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

A DETAILED PROOFS

In this section, we state the proofs of all the theorems and claims in our manuscript.

A.1 Proof of Theorem 1

Here, we provide the proof of Theorem 1. First, we derive an intermediate lemma needed for the final proof.

Lemma 1 (Adapted[3] from Lemma 6 in McAllester, 2007)**.** *Assume that there exists a finite integer value ℓ such that $|\cup_{(x,y)\in S}\,\mathcal{P}(x)| \leq \ell$. Let $Q(w)$ be a unit-variance Gaussian distribution centered at αw for $\alpha = \sqrt{2\log\left(2n\ell/\|w\|_2^2\right)}$. Simultaneously for all $(x,y) \in S$, $y' \in \mathcal{Y}(x)$ and $w \in \mathcal{W}$, we have:*

$$\mathbb{P}_{w'\sim Q(w)}[H(x,y',f_{w'}(x)) - m(x,y',f_{w'}(x),w) < 0] \leq \|w\|_2^2/n$$

or equivalently:

$$\mathbb{P}_{w'\sim Q(w)}[H(x,y',f_{w'}(x)) - m(x,y',f_{w'}(x),w) \geq 0] \geq 1 - \|w\|_2^2/n \tag{7}$$

Proof. First, note that $w' - \alpha w$ is a zero-mean and unit-variance Gaussian random vector. By well-known Gaussian concentration inequalities, for any $p \in \mathcal{P}(x)$ we have:

$$\mathbb{P}_{w'\sim Q(w)}[|w'_p - \alpha w_p| \geq \varepsilon] \leq 2e^{-\varepsilon^2/2}$$

By the union bound and setting $\varepsilon = \alpha = \sqrt{2\log\left(2n\ell/\|w\|_2^2\right)}$, we have:

$$\begin{aligned}\mathbb{P}_{w'\sim Q(w)}[(\exists p \in \cup_{(x,y)\in S}\mathcal{P}(x))\ |w'_p - \alpha w_p| \geq \alpha] &\leq 2|\cup_{(x,y)\in S}\,\mathcal{P}(x)|e^{-\alpha^2/2} \\ &= |\cup_{(x,y)\in S}\,\mathcal{P}(x)|\frac{\|w\|_2^2}{\ell n} \\ &\leq \|w\|_2^2/n\end{aligned}$$

or equivalently:

$$\mathbb{P}_{w'\sim Q(w)}[(\forall p \in \cup_{(x,y)\in S}\mathcal{P}(x))\ |w'_p - \alpha w_p| < \alpha] \geq 1 - \|w\|_2^2/n$$

The high-probability statement in eq.(7) can be written as:

$$\hat{y} = f_{w'}(x) \ \Rightarrow\ H(x,y',\hat{y}) - m(x,y',\hat{y},w) \geq 0$$

Next, we use proof by contradiction, i.e., we will assume:

$$\hat{y} = f_{w'}(x) \text{ and } H(x,y',\hat{y}) - m(x,y',\hat{y},w) < 0$$

[3] We make two small corrections to Lemma 6 of (McAllester, 2007). First, it is only stated for $y' = f_w(x)$ but it does not make use of the optimality of $f_w(x)$, thus, it holds for any $y' \in \mathcal{Y}(x)$. Second, for the union bound over all $p \in \cup_{(x,y)\in S}\mathcal{P}(x)$, we assume that $|\cup_{(x,y)\in S}\,\mathcal{P}(x)| \leq \ell$. Instead, Lemma 6 in (McAllester, 2007) incorrectly assumes $|\mathcal{P}(x)| \leq \ell$ for all $x \in \mathcal{X}$, and thus $|\cup_{(x,y)\in S}\,\mathcal{P}(x)| \leq \sum_{(x,y)\in S}|\mathcal{P}(x)| \leq n\ell$.

and arrive to a contradiction $\hat{y} \neq f_{w'}(x)$. From the above, we have:

$$
\begin{aligned}
m(x, y', \hat{y}, w') &= m(x, y', \hat{y}, \alpha w + (w' - \alpha w)) \\
&= \alpha m(x, y', \hat{y}, w) - (\phi(x, y') - \phi(x, \hat{y})) \cdot (\alpha w - w') \\
&> \alpha H(x, y', \hat{y}) - (\phi(x, y') - \phi(x, \hat{y})) \cdot (\alpha w - w') \\
&= \alpha H(x, y', \hat{y}) - \sum_{p \in \mathcal{P}(x)} (c(p, x, y') - c(p, x, \hat{y}))(\alpha w_p - w'_p) \\
&\geq \alpha H(x, y', \hat{y}) - \sum_{p \in \mathcal{P}(x)} |c(p, x, y') - c(p, x, \hat{y})||\alpha w_p - w'_p| \\
&\geq \alpha H(x, y', \hat{y}) - \sum_{p \in \mathcal{P}(x)} |c(p, x, y') - c(p, x, \hat{y})|\alpha \\
&= 0
\end{aligned}
$$

Note that $m(x, y', \hat{y}, w') > 0$ if and only if $\phi(x, y') \cdot w > \phi(x, \hat{y}) \cdot w$. Therefore $\hat{y} \neq f_{w'}(x)$ since it does not maximize $\phi(x, \cdot) \cdot w$ as defined in eq.(1). Thus, we prove our claim. □

Next, we provide the final proof.

