Inference and Learning with Random Maximum A-Posteriori

Tamir Hazan
TTI Chicago

in collaboration with Tommi Jaakkola, Joseph Keshet, David McAllester, Raquel Urtasun, Koichiro Yamaguchi

Scene Understanding

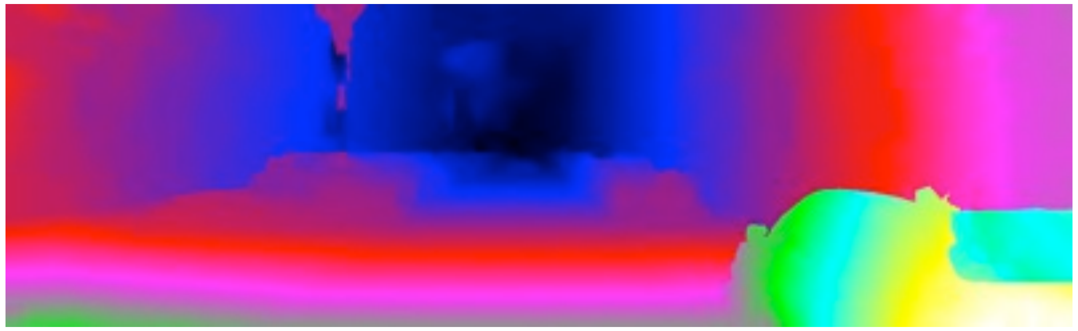
Maximum A-Posteriori (MAP)

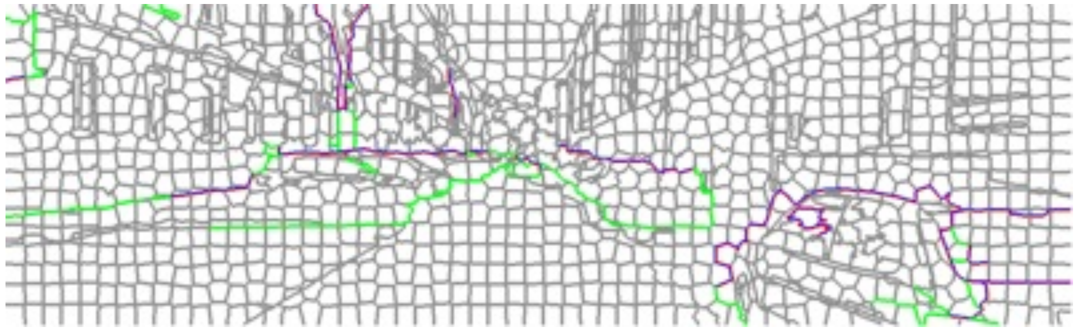
$$x^* = \operatorname*{argmax}_{x_1,\ldots,x_n} \ \phi(x_1,\ldots,x_n)$$

prediction　　　　　　　　scores

- Recently, many message-passing efficient MAP solvers for graphs with cycles: Graph-cuts, Gurobi, MPLP
- (Yamaguchi, Hazan, McAllester, Urtasun 2012)

