Structured Prediction: From Gaussian Perturbations to Linear-Time Principled Algorithms

Jean Honorio, Tommi Jaakkola

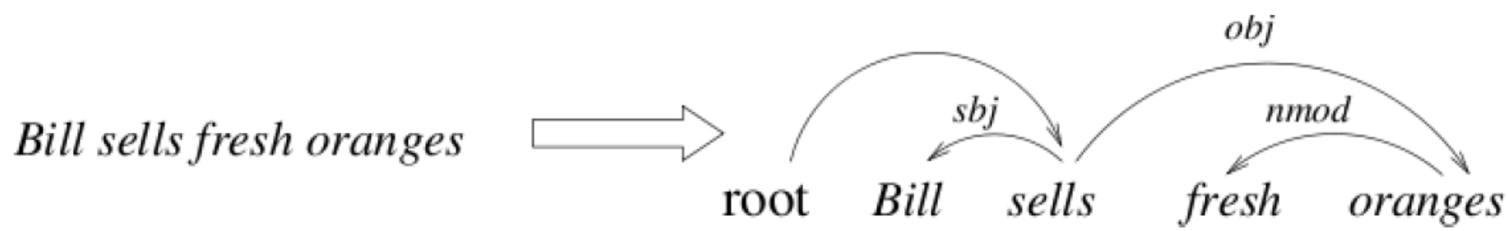
PURDUE UNIVERSITY

MIT CSAIL

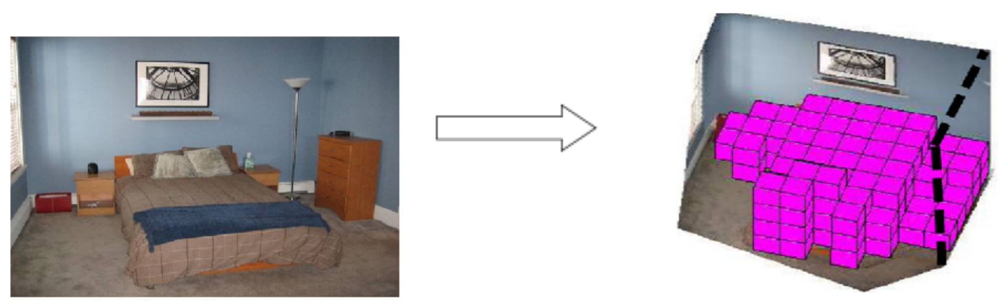
Structured Prediction

(credits: Ivan Titov)

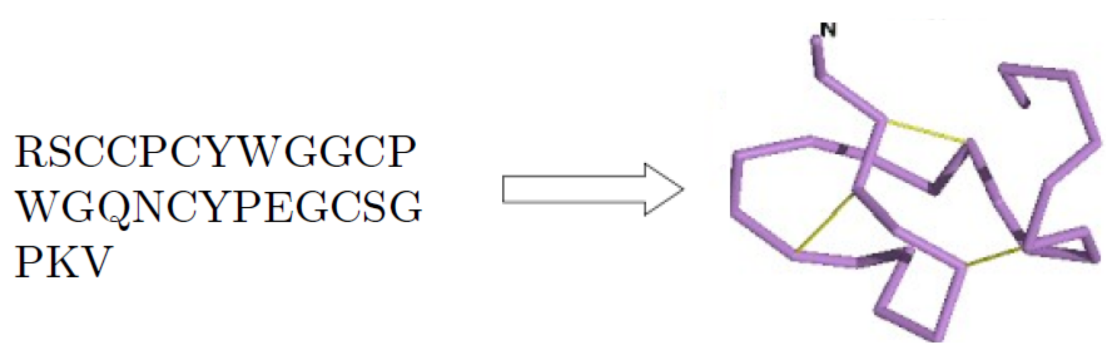
- Syntactic/dependency parsing



- 3D layout prediction



- Protein structure prediction

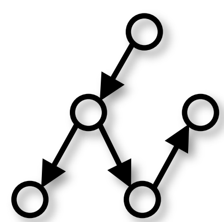
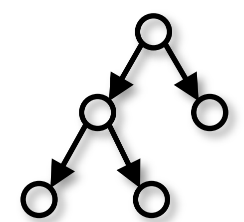


