Structured Learning via Logistic Regression

Justin Domke

NICTA and The Australian National University

NICTA

Australian National University

Structured Prediction: $y^* = \arg\max_y F(x, y)$.

Structured Learning: Fit F to optimize some empirical risk.

$\rightarrow$ Usually, energy is assumed linear: $F(x, y) = w^T \Phi(x, y)$.

Difficulty: Often hard to maximize F. Recent work (Meshi et al., 2010, Hazan and Urtasun 2012) phrases learning as a joint optimization:

$$\min_F \min_{\{\lambda^k\}} \sum_k \left[-F(x^k, y^k) + A(\lambda^k, \theta_F^k) \right].$$

Algorithm. Alternate:

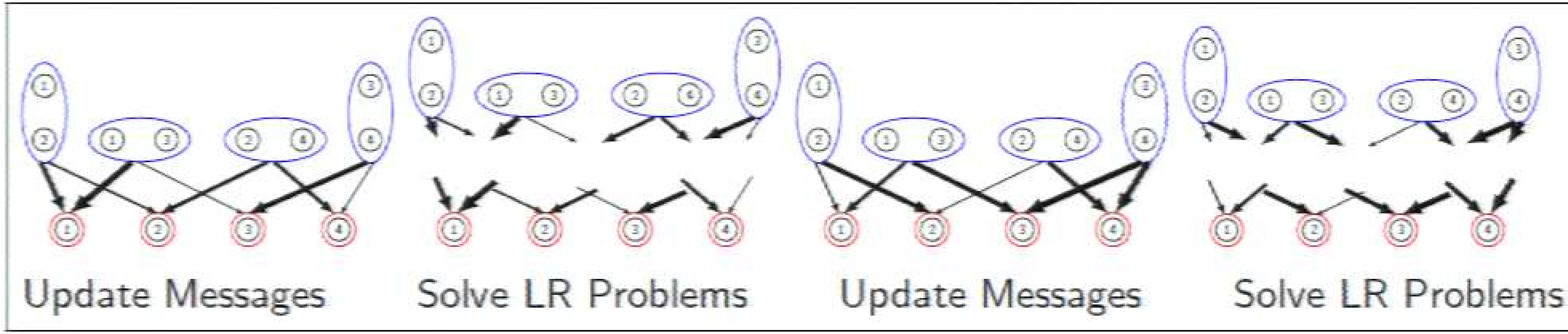
1. For each k, update λ^k using message-passing.
2. Perform gradient updates to F.

λ^k - Messages for datum k

θ_F^k - parameters for datum k (determined by F and Δ)

Main Result: For fixed messages $\{\lambda^k\}$ can optimize F by solving a set of logistic regression problems.

$\rightarrow$ Each datum is biased by the current messages.