	Middlebury (HR)	KITTI
Best other	7.0%	8.86%
Ours	4.4%	6.25%

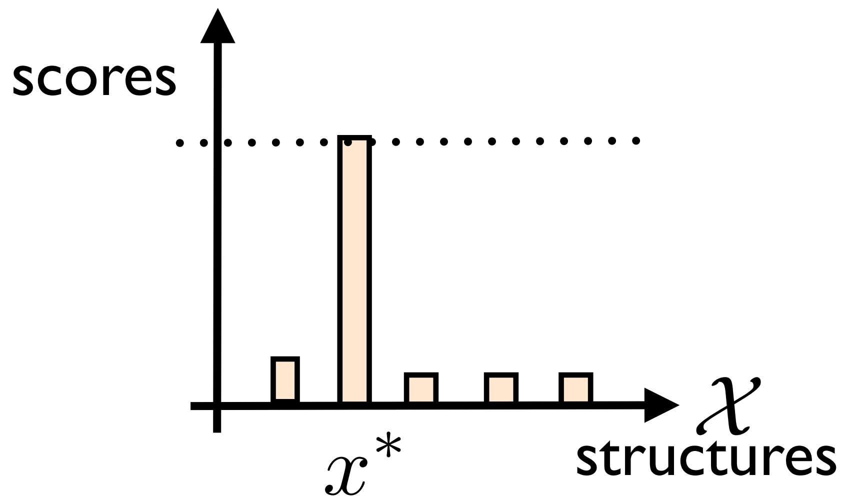




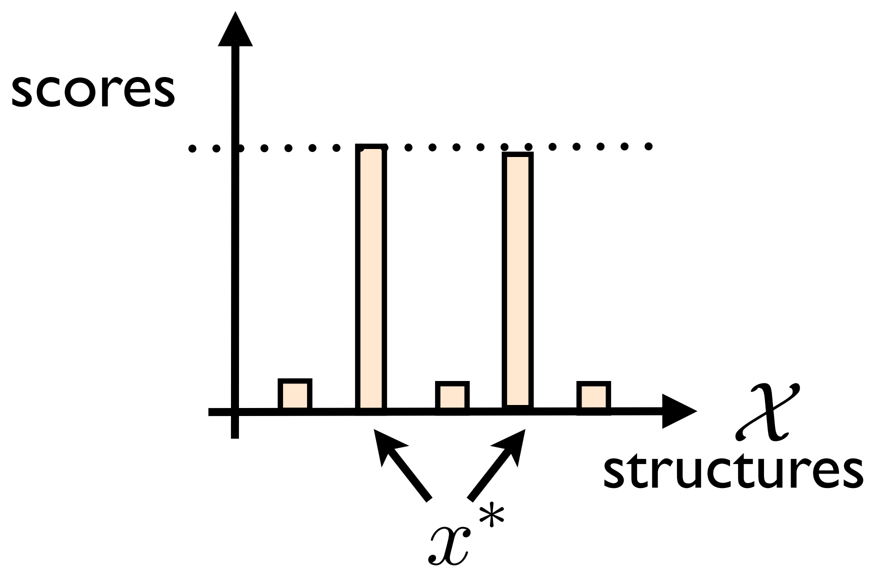
Inference & Learning with MAP

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} \{\phi(x)\}$$

easy

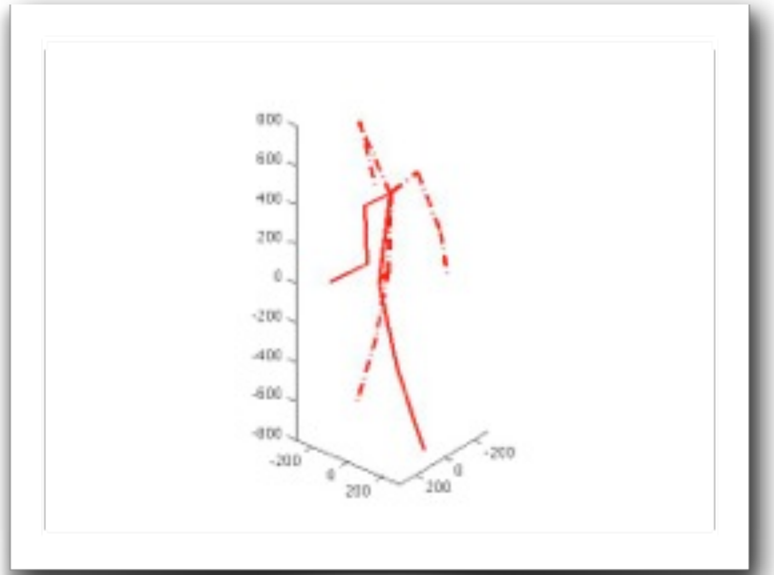


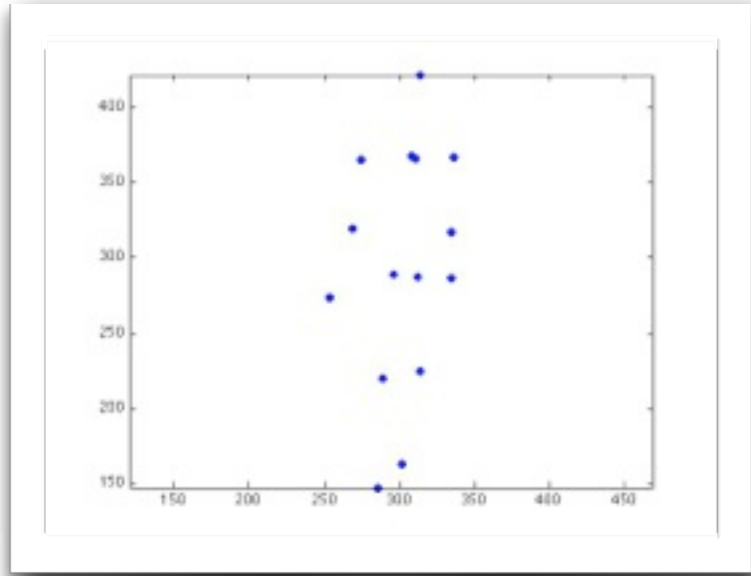
hard



Failures - Ambiguity

- Pose estimation: 3D joint locations from 2D images





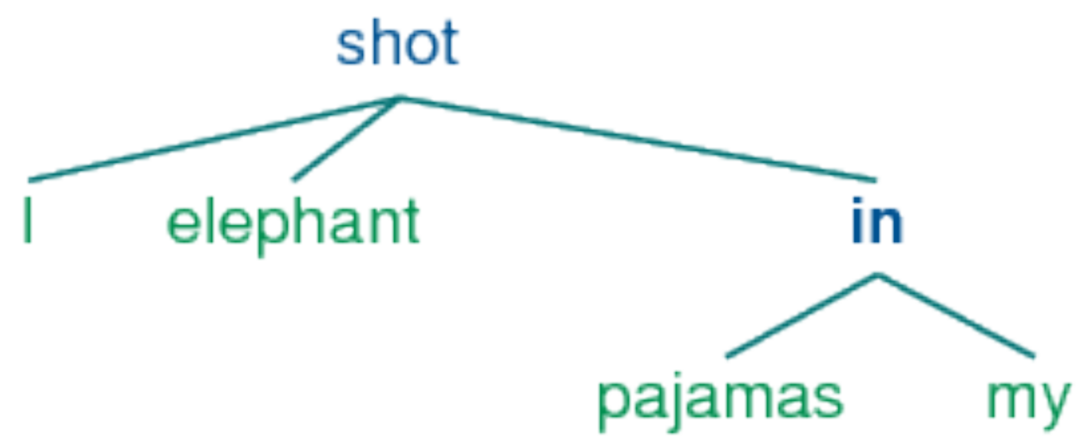
- complex scenes

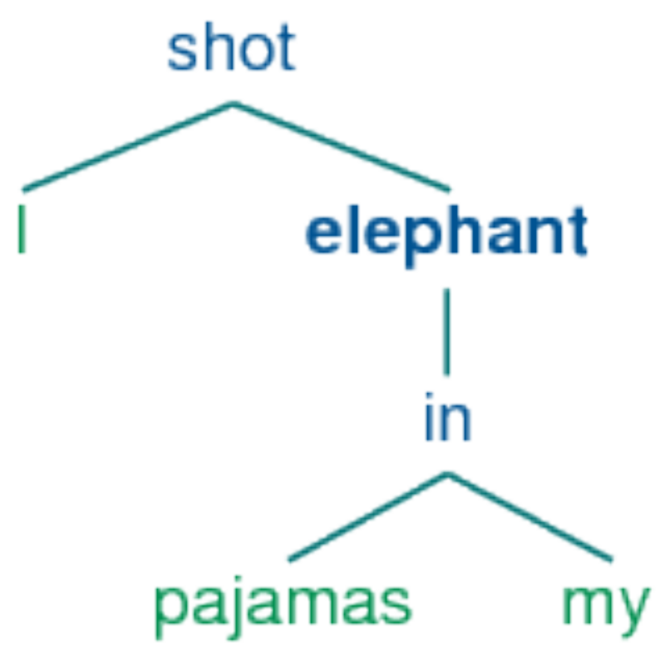
- occlusions

Failures - Ambiguity

- Natural language processing:

  "I shot an elephant in my pajamas" (Groucho Marx)





- and everywhere... (Kulesza et al 07, Finley et al 08)