Structured Prediction

- Observed input $x \in \mathcal{X}$
- Latent structured output $y \in \mathcal{Y}$
- n training samples from a distribution D

$(x, y) \sim D$

$S \sim D^n$

S = ("Bill sells fresh oranges" , ), ("the cat is white" , ), …

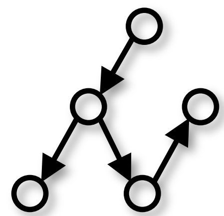
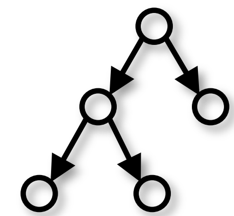
- Countable set of feasible decodings $\mathcal{Y}(x) \neq \emptyset$

Structured Prediction

- Observed input $x \in \mathcal{X}$
- Latent structured output $y \in \mathcal{Y}$
- n training samples from a distribution D

$(x, y) \sim D$

$S \sim D^n$

S = ("Bill sells fresh oranges" , ), ("the cat is white" , ), …

- Countable set of feasible decodings $\mathcal{Y}(x) \neq \emptyset$
- ***Linear decoder***

$$f_w(x) \equiv \arg\max_{y \in \mathcal{Y}(x)} \phi(x, y) \cdot w$$

$\phi(x, y) \in \mathbb{R}^k$
feature vector

$w \in \mathcal{W} \subseteq \mathbb{R}^k \setminus \{0\}$
parameter vector

Structured Prediction

- Distortion function

$$d : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$$

- ***To learn parameter*** w

$$\min_{w \in \mathcal{W}} \mathbb{E}_{(x,y)\sim D} [d(y, f_w(x))]$$

needs access to D

discontinuous w.r.t. w

  - Statistically inefficient
  - Computationally inefficient

Structured Prediction

- Distortion function

$$d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$$

- To learn parameter w

$$\min_{w \in \mathcal{W}} \mathop{\mathbb{E}}_{(x,y)\sim D} [d(y, f_w(x))]$$

  - Statistically inefficient
  - Computationally inefficient

- ***To robustly learn parameter*** w
  - Robust objective under ***Gaussian perturbations***
  - Let $Q(w)$ = *N*($w\alpha$, **I**) for $\alpha > 0$

$$\min_{w \in \mathcal{W}} \mathop{\mathbb{E}}_{(x,y)\sim D} \left[\mathop{\mathbb{E}}_{w' \sim Q(w)} [d(y, f_{w'}(x))] \right]$$

Structured Prediction

- Distortion function

$$d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$$

- To learn parameter w

$$\min_{w \in \mathcal{W}} \mathop{\mathbb{E}}_{(x,y) \sim D} [d(y, f_w(x))]$$

  - Statistically inefficient
  - Computationally inefficient

- To robustly learn parameter w
  - Robust objective under ***Gaussian perturbations***
  - Let $Q(w)$ = *N*($w\alpha$, **I**) for $\alpha > 0$

$$\min_{w \in \mathcal{W}} \boxed{\mathop{\mathbb{E}}_{(x,y) \sim D} \left[\mathop{\mathbb{E}}_{w' \sim Q(w)} [d(y, f_{w'}(x))] \right]}$$

We provide upper bounds for this ***Gibbs distortion***

Structured Prediction

- ***Margin*** $m(x, y, y', w) \equiv \phi(x, y) \cdot w - \phi(x, y') \cdot w$
- $c(p, x, y)$ # times that a part $p \in \mathcal{P}$ appears in (x, y)
  - use as features $\phi_p(x, y) \equiv c(p, x, y)$
  - set of ***active*** parts $\mathcal{P}(x) = \{p \mid (\exists y)\ c(p, x, y) > 0\}$
- ***Hamming distance***

$$H(x, y, y') \equiv \sum_{p \in \mathcal{P}(x)} |c(p, x, y) - c(p, x, y')|$$

Using Randomness

- To learn parameter w **(max all)**

$$\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{(x,y) \in S} \max_{\hat{y} \in \mathcal{Y}(x)} d(y, \hat{y}) \, 1 \begin{pmatrix} H(x, y, \hat{y}) \\ -m(x, y, \hat{y}, w) \geq 0 \end{pmatrix} + \lambda \|w\|_2^2$$