Proof of Theorem 1. Define the Gibbs decoder *empirical* distortion of the perturbation distribution $Q(w)$ and training set S as:

$$
L(Q(w), S) = \frac{1}{n} \sum_{(x,y) \in S} \mathop{\mathbb{E}}_{w' \sim Q(w)} [d(y, f_{w'}(x))]
$$

In PAC-Bayes terminology, $Q(w)$ is the *posterior* distribution. Let the *prior* distribution P be the unit-variance zero-mean Gaussian distribution. Fix $\delta \in (0, 1)$ and $\alpha > 0$. By well-known PAC-Bayes proof techniques, Lemma 4 in (McAllester, 2007) shows that with probability at least $1 - \delta/2$ over the choice of n training samples, simultaneously for all parameters $w \in \mathcal{W}$, and unit-variance Gaussian posterior distributions $Q(w)$ centered at $w\alpha$, we have:

$$
\begin{aligned}
L(Q(w), D) &\leq L(Q(w), S) + \sqrt{\frac{KL(Q(w) \| P) + \log(2n/\delta)}{2(n-1)}} \\
&= L(Q(w), S) + \sqrt{\frac{\|w\|_2^2 \alpha^2/2 + \log(2n/\delta)}{2(n-1)}}
\end{aligned} \tag{8}
$$

Thus, an upper bound of $L(Q(w), S)$ would lead to an upper bound of $L(Q(w), D)$. In order to upper-bound $L(Q(w), S)$, we can upper-bound each of its summands, i.e., we can upper-bound $\mathbb{E}_{w' \sim Q(w)}[d(y, f_{w'}(x))]$ for each $(x, y) \in S$. Define the distribution $Q(w, x)$ with support on $\mathcal{Y}(x)$ in the following form for all $y \in \mathcal{Y}(x)$:

$$
\mathop{\mathbb{P}}_{y' \sim Q(w,x)} [y' = y] \equiv \mathop{\mathbb{P}}_{w' \sim Q(w)} [f_{w'}(x) = y] \tag{9}
$$

For clarity of presentation, define:

$$
u(x, y, y', w) \equiv H(x, y, y') - m(x, y, y', w)
$$

Let $u \equiv u(x, y, f_{w'}(x), w)$. Simultaneously for all $(x, y) \in S$, we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{w' \sim Q(w)}[d(y, f_{w'}(x)] &= \mathop{\mathbb{E}}_{w' \sim Q(w)}[d(y, f_{w'}(x))\, 1\,(u \geq 0) + d(y, f_{w'}(x))\, 1\,(u < 0)] \\
&\leq \mathop{\mathbb{E}}_{w' \sim Q(w)}[d(y, f_{w'}(x))\, 1\,(u \geq 0) + 1\,(u < 0)] && \text{(10.a)} \\
&= \mathop{\mathbb{E}}_{w' \sim Q(w)}[d(y, f_{w'}(x))\, 1\,(u \geq 0)] + \mathop{\mathbb{P}}_{w' \sim Q(w)}[u < 0] \\
&\leq \mathop{\mathbb{E}}_{w' \sim Q(w)}[d(y, f_{w'}(x))\, 1\,(u \geq 0)] + \|w\|_2^2/n && \text{(10.b)} \\
&= \mathop{\mathbb{E}}_{w' \sim Q(w)}[d(y, f_{w'}(x))\, 1\,(u(x, y, f_{w'}(x), w) \geq 0)] + \|w\|_2^2/n \\
&= \mathop{\mathbb{E}}_{y' \sim Q(w,x)}[d(y, y')\, 1\,(u(x, y, y', w) \geq 0)] + \|w\|_2^2/n && \text{(10.c)} \\
&\leq \max_{\hat{y} \in \mathcal{Y}(x)} d(y, \hat{y})\, 1\,(u(x, y, \hat{y}, w) \geq 0) + \|w\|_2^2/n && \text{(10.d)}
\end{aligned}$$

where the step in eq.(10.a) holds since $d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. The step in eq.(10.b) follows from Lemma 1 which states that $\mathbb{P}_{w' \sim Q(w)}[u(x, y', f_{w'}(x), w) < 0] \leq \|w\|_2^2/n$ for $\alpha = \sqrt{2 \log (2n\ell/\|w\|_2^2)}$, simultaneously for all $(x, y) \in S$, $y' \in \mathcal{Y}(x)$ and $w \in \mathcal{W}$. By the definition in eq.(9), then the step in eq.(10.c) holds. Let $g : \mathcal{Y} \to [0, 1]$ be some arbitrary function, the step in eq.(10.d) uses the fact that $\mathbb{E}_y[g(y)] \leq \max_y g(y)$.

By eq.(8) and eq.(10.d), we prove our claim. □

A.2 Proof of Theorem 2

Here, we provide the proof of Theorem 2. First, we derive an intermediate lemma needed for the final proof.