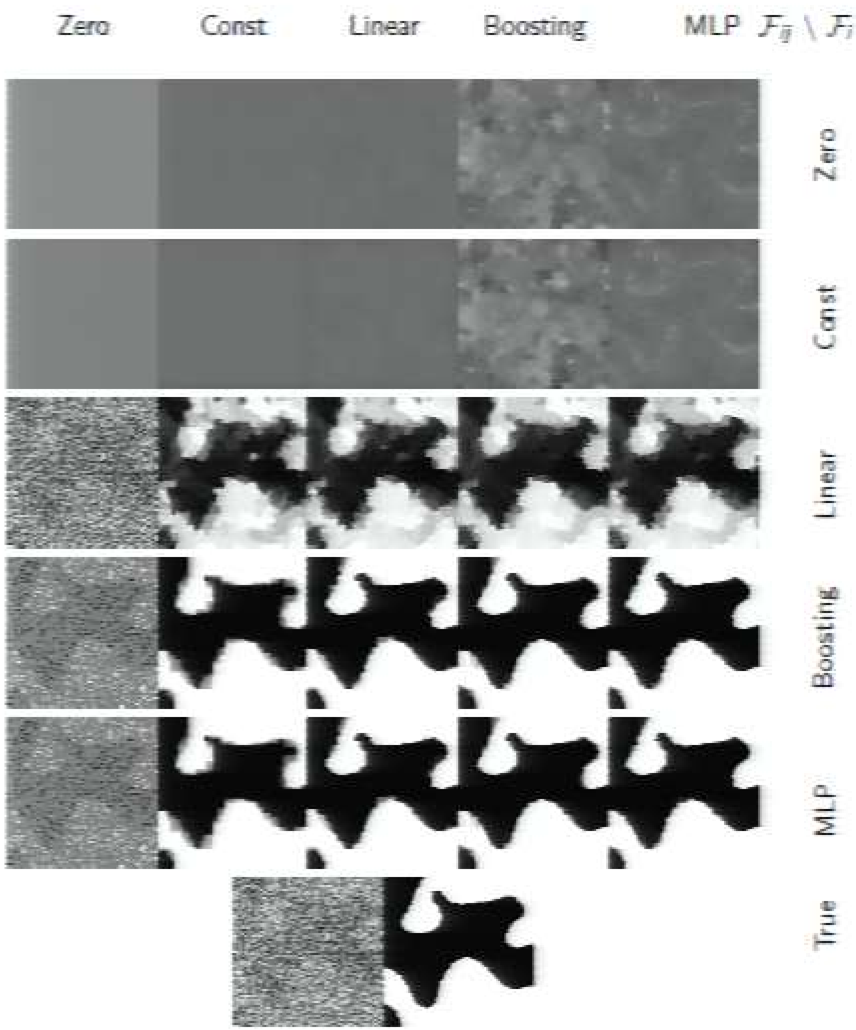


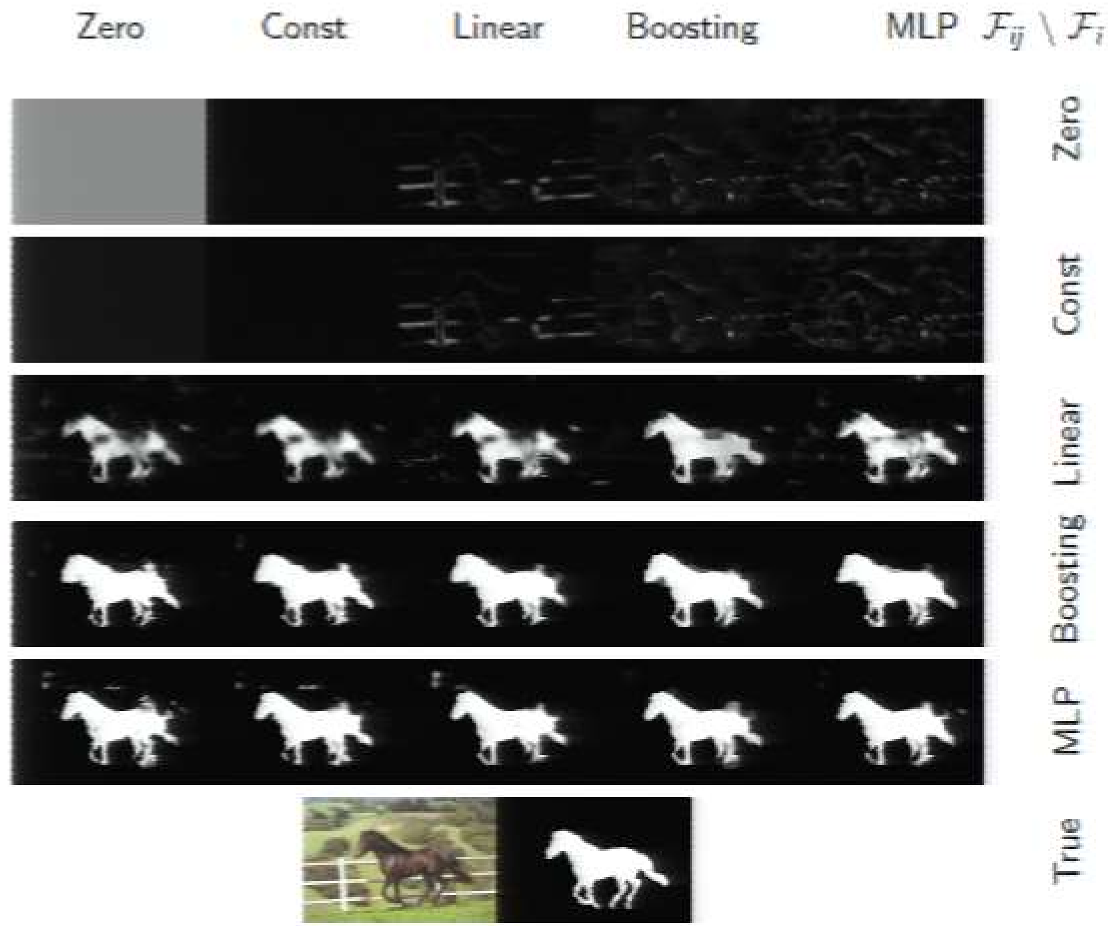
Algorithm:

- For each datum k do message passing to update λ^k
- Set $b_\alpha^k(y_\alpha) \leftarrow \frac{1}{\epsilon}\left(\Delta(y_\alpha^k, y_\alpha) + \sum_{\beta\subset\alpha} \lambda_\alpha^k(y_\beta) - \sum_{\gamma\supset\alpha} \lambda_\gamma^k(y_\alpha)\right).$
- For each region α, maximize logistic loss over $f_\alpha \in \mathcal{F}_\alpha$, biased by b_α.

Advantage: Use any function class you can fit a logistic loss to.

Each iteration similar to piecewise. Messages bias for joint prediction.





Denoising

$\mathcal{F}_i \setminus \mathcal{F}_{ij}$	Zero	Const.	Linear	Boost.	MLP
Zero	.502	.502	.502	.511	.502
Const.	.502	.502	.502	.510	.502
Linear	.444	.077	.059	.049	.034
Boost.	.444	.034	.015	.009	.007
MLP	.445	.032	.015	.009	.008

Horses

$\mathcal{F}_i \setminus \mathcal{F}_{ij}$	Zero	Const.	Linear	Boost.	MLP
Zero	.246	.246	.247	.244	.245
Const.	.246	.246	.247	.244	.245
Linear	.185	.185	.168	.154	.156
Boost.	.103	.098	.092	.084	.086
MLP	.096	.094	.087	.080	.081