- ***To learn parameter*** w **(max random, Zhang'14)**

$$\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{(x,y) \in S} \max_{\hat{y} \in T(w,x)} d(y, \hat{y}) \, 1 \begin{pmatrix} H(x, y, \hat{y}) \\ -m(x, y, \hat{y}, w) \geq 0 \end{pmatrix} + \lambda \|w\|_2^2$$

set $T(w, x)$ of ***random outputs***, i.i.d. from proposal $R(w, x)$

Using Randomness

- To learn parameter w **(max all)**

$$\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{(x,y) \in S} \max_{\hat{y} \in \mathcal{Y}(x)} d(y, \hat{y})\, 1 \begin{pmatrix} H(x, y, \hat{y}) \\ -m(x, y, \hat{y}, w) \geq 0 \end{pmatrix} + \lambda \|w\|_2^2$$

- To learn parameter w **(max random)**

$$\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{(x,y) \in S} \max_{\hat{y} \in T(w,x)} d(y, \hat{y})\, 1 \begin{pmatrix} H(x, y, \hat{y}) \\ -m(x, y, \hat{y}, w) \geq 0 \end{pmatrix} + \lambda \|w\|_2^2$$

set $T(w, x)$ of ***random outputs***, i.i.d. from proposal $R(w, x)$

- ***Gibbs distortion ≤ max random ≤ max all***

We show this — This is obvious

Using Randomness

- To learn parameter w **(max all)**

$$\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{(x,y) \in S} \max_{\hat{y} \in \mathcal{Y}(x)} d(y, \hat{y})\, 1 \begin{pmatrix} H(x, y, \hat{y}) \\ -m(x, y, \hat{y}, w) \geq 0 \end{pmatrix} + \lambda \|w\|_2^2$$

- To learn parameter w **(max randon**

$$\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{(x,y) \in S} \max_{\hat{y} \in T(w,x)} d(y, \hat{y})\, 1 \begin{pmatrix} H(x, y, \hat{y}) \\ -m(x, y, \hat{y} \end{pmatrix}$$

set $T(w, x)$ of ***random outputs***, i.i.d. from

Counterintuitive for optimization: minimizing a lower bound

- ***Gibbs distortion ≤ max random ≤ max all***

We show this

This is obvious

Using Randomness

- ***To learn parameter*** w **(max random)**

$$\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{(x,y)\in S} \max_{\hat{y} \in T(w,x)} d(y,\hat{y})\, 1 \begin{pmatrix} H(x,y,\hat{y}) \\ -m(x,y,\hat{y},w) \geq 0 \end{pmatrix} + \lambda \|w\|_2^2$$