Lemma 2. *Let $\Delta \in \mathbb{R}^k$ be a random variable, and $w \in \mathbb{R}^k$ be a constant. If $\mathbb{E}[\mu(\Delta)] \cdot w \leq 1/2$ then we have:*

$$\mathbb{P}[\|\Delta\|_1 - \Delta \cdot w < 0] \leq \exp\left(\frac{-1}{32\|w\|_2^2}\right)$$

Proof. Let $t > 0$, we have that:

$$\begin{aligned}
\mathbb{P}[\|\Delta\|_1 - \Delta \cdot w < 0] &= \mathbb{P}[\mu(\Delta) \cdot w > 1] && \text{(11.a)} \\
&= \mathbb{P}[(\mu(\Delta) - \mathbb{E}[\mu(\Delta)]) \cdot w > 1 - \mathbb{E}[\mu(\Delta)] \cdot w] \\
&\leq \mathbb{P}[(\mu(\Delta) - \mathbb{E}[\mu(\Delta)]) \cdot w \geq 1/2] && \text{(11.b)} \\
&= \mathbb{P}[\exp(t(\mu(\Delta) - \mathbb{E}[\mu(\Delta)]) \cdot w) \geq e^{t/2}] \\
&\leq e^{-t/2}\, \mathbb{E}[\exp(t(\mu(\Delta) - \mathbb{E}[\mu(\Delta)]) \cdot w)] && \text{(11.c)} \\
&\leq \exp\left(-t/2 + 2t^2\|w\|_2^2\right) && \text{(11.d)}
\end{aligned}$$

where the step in eq.(11.a) follows from dividing $\|\Delta\|_1 - \Delta \cdot w$ by $\|\Delta\|_1$. Note that $\Delta = 0$ does not fulfill either of the two expressions $\|\Delta\|_1 - \Delta \cdot w < 0$, or $\mu(\Delta) \cdot w > 1$. The step in eq.(11.b) follows from $\mathbb{E}[\mu(\Delta)] \cdot w \leq 1/2$ and thus $1 - \mathbb{E}[\mu(\Delta)] \cdot w \geq 1/2$. The step in eq.(11.c) follows from Markov's inequality. The step in eq.(11.d) follows from Hoeffding's lemma and the fact that the random variable $z = (\mu(\Delta) - \mathbb{E}[\mu(\Delta)]) \cdot w$ fulfills $\mathbb{E}[z] = 0$ as well as $z \in [-2\|w\|_2, +2\|w\|_2]$. In more detail, note that $\|\mu(\Delta)\|_2 \leq 1$ since it holds trivially for $\Delta = 0$, and for $\Delta \neq 0$ we have that $\|\mu(\Delta)\|_2 = \|\Delta\|_2/\|\Delta\|_1 \leq 1$. By Jensen's inequality $\|\mathbb{E}[\mu(\Delta)]\|_2 \leq \mathbb{E}[\|\mu(\Delta)\|_2] \leq 1$. Then, note that by Cauchy-Schwarz inequality $|(\mu(\Delta) - \mathbb{E}[\mu(\Delta)]) \cdot w| \leq \|\mu(\Delta) - \mathbb{E}[\mu(\Delta)]\|_2\|w\|_2 \leq (\|\mu(\Delta)\|_2 + \|\mathbb{E}[\mu(\Delta)]\|_2)\|w\|_2 \leq 2\|w\|_2$. Finally, let $g(t) = -t/2 + 2t^2\|w\|_2^2$. By making $\partial g/\partial t = 0$, we get the optimal setting $t^* = 1/(8\|w\|_2^2)$. Thus, $g(t^*) = -1/(32\|w\|_2^2)$ and we prove our claim. □

Next, we provide the final proof.

Proof of Theorem 2. Note that sampling from the distribution $Q(w, x)$ as defined in eq.(9) is NP-hard in general, thus our plan is to upper-bound the expectation in eq.(10.c) by using the maximum over random structured outputs sampled independently from a proposal distribution $R(w, x)$ with support on $\mathcal{Y}(x)$.

Let $T(w,x)$ be a set of n' i.i.d. random structured outputs drawn from the proposal distribution $R(w,x)$, i.e., $T(w,x) \sim R(w,x)^{n'}$. Furthermore, let $\mathbb{T}(w)$ be the collection of the n sets $T(w,x)$ for all $(x,y) \in S$, i.e. $\mathbb{T}(w) \equiv \{T(w,x)\}_{(x,y)\in S}$ and thus $\mathbb{T}(w) \sim \{R(w,x)^{n'}\}_{(x,y)\in S}$. For clarity of presentation, define:

$$v(x,y,y',w) \equiv d(y,y')\, 1\left(H(x,y,y') - m(x,y,y',w) \geq 0\right)$$

For sets $T(w,x)$ of sufficient size n', our goal is to upper-bound eq.(10.c) in the following form for all parameters $w \in \mathcal{W}$:

$$\frac{1}{n}\sum_{(x,y)\in S} \mathop{\mathbb{E}}_{y'\sim Q(w,x)}[v(x,y,y',w)] \leq \frac{1}{n}\sum_{(x,y)\in S} \max_{\hat{y}\in T(w,x)} v(x,y,\hat{y},w) + \mathcal{O}(\log^{3/2} n/\sqrt{n})$$