Our Approach

- Inference & Learning with Random Maximum A-Posteriori Perturbations

Inference and Learning

$$x^* = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \ \{\phi(x)\}$$

prediction — possible structures — scores

- Probabilistic predictions (e.g., Gibbs' distribution) over structures

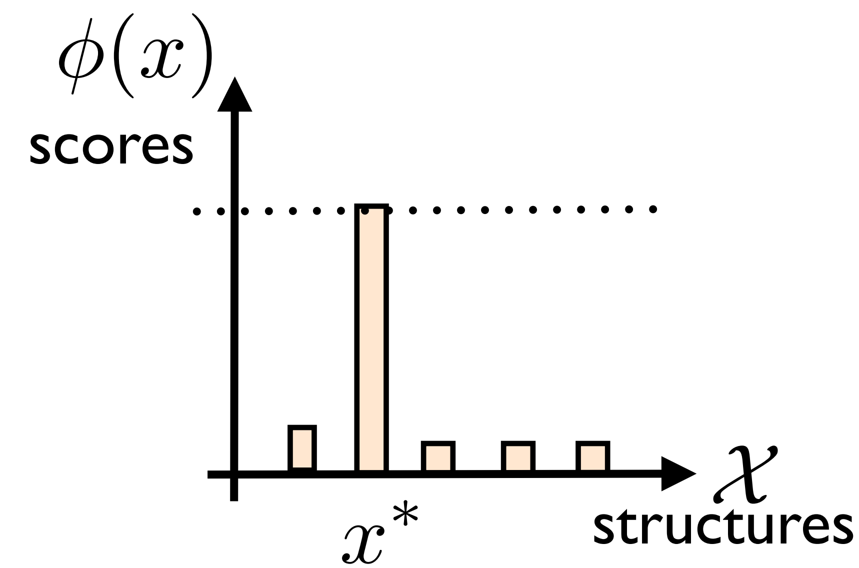
$$p(x) = \frac{1}{Z} \exp(\phi(x))$$

- partition function

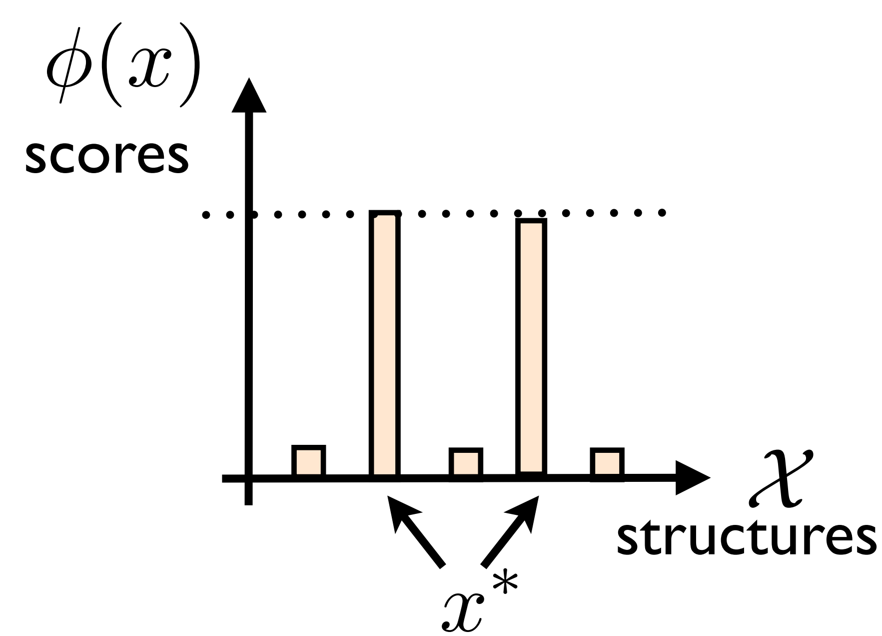
$$Z = \sum_{x \in \mathcal{X}} \exp(\phi(x))$$

  - Often hard, even when the max is easy

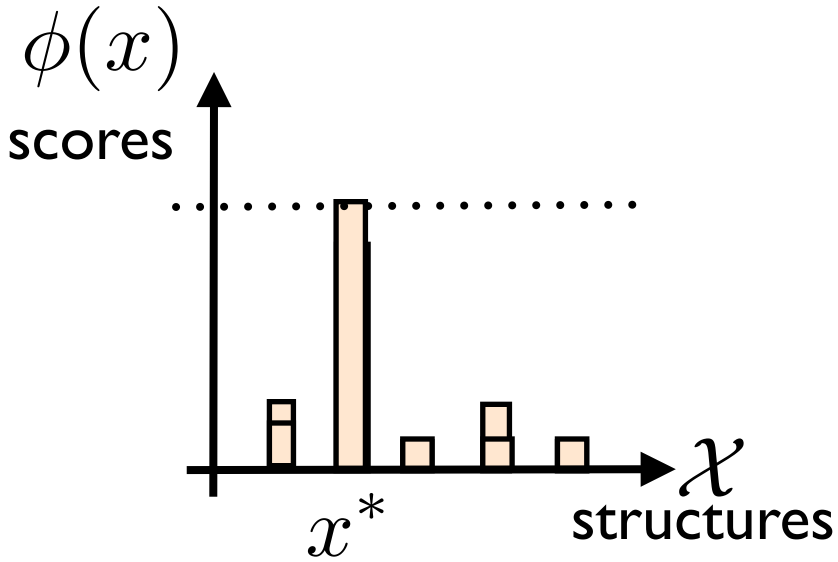
- Success: dominant solution



- Failures: multiple high scoring alternatives



Random Maximum A-Posteriori



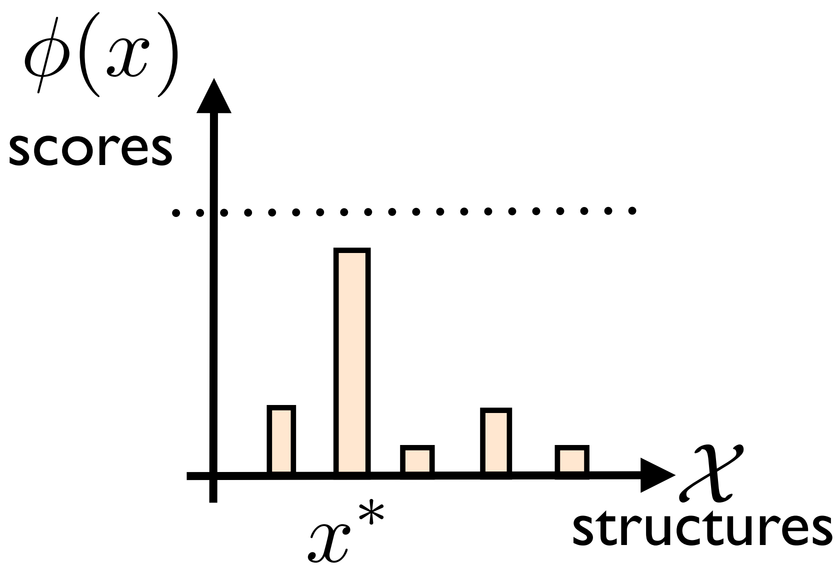
$$x^* = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \; \{\phi(x) + \gamma(x)\}$$