set $T(w,x)$ of ***random outputs***, i.i.d. from proposal $R(w,x)$

Procedure for sampling $y' \sim R(w,x)$

Input: $w \in \mathcal{W}$, $x \in \mathcal{X}$

Initialize uniformly at random $\hat{y} \in \mathcal{Y}(x)$

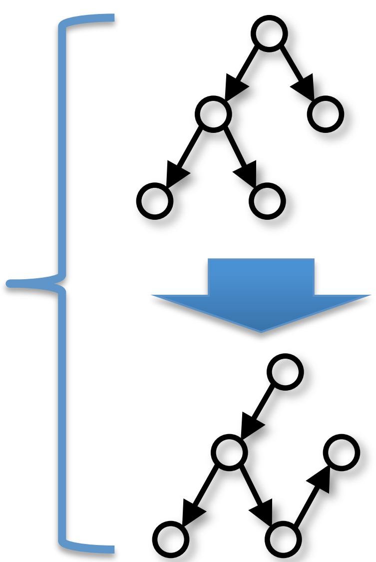
repeat

Make a local change to $\hat{y}$

in order to increase $\phi(x,\hat{y}) \cdot w$

until no refinement is possible

Output: $y' \leftarrow \hat{y}$



Using Randomness

- ***Synthetic results***
  - over 30 repetitions, 95% confidence level

Problem	**Method**	**Training runtime**	**Training distortion**	**Test runtime**	**Test distortion**	**Distance to ground truth**	**Angle with ground truth**
Directed spanning trees	All	1000	52% ± 1.1%	12.4 ± 0.4	61% ± 1.8%	0.56 ± 0.004	74° ± 0.3°
	Random	104 ± 3	38% ± 2.1%	2.4 ± 0.1	56% ± 1.9%	0.51 ± 0.005	49° ± 0.6°
	Random/All			12.4 ± 0.3	56% ± 1.9%		
Directed acyclic graphs	All	1000	41% ± 1.2%	10.8 ± 0.2	45% ± 1.5%	0.60 ± 0.020	61° ± 1.0°
	Random	386 ± 21	30% ± 1.3%	8.5 ± 0.2	39% ± 1.6%	0.40 ± 0.008	37° ± 1.0°
	Random/All			10.8 ± 0.2	39% ± 1.6%		
Cardinality constrained sets	All	1000	42% ± 1.4%	11.1 ± 0.4	45% ± 1.8%	0.58 ± 0.011	65° ± 0.6°
	Random	272 ± 9	21% ± 1.2%	6.0 ± 0.2	30% ± 1.9%	0.44 ± 0.008	30° ± 0.8°
	Random/All			10.9 ± 0.3	29% ± 2.1%		

- ***Real-world datasets***
  - See (Zhang'14, Zhang'15) for natural language processing

Prior Generalization Result

- **(McAllester'07)** assumes bounded ***active*** parts

$$|\cup_{(x,y)\in S}\mathcal{P}(x)| \le \ell$$

- Let $\delta \in (0,1)$, with probability at least $1-\delta/2$ over the choice of n training samples, for all $w \in \mathcal{W}$ and perturbations $Q(w)$ = $N(w\sqrt{2\log(2n\ell/\|w\|_2^2)}, I)$:

$$\mathop{\mathbb{E}}_{(x,y)\sim D}\left[\mathop{\mathbb{E}}_{w'\sim Q(w)}[d(y,f_{w'}(x))]\right]$$

$$\le \frac{1}{n}\sum_{(x,y)\in S}\max_{\hat{y}\in\mathcal{Y}(x)} d(y,\hat{y})\,1\begin{pmatrix}H(x,y,\hat{y})\\ -m(x,y,\hat{y},w)\ge 0\end{pmatrix}$$

$$+\frac{\|w\|_2^2}{n}+\sqrt{\frac{\|w\|_2^2\log(2n\ell/\|w\|_2^2)+\log(2n/\delta)}{2(n-1)}}$$

Gibbs distortion

Gaussian concentration inequalities

PAC-Bayes theorem (KL divergence)

Prior Generalization Result

- **(McAllester'07)** assumes bounded active parts

$$|\cup_{(x,y)\in S}\mathcal{P}(x)| \leq \ell$$

- Let $\delta \in (0,1)$, with probability at least $1-\delta/2$ over the choice of n training samples, for all $w \in \mathcal{W}$ and perturbations $Q(w)$ = N($w\sqrt{2\log(2n\ell/\|w\|_2^2)}$,I):

$$\mathop{\mathbb{E}}_{(x,y)\sim D}\left[\mathop{\mathbb{E}}_{w'\sim Q(w)}[d(y,f_{w'}(x))]\right]$$

$$\leq \frac{1}{n}\sum_{(x,y)\in S} \max_{\hat{y} \ldots (x)} \ldots (y,\hat{y})\, 1\left(\ldots -m(x,y,\hat{y},w) \geq 0\right)$$

$$+\frac{\|w\|_2^2}{n} + \sqrt{\frac{\|w\|_2^2\log(2n\ell/\|w\|_2^2)+\log(2n/\delta)}{2(n-1)}}$$