Note that the above expression would produce a tighter upper bound than the maximum loss over all possible structured outputs since $\max_{\hat{y}\in T(w,x)} v(x,y,\hat{y},w) \leq \max_{\hat{y}\in \mathcal{Y}(x)} v(x,y,\hat{y},w)$. For analysis purposes, we decompose the latter equation into two quantities:

$$A(w,S) \equiv \frac{1}{n}\sum_{(x,y)\in S}\left(\mathop{\mathbb{E}}_{y'\sim Q(w,x)}[v(x,y,y',w)] - \mathop{\mathbb{E}}_{T(w,x)\sim R(w,x)^{n'}}\left[\max_{\hat{y}\in T(w,x)} v(x,y,\hat{y},w)\right]\right) \tag{12}$$

$$B(w,S,\mathbb{T}(w)) \equiv \frac{1}{n}\sum_{(x,y)\in S}\left(\mathop{\mathbb{E}}_{T(w,x)\sim R(w,x)^{n'}}\left[\max_{\hat{y}\in T(w,x)} v(x,y,\hat{y},w)\right] - \max_{\hat{y}\in T(w,x)} v(x,y,\hat{y},w)\right) \tag{13}$$

Thus, we will show that $A(w,S) \leq \sqrt{1/n}$ and $B(w,S,\mathbb{T}(w)) \leq \mathcal{O}(\log^{3/2} n/\sqrt{n})$ for all parameters $w \in \mathcal{W}$, any training set S and all collections $\mathbb{T}(w)$, and therefore $A(w,S) + B(w,S,\mathbb{T}(w)) \leq \mathcal{O}(\log^{3/2} n/\sqrt{n})$. Note that while the value of $A(w,S)$ is deterministic, the value of $B(w,S,\mathbb{T}(w))$ is stochastic given that $\mathbb{T}(w)$ is a collection of sampled random structured outputs.

Fix a specific $w \in \mathcal{W}$. If data is separable then $v(x,y,y',w) = 0$ for all $(x,y) \in S$ and $y' \in \mathcal{Y}(x)$. Thus, we have $A(w,S) = B(w,S,\mathbb{T}(w)) = 0$ and we complete our proof for the separable case.[4] In what follows, we focus on the nonseparable case.

Bounding the Deterministic Expectation $A(w,S)$. Here, we show that in eq.(12), $A(w,S) \leq \sqrt{1/n}$ for all parameters $w \in \mathcal{W}$ and any training set S, provided that we use a sufficient number n' of random structured outputs sampled from the proposal distribution.

[4]The same result can be obtained for any subset of S for which the "separability" condition holds. Therefore, our analysis with the "nonseparability" condition can be seen as a worst case scenario.

By well-known identities, we can rewrite:

$$
\begin{aligned}
A(w,S) &= \frac{1}{n}\sum_{(x,y)\in S}\int_0^1 \left(\mathop{\mathbb{P}}_{y'\sim R(w,x)}[v(x,y,y',w)\leq z]^{n'} - \mathop{\mathbb{P}}_{y'\sim Q(w,x)}[v(x,y,y',w)\leq z]\right)dz && \text{(14.a)}\\
&\leq \frac{1}{n}\sum_{(x,y)\in S}\mathop{\mathbb{P}}_{y'\sim R(w,x)}[v(x,y,y',w)<1]^{n'}\\
&= \frac{1}{n}\sum_{(x,y)\in S}\mathop{\mathbb{P}}_{y'\sim R(w,x)}[d(y,y')<1 \vee H(x,y,y')-m(x,y,y',w)<0]^{n'}\\
&= \frac{1}{n}\sum_{(x,y)\in S}\left(1-\mathop{\mathbb{P}}_{y'\sim R(w,x)}[d(y,y')=1 \wedge H(x,y,y')-m(x,y,y',w)\geq 0]\right)^{n'}\\
&\leq \frac{1}{n}\sum_{(x,y)\in S}\left(1-\min\left(\mathop{\mathbb{P}}_{y'\sim R(w,x)}[d(y,y')=1]\,,\ \mathop{\mathbb{P}}_{y'\sim R(w,x)}[H(x,y,y')-m(x,y,y',w)\geq 0]\right)\right)^{n'}\\
&= \frac{1}{n}\sum_{(x,y)\in S}\max\left(1-\mathop{\mathbb{P}}_{y'\sim R(w,x)}[d(y,y')=1]\,,\ \mathop{\mathbb{P}}_{y'\sim R(w,x)}[H(x,y,y')-m(x,y,y',w)<0]\right)^{n'}\\
&\leq \max\left(\beta\,,\ \exp\left(\frac{-1}{32\|w\|_2^2}\right)\right)^{n'} && \text{(14.b)}\\
&= \sqrt{1/n} && \text{(14.c)}
\end{aligned}
$$