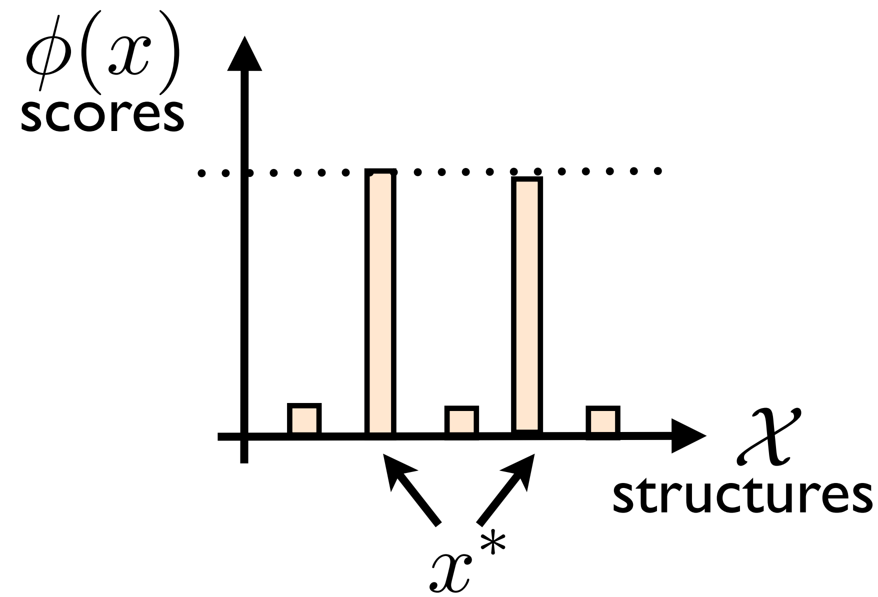
prediction

possible structures

perturbed scores

Random Maximum A-Posteriori

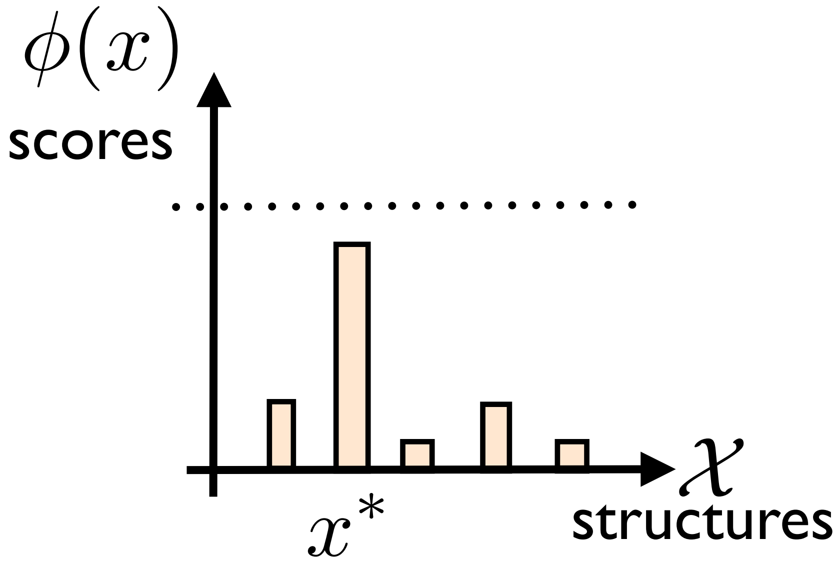


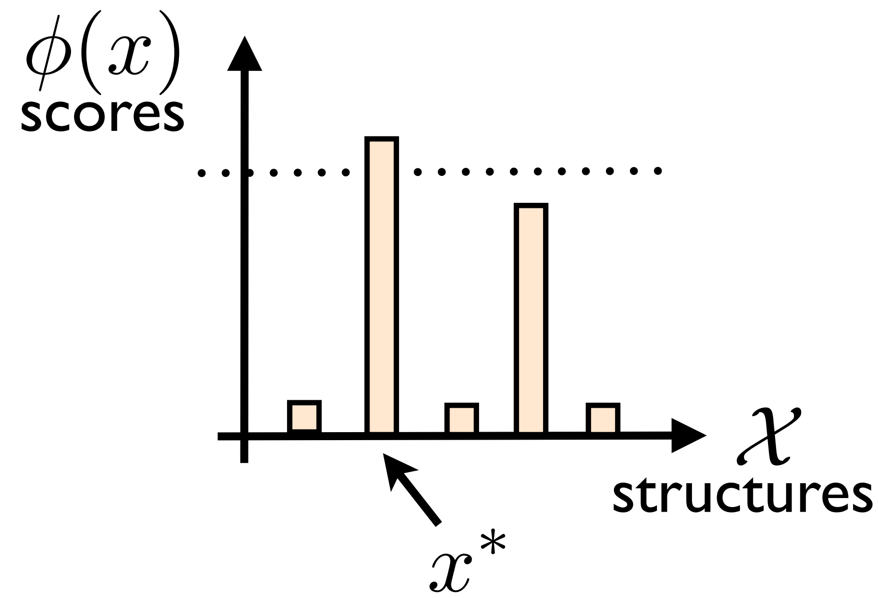


$$x^* = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \; \{\phi(x) + \gamma(x)\}$$

prediction — possible structures — perturbed scores

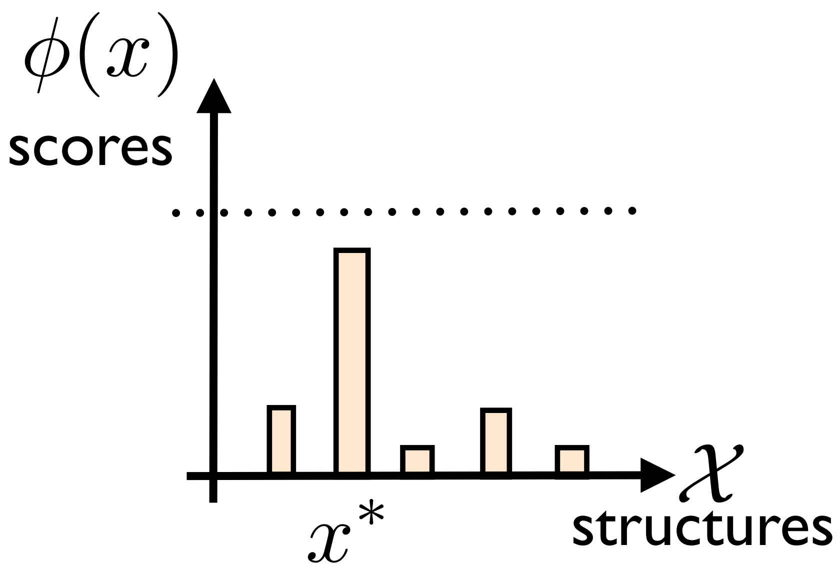
Random Maximum A-Posteriori

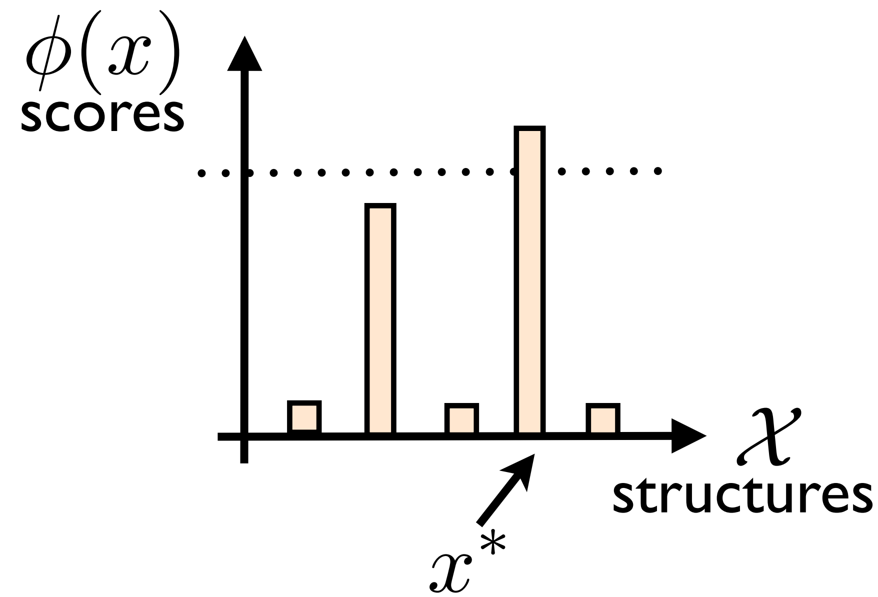




$$\underbrace{x^*}_{\text{prediction}} = \underbrace{\operatorname*{argmax}_{x \in \mathcal{X}}}_{\text{possible structures}} \underbrace{\{\phi(x) + \gamma(x)\}}_{\text{perturbed scores}}$$