Gibbs distortion

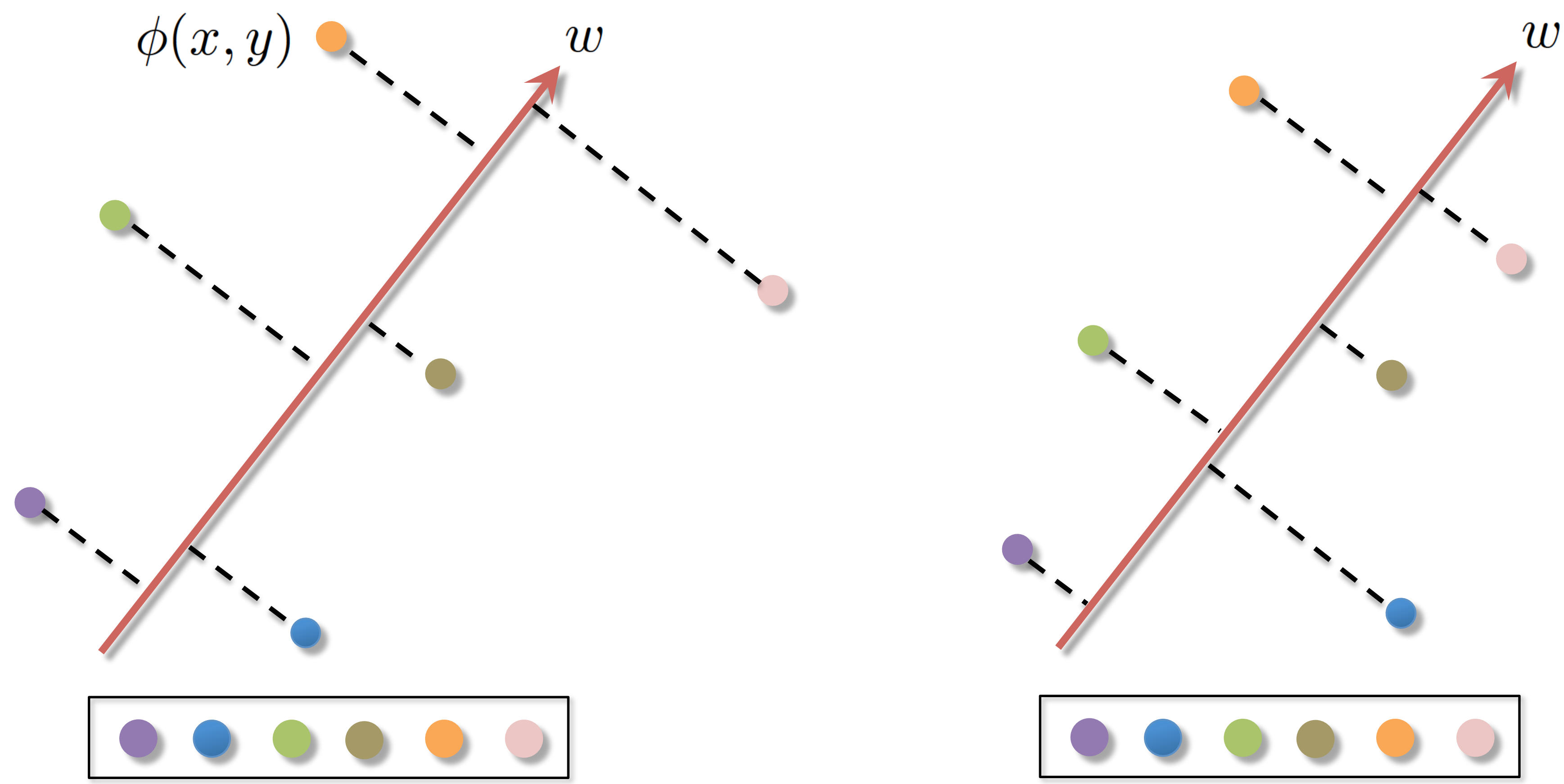
Gaussian concentration inequalities

PAC-Bayes theorem (KL divergence)

But now randomness comes from S and $T(w,x)$ for all $(x,y)\in S$.