where the step in eq.(14.a) holds since for two independent random variables $g,h\in[0,1]$, we have $\mathbb{E}[g]=1-\int_0^1\mathbb{P}[g\leq z]dz$ and $\mathbb{P}[\max(g,h)\leq z]=\mathbb{P}[g\leq z]\,\mathbb{P}[h\leq z]$. Therefore, $\mathbb{E}[\max(g,h)] = 1 - \int_0^1\mathbb{P}[g\leq z]\,\mathbb{P}[h\leq z]dz$. For the step in eq.(14.b), we used Assumption A for the first term in the $\max$. For the second term in the $\max$, we used Assumption B. More formally, let $\Delta\equiv\phi(x,y)-\phi(x,y')$ then $H(x,y,y')=\|\Delta\|_1$ and $m(x,y,y',w)=\Delta\cdot w$. By Assumption B, we have that $\|\mathbb{E}[\mu(\Delta)]\|_2\leq 1/(2\sqrt{n})\leq 1/(2\|w\|_2)$. By Cauchy-Schwarz inequality we have $\mathbb{E}[\mu(\Delta)]\cdot w\leq\|\mathbb{E}[\mu(\Delta)]\|_2\|w\|_2\leq\|w\|_2/(2\|w\|_2)\leq 1/2$. Since $\mathbb{E}[\mu(\Delta)]\cdot w\leq 1/2$, we apply Lemma 2 in the step in eq.(14.b). For the step in eq.(14.c), let $\alpha\equiv\max\left(\frac{1}{\log(1/\beta)},32\|w\|_2^2\right)$. Note that $\max\left(\beta\,,\ \exp\left(\frac{-1}{32\|w\|_2^2}\right)\right)=e^{-1/\alpha}$. Furthermore, let $n'=\frac{1}{2}\alpha\log n$. Therefore, $\max\left(\beta\,,\ \exp\left(\frac{-1}{32\|w\|_2^2}\right)\right)^{n'}=(e^{-1/\alpha})^{\frac{1}{2}\alpha\log n}=e^{\frac{-1}{2}\log n}=\sqrt{1/n}$.

Bounding the Stochastic Quantity $B(w,S,\mathbb{T}(w))$. Here, we show that in eq.(13), $B(w,S,\mathbb{T}(w))\leq\mathcal{O}(\log^{3/2}n/\sqrt{n})$ for all parameters $w\in\mathcal{W}$, any training set S and all collections $\mathbb{T}(w)$. For clarity of presentation, define:

$$g(x,y,T,w)\equiv\max_{\hat{y}\in T}v(x,y,\hat{y},w)$$

Thus, we can rewrite:

$$B(w,S,\mathbb{T}(w))=\frac{1}{n}\sum_{(x,y)\in S}\left(\mathop{\mathbb{E}}_{T(w,x)\sim R(w,x)^{n'}}[g(x,y,T(w,x),w)]-g(x,y,T(w,x),w)\right)$$

Let $r(x)\equiv|\mathcal{Y}(x)|$ and thus $\mathcal{Y}(x)\equiv\{y_1\dots y_{r(x)}\}$. Let $\pi(x)=(\pi_1\dots\pi_{r(x)})$ be a permutation of $\{1\dots r(x)\}$ such that $\phi(x,y_{\pi_1})\cdot w<\cdots<\phi(x,y_{\pi_{r(x)}})\cdot w$. Let Π be the collection of the n permutations $\pi(x)$ for all $(x,y)\in S$, i.e. $\Pi=\{\pi(x)\}_{(x,y)\in S}$. From Assumption C, we have that $R(\pi(x),x)\equiv R(w,x)$. Similarly, we rewrite $T(\pi(x),x)\equiv T(w,x)$ and $\mathbb{T}(\Pi)\equiv\mathbb{T}(w)$.

Furthermore, let $\mathcal{W}_{\Pi,S}$ be the set of all $w\in\mathcal{W}$ that induce Π on the training set S. For the parameter space $\mathcal{W}$, collection Π and training set S, define the function class $\mathfrak{G}_{\mathcal{W},\Pi,S}$ as follows:

$$\mathfrak{G}_{\mathcal{W},\Pi,S}\equiv\{g(x,y,T,w)\mid w\in\mathcal{W}_{\Pi,S}\wedge(x,y)\in S\}$$

Note that since $|\mathcal{Y}(x)|\leq r$ for all $(x,y)\in S$, then $|\cup_{(x,y)\in S}\mathcal{Y}(x)|\leq\sum_{(x,y)\in S}|\mathcal{Y}(x)|\leq nr$. Note that each ordering of the nr structured outputs completely determines a collection Π and thus the collection of proposal distributions $R(w,x)$

for each $(x, y) \in S$. Note that since $|\cup_{(x,y)\in S} \mathcal{P}(x)| \leq \ell$, we need to consider $\phi(x, y) \in \mathbb{R}^\ell$. Although we can consider $w \in \mathbb{R}^\ell$, the vector w is sparse with at most $\mathfrak{s}$ non-zero entries. Thus, we take into account all possible subsets of $\mathfrak{s}$ features from ℓ possible features. From results in (Bennett, 1956; Bennett & Hays, 1960; Cover, 1967), we can conclude that there are at most $(nr)^{2(\mathfrak{s}-1)}$ linearly inducible orderings, for a fixed set of $\mathfrak{s}$ features. Therefore, there are at most $\binom{\ell}{\mathfrak{s}}(nr)^{2(\mathfrak{s}-1)} \leq \ell^{\mathfrak{s}}(nr)^{2\mathfrak{s}}$ collections Π.