Random Maximum A-Posteriori





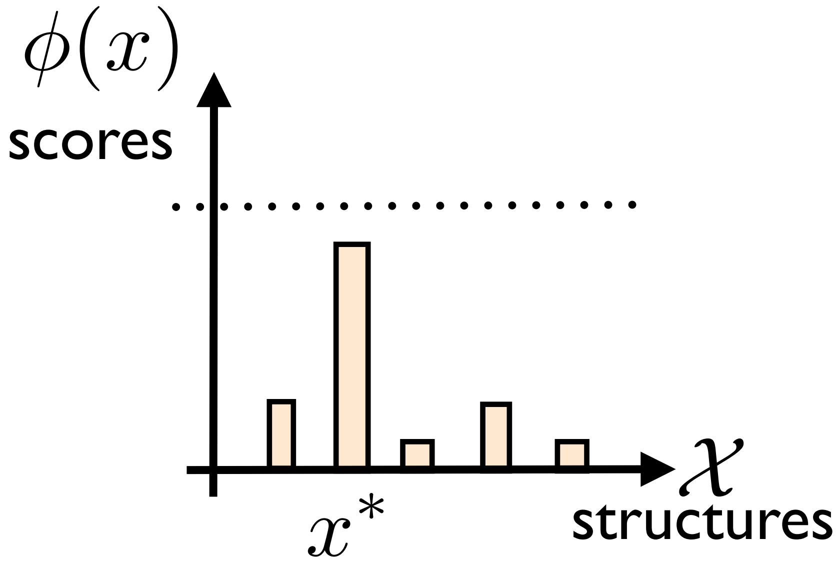
$$x^* = \operatorname*{argmax}_{x\in\mathcal{X}} \{\phi(x)+\gamma(x)\}$$

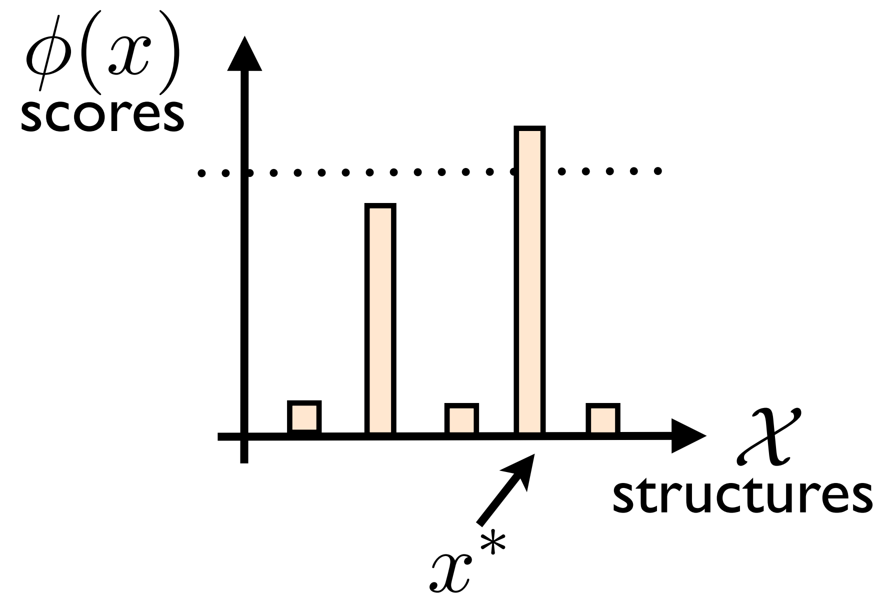
prediction

possible structures

perturbed scores

Random Maximum A-Posteriori





$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} \; \{\phi(x) + \gamma(x)\}$$

prediction

possible structures

perturbed scores

- **Theorem:** There is a distribution over perturbations $\gamma(x)$

$$P_\gamma\left[x^* = \arg\max_{x\in\mathcal{X}}\{\phi(x)+\gamma(x)\}\right] = \frac{1}{Z}\exp(\phi(x^*))$$

(cf. Papandreou & Yuille 2011, Tarlow & Adams & Zemel 2012)

Why the Partition Function?

$$\log \sum_{x \in \mathcal{X}} \exp(\phi(x))$$

partition function

gradient to statistics

$\longleftrightarrow$

$$P(x) = \frac{1}{Z} \exp(\phi(x))$$

Gibbs' distribution

Max-Statistics

- **Lemma:**

Let $\gamma(x)$ be i.i.d with Gumbel distribution with zero mean

$$F(t) \stackrel{def}{=} P[\gamma(x) \leq t] = \exp(-\exp(-t))$$

then the random MAP perturbation

$$\max_{x \in \mathcal{X}} \{\phi(x) + \gamma(x)\}$$

has Gumbel distribution whose mean is $\log Z$

- **Proof:** $P[\max_{x \in \mathcal{X}} \{\phi(x) + \gamma(x)\} \leq t] = \prod_{x \in \mathcal{X}} F(t - \phi(x)) =$

$$\exp(-\sum_{x \in \mathcal{X}} \exp(-(t - \phi(x)))) = \exp(-\exp(-t)Z) = F(t - \log Z)$$

Random MAP Perturbations

- (Hazan and Jaakkola 2012)
- **Theorem (low dimension perturbations):**

Let $\gamma_i(x_i)$ be i.i.d with Gumbel distribution. Then

$$\log Z = E_{\gamma_1(x_1)} \max_{x_1} \cdots E_{\gamma_n(x_n)} \max_{x_n} \{\phi(x) + \sum_{i=1}^{n} \gamma_i(x_i)\}$$

- **Proof:** $$Z = \sum_{x_1} \cdots \sum_{x_n} \exp(\phi(x))$$

and previous theorem implies $$E_{\gamma_i(x_i)} \max_{x_i} \Longleftrightarrow \sum_{x_i}$$

Upper Bounds

- **Corollary:**

Let $\gamma_i(x_i)$ be i.i.d with Gumbel distribution. Then

$$\log Z \leq E_\gamma \Big[\max_{x_1,...,x_n} \{\phi(x) + \sum_{i=1}^n \gamma_i(x_i)\} \Big]$$

- **Proof:**

$$\log Z = E_{\gamma_1(x_1)} \max_{x_1} \cdots E_{\gamma_n(x_n)} \max_{x_n} \{\phi(x) + \sum_{i=1}^n \gamma_i(x_i)\}$$

