(\cdot)$*, with* $f_m(\cdot)$ *convex and differentiable. A subgradient of* f *at* $\mathbf{y}$ *is given by* $\nabla f_{\hat{m}}(\mathbf{y})$*, where* $\hat{m}$ *is any index for which* $f(\mathbf{y}) = f_{\hat{m}}(\mathbf{y})$.

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) - \min_{\mathbf{x}} \left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)$$

$$\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) = \mathbf{w}^T g(\mathbf{x}, \mathbf{z}^k)$$

subgradient of L_G = $g(\mathbf{x}^k, \mathbf{z}^k) - g(\hat{\mathbf{x}}^k, \mathbf{z}^k)$

$$\hat{\mathbf{x}}^k = \arg\min_{\mathbf{x}} \left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)$$

Max-margin learning (UNCONSTRAINED)

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) - \min_{\mathbf{x}}\left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)$$

Subgradient algorithm

Repeat

1. compute global minimizers $\hat{\mathbf{x}}^k$ at current $\mathbf{w}$
2. compute **total subgradient** at current $\mathbf{w}$
3. update $\mathbf{w}$ by taking a step in the negative total subgradient direction

until convergence

total subgr. $= \mathrm{subgradient}_{\mathbf{w}}[R(\mathbf{w})] + \sum_k \left(g(\mathbf{x}^k, \mathbf{z}^k) - g(\hat{\mathbf{x}}^k, \mathbf{z}^k)\right)$

Max-margin learning (UNCONSTRAINED)

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{k=1}^{K} L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right)$$

$$L_G\left(\mathbf{x}^k, \mathbf{z}^k; \mathbf{w}\right) = \mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) - \min_{\mathbf{x}}\left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)$$

Stochastic subgradient algorithm

Repeat

1. pick k at random
2. compute global minimizer $\hat{\mathbf{x}}^k$ at current $\mathbf{w}$
3. compute **partial subgradient** at current $\mathbf{w}$
4. update $\mathbf{w}$ by taking a step in the negative partial subgradient direction

until convergence

MRF-MAP estimation per iteration
(unfortunately NP-hard)

partial subgradient = $\mathrm{subgradient}_{\mathbf{w}}[R(\mathbf{w})] + g(\mathbf{x}^k, \mathbf{z}^k) - g(\hat{\mathbf{x}}^k, \mathbf{z}^k)$

Max-margin learning (CONSTRAINED)

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_k \xi_k$$

subject to the constraints:

$$\mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) \leq \mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k) + \xi_k$$

Max-margin learning (CONSTRAINED)

$$\min_{\mathbf{w}} \frac{\mu}{2}||\mathbf{w}||^2 + \sum_k \xi_k$$

subject to the constraints:

$$\mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) \leq \mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k) + \xi_k$$

linear in $\mathbf{w}$

- Quadratic program (great!)
- But exponentially many constraints (not so great)

Max-margin learning (CONSTRAINED)

- What if we use only a small number of constraints?
  - Resulting QP can be solved
  - But solution may be infeasible
- **Constraint generation** to the rescue
  - only few constraints **active** at optimal solution !!
    (variables much fewer than constraints)
  - Given the active constraints, rest can be ignored
  - Then let us try to find them!

Constraint generation

1. Start with some constraints
2. Solve QP
3. Check if solution is feasible w.r.t. to **all** constraints
4. If yes, we are done!
5. If not, pick a violated constraint and add it to the current set of constraints. Repeat from step 2.

   (optionally, we can also remove inactive constraints)

Constraint generation

- **Key issue:** we must always be able to find a violated constraint if one exists

- Recall the constraints for max-margin learning

$$\mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) \leq \mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k) + \xi_k$$

- To find violated constraint, we therefore need to compute:

$$\hat{\mathbf{x}}^k = \arg\min_{\mathbf{x}} \left(\mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k)\right)$$

(just like subgradient method!)

Constraint generation

1. Initialize set of constraints C to empty
2. Solve QP using current constraints C and obtain new $(\mathbf{w},\xi)$
3. Compute global minimizers $\hat{\mathbf{x}}^k$ at current $\mathbf{w}$
4. For each k, if the following constraint is violated then add it to set C:

$$\mathrm{MRF}_G(\mathbf{x}^k;\mathbf{w},\mathbf{z}^k) \leq \mathrm{MRF}_G(\hat{\mathbf{x}}^k;\mathbf{w},\mathbf{z}^k) - \Delta(\hat{\mathbf{x}}^k,\mathbf{x}^k) + \xi_k$$

5. If no new constraint was added then terminate. Otherwise go to step 2.

MRF-MAP estimation **per sample**
(unfortunately **NP-hard**)

Max-margin learning (CONSTRAINED)

$$\min_{\mathbf{w}} \frac{\mu}{2}||\mathbf{w}||^2 + \sum_k \xi_k$$

subject to the constraints:

$$\mathrm{MRF}_G(\mathbf{x}^k; \mathbf{w}, \mathbf{z}^k) \leq \mathrm{MRF}_G(\mathbf{x}; \mathbf{w}, \mathbf{z}^k) - \Delta(\mathbf{x}, \mathbf{x}^k) + \xi_k$$

- Alternatively, we can solve above QP in the **dual domain**
- dual variables $\leftrightarrow$ primal constraints
- Too many variables, but most of them zero at optimal solution
- Use a **working-set** method
(essentially dual to constraint generation)