Fix $\delta \in (0, 1)$. By Rademacher-based uniform convergence[5] and by a union bound over all $\ell^{\mathfrak{s}}(nr)^{2\mathfrak{s}}$ collections Π, with probability at least $1 - \delta/2$ over the choice of n sets of random structured outputs, simultaneously for all parameters $w \in \mathcal{W}$:

$$B(w, S, \mathbb{T}(w)) \leq 2\, \mathfrak{R}_{\mathbb{T}(\Pi)}(\mathfrak{G}_{\mathcal{W},\Pi,S}) + 3\sqrt{\frac{\mathfrak{s}(\log \ell + 2\log(nr)) + \log(4/\delta)}{n}} \tag{15}$$

where $\mathfrak{R}_{\mathbb{T}(\Pi)}(\mathfrak{G}_{\mathcal{W},\Pi,S})$ is the *empirical* Rademacher complexity of the function class $\mathfrak{G}_{\mathcal{W},\Pi,S}$ with respect to the collection $\mathbb{T}(\Pi)$ of the n sets $T(\pi(x), x)$ for all $(x, y) \in S$. For clarity, define:

$$\Delta_p(x, y, y') \equiv \begin{cases} c(p, x, y) - c(p, x, y') & \text{if } p \in \mathcal{P}(x) \\ 0 & \text{otherwise} \end{cases}$$

Let σ be an n-dimensional vector of independent Rademacher random variables indexed by $(x, y) \in S$, i.e., $\mathbb{P}[\sigma_{(x,y)} = +1] = \mathbb{P}[\sigma_{(x,y)} = -1] = 1/2$. The empirical Rademacher complexity is defined as:

$$\begin{aligned}
\mathfrak{R}_{\mathbb{T}(\Pi)}(\mathfrak{G}_{\mathcal{W},\Pi,S}) &\equiv \mathop{\mathbb{E}}_{\sigma}\left[\sup_{g\in\mathfrak{G}_{\mathcal{W},\Pi,S}}\left(\frac{1}{n}\sum_{(x,y)\in S}\sigma_{(x,y)}\, g(x, y, T(\pi(x), x), w)\right)\right] \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{w\in\mathcal{W}_{\Pi,S}}\left(\frac{1}{n}\sum_{(x,y)\in S}\sigma_{(x,y)} \max_{\hat{y}\in T(\pi(x),x)} d(y, \hat{y})\, 1\left(H(x, y, \hat{y}) - m(x, y, \hat{y}, w) \geq 0\right)\right)\right] \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{w\in\mathcal{W}_{\Pi,S}}\left(\frac{1}{n}\sum_{(x,y)\in S}\sigma_{(x,y)} \max_{\hat{y}\in T(\pi(x),x)} d(y, \hat{y})\, 1\left(\|\Delta(x, y, \hat{y})\|_1 - \Delta(x, y, \hat{y})\cdot w \geq 0\right)\right)\right] \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{w\in\mathbb{R}^\ell\setminus\{0\}}\left(\frac{1}{n}\sum_{i\in\{1\ldots n\}}\sigma_i \max_{j\in\{1\ldots n'\}} d_{ij}\, 1\left(\|z_{ij}\|_1 - z_{ij}\cdot w \geq 0\right)\right)\right] && \text{(16.a)} \\
&\leq \sum_{j\in\{1\ldots n'\}}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{w\in\mathbb{R}^\ell\setminus\{0\}}\left(\frac{1}{n}\sum_{i\in\{1\ldots n\}}\sigma_i\, d_{ij}\, 1\left(\|z_{ij}\|_1 - z_{ij}\cdot w \geq 0\right)\right)\right] && \text{(16.b)} \\
&\leq \sum_{j\in\{1\ldots n'\}}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{w\in\mathbb{R}^\ell\setminus\{0\}}\left(\frac{1}{n}\sum_{i\in\{1\ldots n\}}\sigma_i\, 1\left(\|z_{ij}\|_1 - z_{ij}\cdot w \geq 0\right)\right)\right] && \text{(16.c)} \\
&\leq \sum_{j\in\{1\ldots n'\}}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{w\in\mathbb{R}^{\ell+1}\setminus\{0\}}\left(\frac{1}{n}\sum_{i\in\{1\ldots n\}}\sigma_i\, 1\left(z_{ij}\cdot w \geq 0\right)\right)\right] && \text{(16.d)} \\
&\leq 2n'\sqrt{\frac{\mathfrak{s}\log(\ell+1)\log(n+1)}{n}} && \text{(16.e)}
\end{aligned}$$

where in the step in eq.(16.a), the terms σ_i, d_{ij} and z_{ij} correspond to $\sigma_{(x,y)}$, $d(y, \hat{y})$ and $\Delta(x, y, \hat{y})$ respectively. Thus, we assume that index i corresponds to the training sample $(x, y) \in S$, and that index j corresponds to the structured output $\hat{y} \in T(\pi(x), x)$. Note that since $|\cup_{(x,y)\in S} \mathcal{P}(x)| \leq \ell$, thus the step in eq.(16.a) considers $w, z_{ij} \in \mathbb{R}^\ell \setminus \{0\}$ without loss of generality. The step in eq.(16.b) follows from the fact that for any two function classes $\mathfrak{G}$ and $\mathfrak{H}$, we have that $\mathfrak{R}(\{\max(g, h) \mid g \in \mathfrak{G} \wedge h \in \mathfrak{H}\}) \leq \mathfrak{R}(\mathfrak{G}) + \mathfrak{R}(\mathfrak{H})$. The step in eq.(16.c) follows from the composition lemma and the

[5] Note that for the analysis of $B(w, S, \mathbb{T}(w))$, the training set S is fixed and randomness stems from the collection $\mathbb{T}(w)$. Also, note that for applying McDiarmid's inequality, independence of each set $T(w, x)$ for all $(x, y) \in S$ is a sufficient condition, and identically distributed sets $T(w, x)$ are not necessary.