Move maximizations inside

- Related work (Counting): $x_i \in \{0,1\}, \ \phi(x) \in \{-\infty, 0\}$
  - Talagrand 94: Bounds on canonical processes. Laplace distribution
  - Barvinok & Samorodnitsky 07: Approximate counting. Logistic distribution

Lower Bounds

- **Corollary:**

Let $\gamma_i(x_i)$ be i.i.d with Gumbel distribution. Then

$$\log Z \geq E_\gamma\Big[\max_{x_1,...,x_n}\{\phi(x) + \gamma_i(x_i)\}\Big]$$

- **Proof:**

$$\log Z = E_{\gamma_1(x_1)} \max_{x_1} \cdots E_{\gamma_n(x_n)} \max_{x_n}\{\phi(x) + \sum_{i=1}^{n} \gamma_i(x_i)\}$$

Move expectation inside, while $E_\gamma[\gamma_i(x_i)] = 0$

Results (Upper bounds & Approx)

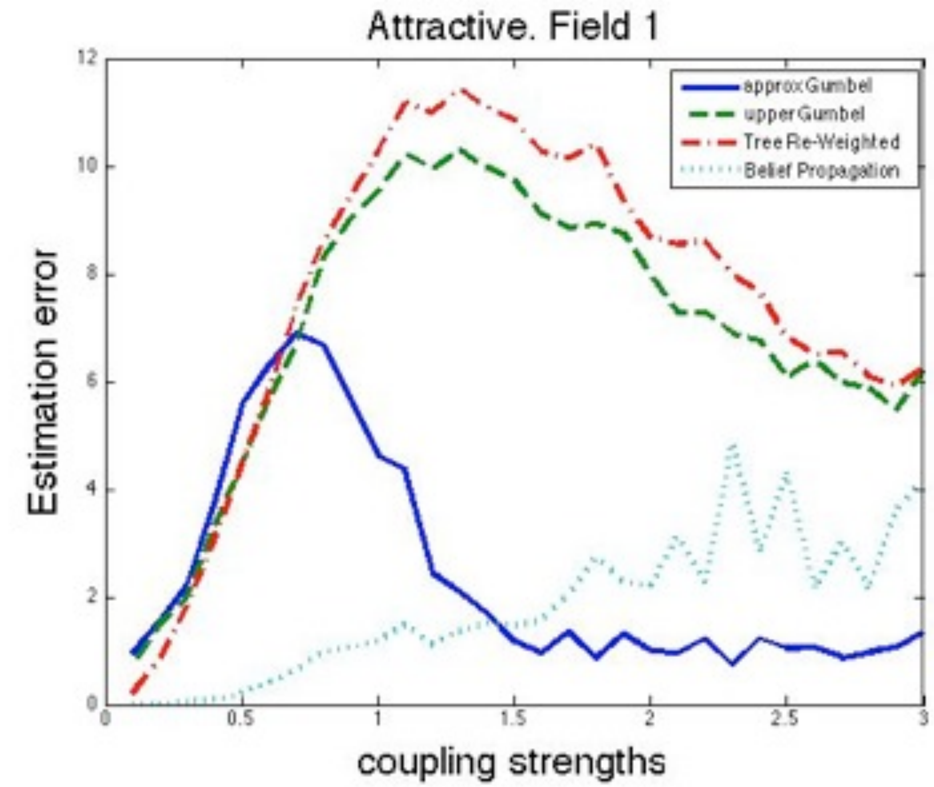
- Spin glass, 10x10 grid

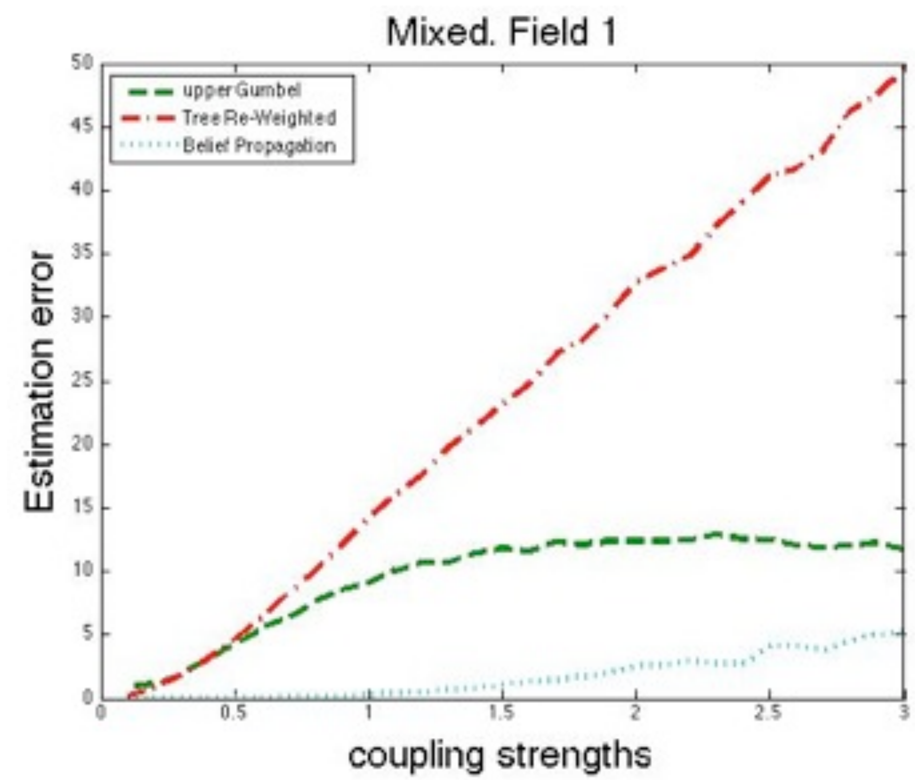
$$\sum_i w_i \phi_i(x_i) + \sum_{i,j} w_{i,j} \phi_{i,j}(x_i, x_j)$$

- $\phi_i(x_i) = x_i, \quad x_i \in \{-1, 1\}$

- $\phi_{i,j}(x_i, x_j) = x_i x_j$

- Field w_i
- attractive $w_{i,j} \geq 0$. Graph-cuts.
- mixed $w_{i,j} \lessgtr 0$. MPLP.