Assumption: Linearly Inducible Ordering

- ***How many*** $R(w, x)$ **?**
- For **(max random)** the exact value of $\phi(x, y) \cdot w$ is unimportant, only their ***induced linear ordering***
  - These 2 examples lead to the same proposal $R(w, x)$



Assumption: Linearly Inducible Ordering

- Linear ordering induced by $w \in \mathcal{W}$ and $\phi(x, \cdot)$

  $r(x) \equiv |\mathcal{Y}(x)|$ $\mathcal{Y}(x) \equiv \{y_1 \dots y_{r(x)}\}$

  linear ordering $\phi(x, y_{\pi_1}) \cdot w < \cdots < \phi(x, y_{\pi_{r(x)}}) \cdot w$

  induces a permutation $\pi(x) = (\pi_1 \dots \pi_{r(x)})$ of $\{1 \dots r(x)\}$

  if $\pi(x) = \pi'(x)$ then $KL(R(w, x) \| R(w', x)) = 0$

- ***How many*** $R(w, x)$ **?**
  - Note that $w, \phi(x, y) \in \mathbb{R}^\ell$
  - Assume $|\mathcal{Y}(x)| \leq r$ and w being $\mathfrak{s}$-sparse
  - n training samples, then nr points in $\mathbb{R}^\ell$
  - At most $(nr)^{2\mathfrak{s}}$ orderings (Bennett'56) for $\mathfrak{s}$ fixed features
  - At most $\binom{\ell}{\mathfrak{s}}(nr)^{2\mathfrak{s}}$ proposals $R(w, x)$

Assumption: Maximal Distortion

- There exist a value $\beta \in [0, 1)$ such that for all $(x, y) \in S$ and $w \in \mathcal{W}$:

$$\mathop{\mathbb{P}}_{y' \sim R(w,x)} [d(y, y') = 1] \geq 1 - \beta$$

- ***Examples:***
  - binary distortion $d(y, y') = 1\,(y \neq y')$, arbitrary $\mathcal{Y}(x)$: $\beta = 1/2$
  - $d(y, y')$ = number of different edges/elements
    - $\mathcal{Y}(x)$ directed spanning trees of v nodes: $\beta = \frac{v-2}{v-1}$
    - $\mathcal{Y}(x)$ directed acyclic graphs of v nodes, and b parents per node: $\beta = \frac{b^2+2b+2}{b^2+3b+2}$
    - $\mathcal{Y}(x)$ sets of b elements from v: $\beta = 1/2$

Our Generalization Result

- Let $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of n training samples and n sets of random structured outputs, for all $w \in \mathcal{W}$, perturbations $Q(w)$ and for sets with $|T(w,x)| = \left\lceil \frac{1}{2} \max\left(\frac{1}{\log(1/\beta)}, 32\|w\|_2^2\right) \log n \right\rceil$

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim D}\left[\mathop{\mathbb{E}}_{w'\sim Q(w)}[d(y, f_{w'}(x))]\right] \\
&\le \frac{1}{n}\sum_{(x,y)\in S} \max_{\hat{y}\in T(w,x)} d(y,\hat{y})\, 1\begin{pmatrix} H(x,y,\hat{y}) \\ -m(x,y,\hat{y},w) \ge 0 \end{pmatrix} \\
&+ \frac{\|w\|_2^2}{n} + \sqrt{\frac{\|w\|_2^2 \log(2n\ell/\|w\|_2^2) + \log(2n/\delta)}{2(n-1)}} + \sqrt{\frac{1}{n}} \\
&+ \max\left(\frac{1}{\log(1/\beta)}, 32\|w\|_2^2\right)\sqrt{\frac{\mathfrak{s}\log(\ell+1)\log^3(n+1)}{n}} \\
&+ 3\sqrt{\frac{\mathfrak{s}(\log\ell + 2\log(nr)) + \log(4/\delta)}{n}}
\end{aligned}
$$

Gibbs distortion

deterministic quantity

empirical Rademacher complexity

linear orderings

uniform convergence

Concluding Remarks

- ***Gibbs distortion ≤ max random ≤ max all***
  - Using randomness is a principled and better way!

- Future work:
  - Non-Gaussian perturbations
  - Latent models **(Ping'14, Yu'09)**
  - Maximum a-posteriori perturbation models **(Gane'14, Papandreou'11)**
  - Approximate inference

$$\widetilde{f}_w(x) \equiv \arg\max_{y \in T(w,x)} \phi(x,y) \cdot w$$

Thanks!