fact that $d_{ij} \in [0, 1]$ for all i and j. The step in eq.(16.d) considers a larger function class, since the value of $\|z_{ij}\|_1$ can be taken as an additional entry in the vector z_{ij} we consider $w, z_{ij} \in \mathbb{R}^{\ell+1} \setminus \{0\}$. The step in eq.(16.e) follows from the Massart lemma, the Sauer-Shelah lemma and the VC-dimension of sparse linear classifiers. That is, for any function class $\mathfrak{G}$, we have that $\mathfrak{R}(\mathfrak{G}) \leq \sqrt{\frac{2VC(\mathfrak{G}) \log(n+1)}{n}}$ where $VC(\mathfrak{G})$ is the VC-dimension of $\mathfrak{G}$. Furthermore, by Theorem 20 of (Neylon, 2006), $VC(\mathfrak{G}) \leq 2\mathfrak{s} \log(\ell + 1)$ for the class $\mathfrak{G}$ of sparse linear classifiers on $\mathbb{R}^{\ell+1}$, with $3 \leq \mathfrak{s} \leq \frac{9}{20}\sqrt{\ell + 1}$.

By eq.(8), eq.(10.c), eq.(14.c), eq.(15) and eq.(16.e), we prove our claim. □

A.3 Proof of Claim i

Proof. For all $(x, y) \in S$ and $w \in \mathcal{W}$, by definition of the total variation distance, we have for any event $\mathcal{A}(x, y, y', w)$:

$$\left| \mathop{\mathbb{P}}_{y' \sim R(w,x)}[\mathcal{A}(x, y, y', w)] - \mathop{\mathbb{P}}_{y' \sim R'(w,x)}[\mathcal{A}(x, y, y', w)] \right| \leq TV(R(w, x) \| R'(w, x))$$

Let the event $\mathcal{A}(x, y, y', w) : d(y, y') = 1 \wedge H(x, y, y') - m(x, y, y', w) \geq 0$. Since $R(w, x)$ fulfills Assumption A with value β_1 and since $TV(R(w, x) \| R'(w, x)) \leq \beta_2$, we have that for all $(x, y) \in S$ and $w \in \mathcal{W}$:

$$\begin{aligned}
\mathop{\mathbb{P}}_{y' \sim R'(w,x)}[\mathcal{A}(x, y, y', w)] &\geq \mathop{\mathbb{P}}_{y' \sim R(w,x)}[\mathcal{A}(x, y, y', w)] - TV(R(w, x) \| R'(w, x)) \\
&\geq 1 - \beta_1 - \beta_2
\end{aligned}$$

which proves our claim. □

A.4 Proof of Claim ii

Proof. Since $d(y, y') = 1\,(y \neq y')$ and since $R(x)$ is a uniform proposal distribution with support on $\mathcal{Y}(x)$, we have:

$$\begin{aligned}
\mathop{\mathbb{P}}_{y' \sim R(x)}[d(y, y') = 1] &= \frac{1}{|\mathcal{Y}(x)|} \sum_{\hat{y} \in \mathcal{Y}(x)} 1\,(d(y, \hat{y}) = 1) \\
&= 1 - \frac{1}{|\mathcal{Y}(x)|} \\
&\geq 1 - 1/2 \qquad \text{(17.a)}
\end{aligned}$$

where the step in eq.(17.a) follows since $|\mathcal{Y}(x)| \geq 2$. □

A.5 Proof of Claim iii

Proof. Let $s = (s_1, s_2, s_3 \dots s_v)$ be the pre-order traversal of y. Let $s' = (s_2, s_1, s_3 \dots s_v)$ be a node ordering where we switched s_1 with s_2. Let $\mathcal{Y}'(x)$ be the set of directed spanning trees of v nodes with node ordering s'.[6] Let $R'(x)$ be the uniform proposal distribution with support on $\mathcal{Y}'(x)$. Since $\mathcal{Y}'(x)$ is the set of directed spanning trees of v nodes with a specific node ordering, then $|\mathcal{Y}'(x)| = \prod_{i=2}^{v} (i - 1) = (v - 1)!$. Moreover, since $d(y, y') = \frac{1}{2(v-1)} \sum_{ij} |A(y)_{ij} - A(y')_{ij}|$

[6] We use the node ordering s' in order to have trees in $\mathcal{Y}'(x)$ with all edges different from y. If we use the node ordering s instead, every tree in $\mathcal{Y}'(x)$ will contain the edge (s_2, s_1), thus no tree in $\mathcal{Y}'(x)$ will have all edges different from y.

and since $R'(x)$ is a uniform proposal distribution with support on $\mathcal{Y}'(x)$, we have:

$$\begin{aligned}
\mathop{\mathbb{P}}_{y'\sim R(x)}[d(y,y')=1] &\geq \mathop{\mathbb{P}}_{y'\sim R'(x)}[d(y,y')=1] \\
&= \mathop{\mathbb{P}}_{y'\sim R'(x)}\left[\sum_{ij}|A(y)_{ij}-A(y')_{ij}|=2(v-1)\right] \\
&= \frac{1}{(v-1)!}\sum_{\hat{y}\in\mathcal{Y}'(x)} 1\left(\sum_{ij}|A(y)_{ij}-A(\hat{y})_{ij}|=2(v-1)\right) \\
&= \frac{1}{(v-1)!}\prod_{i=3}^{v}(i-2) \qquad (18.a)\\
&= 1-\frac{v-2}{v-1}
\end{aligned}$$

where the step in eq.(18.a) follows from the fact that when choosing the parent for the node in position i in the ordering s', we have one option less (i.e., the option that is in y). □