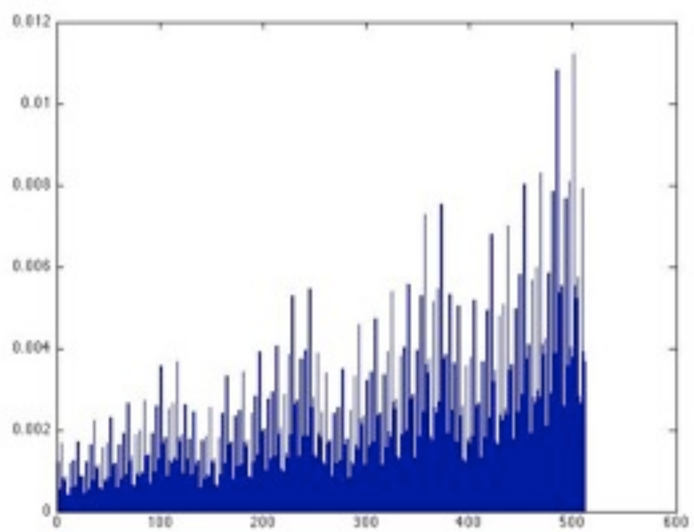




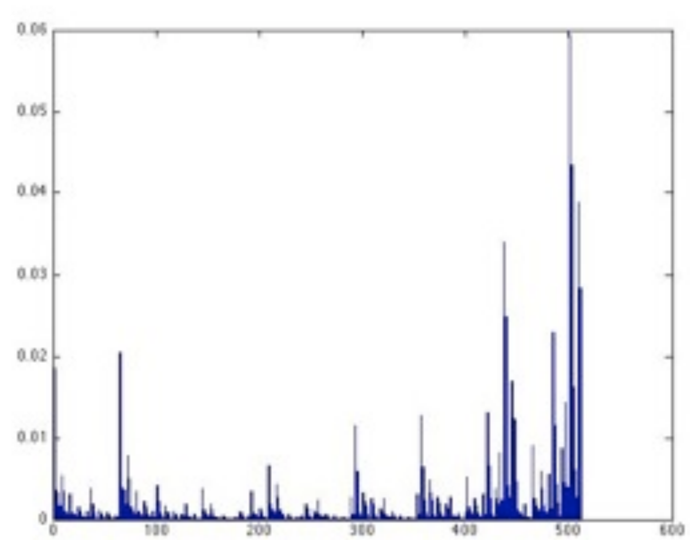
When it works? The "hi-domain"

$$p(x) \propto \exp\left(\sum_i w_i \phi_i(x_i) + \sum_{i,j} w_{i,j} \phi_{i,j}(x_i, x_j)\right)$$

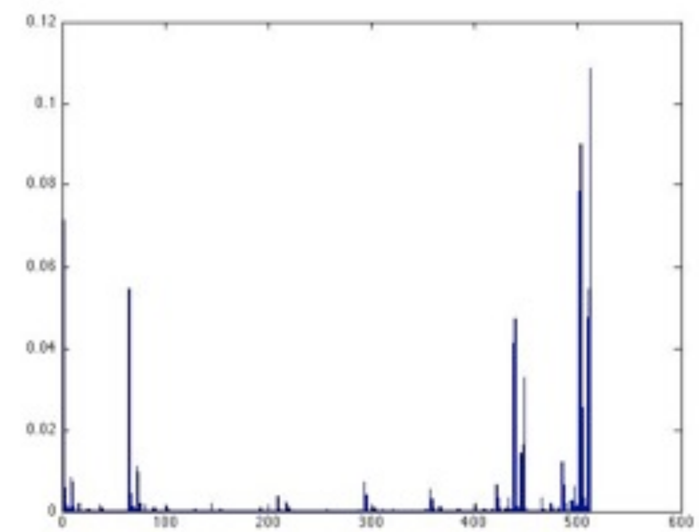
- $w_i = 1, w_{i,j} = 0$



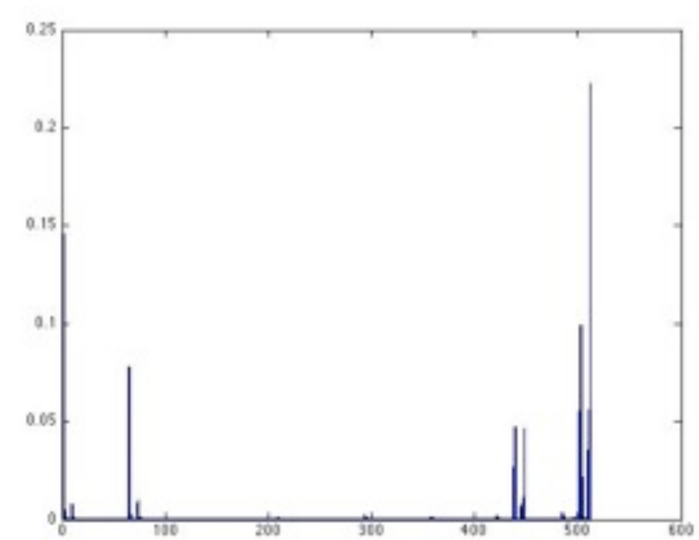
- $w_i = 1, w_{i,j} \in [-1, 1]$



- $w_i = 1, w_{i,j} \in [-2, 2]$



- $w_i = 1, w_{i,j} \in [-3, 3]$



Inference and Learning

$$\log \sum_{x \in \mathcal{X}} \exp(\phi(x))$$

partition function

gradient to statistics

$$P(x) = \frac{1}{Z} \exp(\phi(x))$$

Gibbs' distribution

- hard to compute, even if the max is easy

Inference and Learning

$$E_\gamma\left[\max_{x_1,\ldots,x_n}\{\phi(x)+\sum_{i=1}^{n}\gamma_i(x_i)\}\right] \quad \xleftrightarrow{\text{gradient to statistics}} \quad P\left[x \in \operatorname*{argmax}_{\hat{x}_1,\ldots,\hat{x}_n}\{\phi(\hat{x})+\sum_i \gamma_i(\hat{x}_i)\}\right]$$

surrogate partition induced distribution

- (Hazan and Jaakkola 2012)
- Unbiased sampling is efficient.
- These models were introduced in (Keshet, McAllester, Hazan 2011, Papandreau, Yuille 2011, Tarlow, Adams, Zemel 2012).

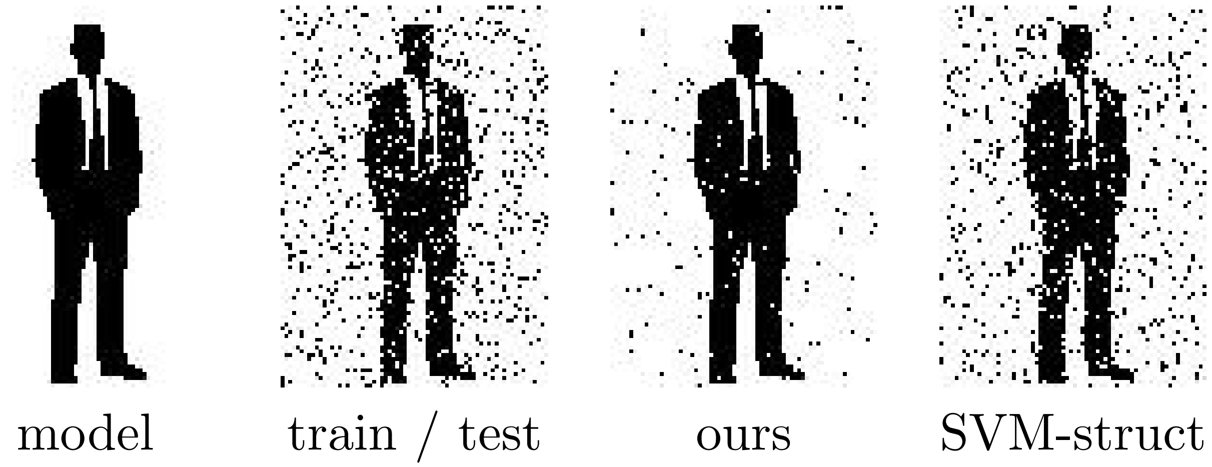
Learning with Likelihood

- Learning spin glass parameters

$$\sum_i w_i \phi_i(x_i) + \sum_{i,j} w_{i,j} \phi_{i,j}(x_i, x_j)$$

- $(x_1, \ldots, x_n)$ are binary pixel values of 70x100 image + 10% noise

- Surrogate partition + MPLP



Ours	SVM-struct
2%	8%

Learning with Loss Minimization

- Learning measured by loss

$$loss(w,x) \stackrel{def}{=} \sum_{\hat{x}} p(\hat{x}|w) loss(\hat{x},x)$$

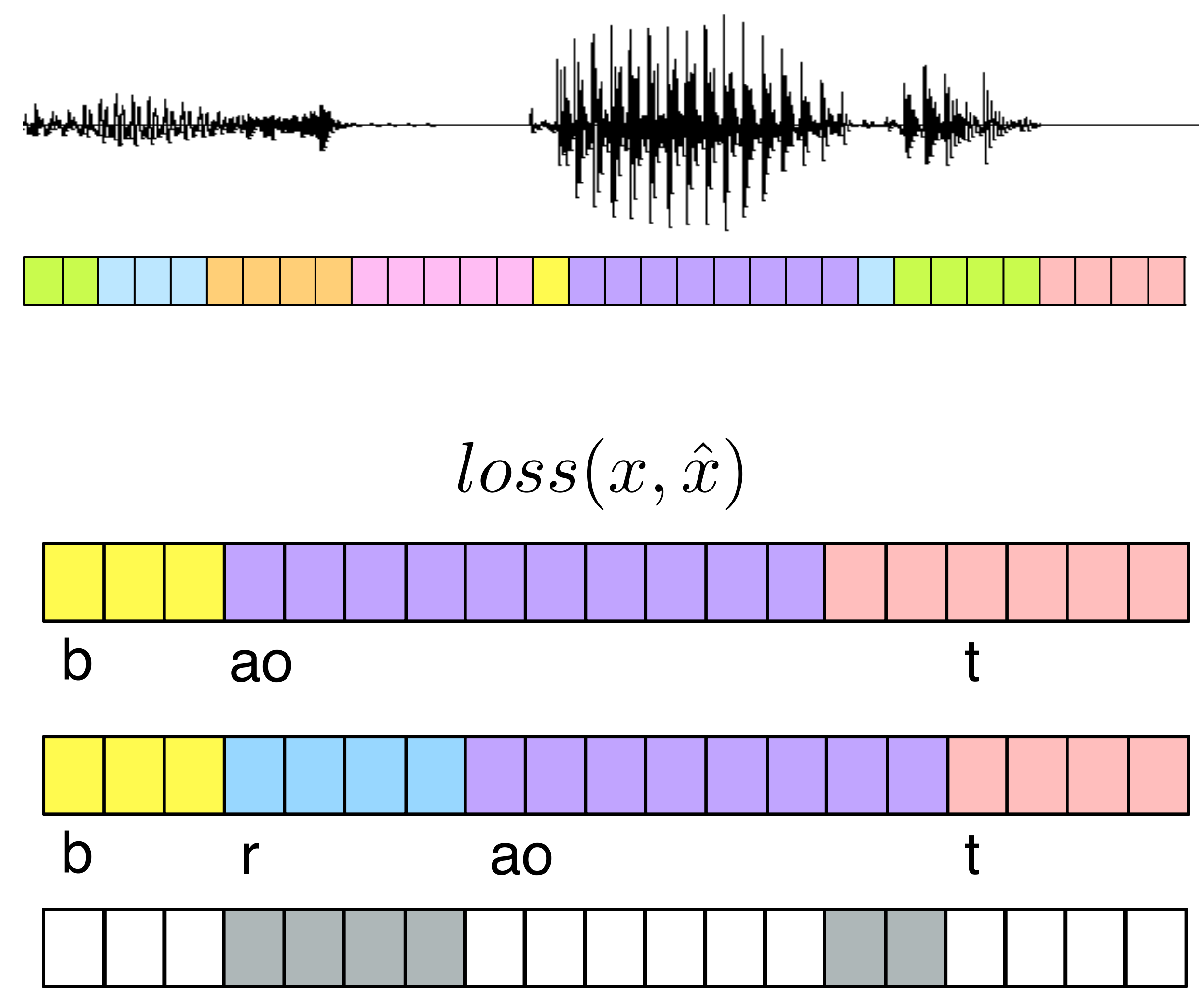
- Perturbed MAP predictions give uniform generalization bounds

$$P_\gamma \left[\hat{x} \in argmax_{x'} \{ (w+\gamma)^\top \phi(x') \} \right]$$

- Theorem: $\forall$ w simultaneously

$$E_{x\sim D} loss(w,x) \leq \frac{2}{|S|} \sum_{x \in S} loss(w,x) + \frac{1}{m-1} \left(\|w\|^2 + 2\log(m/\delta) \right)$$

(Keshet, McAllester, Hazan 2011)



	Ours	SVM-struct
TIMIT	28.6%	30.2%

Our Approach

- Inference & Learning with Random Maximum A-Posteriori Perturbations

Thank You

Panel Discussion

- Compare learning rules:
  - log-likelihood
  - max-margin
  - herding
  - loss minimization?
  - others?

- Optimization and statistics point of views

Panel Discussion

- Why does dropout works?

- Are there other regularization schemes that involve the injection of noise that should be equally effective?

- Can dropout be explained using known perturbation learning techniques (e.g., robust learning / PAC-Bayes?)

Panel Discussion

- Agree or disagree: The Gibbs distribution is special.

- What do we gain in exchange for the hard computation that go into the Gibbs distribution / partition function?

Panel Discussion

From Vincent's abstract: "I will be going back and forth between stochastic perturbations and related deterministic analytic criteria, which I hope may spawn interesting discussions on the interface between, and merits of, both these outlooks."

Are there benefits to thinking in terms of stochastic perturbations versus deterministic analytic criteria? In what cases are they equivalent? Are there cases where one works but the other does not?

Panel Discussion

- Robust optimization versus stochastic perturbations?

Panel Discussion

- We know there is a close relationship between the Gibbs distribution and Perturb & MAP models. We also know there is a close relationship between Perturb & MAP and regularization via PAC Bayes. Can we then view the Gibbs distribution in regularization terms?

Panel Discussion

- Approximate methods?

Panel Discussion

- Applications? vision, NLP, information retrieval

Panel Discussion

- Where do the ideas at the center of this workshop have their historical roots?

Panel Discussion - Question?