A.6 Proof of Claim iv

Proof. Let $s=(s_1,s_2,s_3\ldots s_v)$ be the pre-order traversal of y. Let $s'=(s_2,s_1,s_3\ldots s_v)$ be a node ordering where we switched s_1 with s_2. Let $\mathcal{Y}'(x)$ be the set of directed acyclic graphs of v nodes and b parents per node, and with node ordering s'.[7] Let $R'(x)$ be the uniform proposal distribution with support on $\mathcal{Y}'(x)$. Since $\mathcal{Y}'(x)$ is the set of directed acyclic graphs of v nodes and b parents per node, and with a specific node ordering, then $|\mathcal{Y}'(x)|=\prod_{i=2}^{b+1}(i-1)\prod_{i=b+2}^{v}\binom{i-1}{b}=b!\prod_{i=b+2}^{v}\binom{i-1}{b}$. Moreover, since $d(y,y')=\frac{1}{b(2v-b-1)}\sum_{ij}|A(y)_{ij}-A(y')_{ij}|$ and since $R'(x)$ is a uniform proposal distribution with support on $\mathcal{Y}'(x)$, we have:

$$\begin{aligned}
\mathop{\mathbb{P}}_{y'\sim R(x)}[d(y,y')=1] &\geq \mathop{\mathbb{P}}_{y'\sim R'(x)}[d(y,y')=1] \\
&= \mathop{\mathbb{P}}_{y'\sim R'(x)}\left[\sum_{ij}|A(y)_{ij}-A(y')_{ij}|=b(2v-b-1)\right] \\
&= \left(b!\prod_{i=b+2}^{v}\binom{i-1}{b}\right)^{-1}\sum_{\hat{y}\in\mathcal{Y}'(x)} 1\left(\sum_{ij}|A(y)_{ij}-A(\hat{y})_{ij}|=b(2v-b-1)\right) \\
&= \left(b!\prod_{i=b+2}^{v}\binom{i-1}{b}\right)^{-1}\prod_{i=3}^{b+1}(i-2)\prod_{i=b+2}^{v}\left(\binom{i-1}{b}-1\right) \qquad (19.a)\\
&= \frac{1}{b}\,\frac{\binom{b+1}{b}-1}{\binom{b+1}{b}}\prod_{i=b+3}^{v}\frac{\binom{i-1}{b}-1}{\binom{i-1}{b}} \\
&\geq \frac{1}{b}\,\frac{\binom{b+1}{b}-1}{\binom{b+1}{b}}\prod_{i=b+3}^{v}\frac{\binom{i-1}{2}-1}{\binom{i-1}{2}} \qquad (19.b)\\
&= \frac{bv}{(b^2+3b+2)(v-2)} \qquad (19.c)\\
&\geq 1-\frac{b^2+2b+2}{b^2+3b+2}
\end{aligned}$$

where the step in eq.(19.a) follows from the fact that when choosing the b parents for the node in position i in the ordering s', we have one option less (i.e., the option that is in y). The step in eq.(19.b) follows from the fact that the function $\frac{z-1}{z}$ is nondecreasing as well as $\binom{a}{2}\leq\binom{a}{b}$ for $a\geq b+2$ and $b\geq 2$. The step in eq.(19.c) follows from the fact $v/(v-2)\geq 1$ for $v>2$. □

[7] We use the node ordering s' in order to have graphs in $\mathcal{Y}'(x)$ with all edges different from y. If we use the node ordering s instead, every graph in $\mathcal{Y}'(x)$ will contain the edge (s_2,s_1), thus no graph in $\mathcal{Y}'(x)$ will have all edges different from y.

A.7 Proof of Claim v

Proof. Since $\mathcal{Y}(x)$ is the set of sets of b elements chosen from v possible elements, then $|\mathcal{Y}(x)| = \binom{v}{b}$. Moreover, since $d(y, y') = \frac{1}{2b}(|y - y'| + |y' - y|)$ and since $R(x)$ is a uniform proposal distribution with support on $\mathcal{Y}(x)$, we have:

$$
\begin{aligned}
\mathop{\mathbb{P}}_{y' \sim R(x)}[d(y, y') = 1] &= \mathop{\mathbb{P}}_{y' \sim R(x)}[|y - y'| + |y' - y| = 2b] \\
&= 1 - \mathop{\mathbb{P}}_{y' \sim R(x)}[|y - y'| + |y' - y| < 2b] \\
&= 1 - \binom{v}{b}^{-1} \sum_{\hat{y} \in \mathcal{Y}(x)} 1\left(|y - \hat{y}| + |\hat{y} - y| < 2b\right) \\
&= 1 - \binom{v}{b}^{-1} \sum_{i=0}^{b-1} \binom{v-b}{i} && \text{(20.a)} \\
&\geq 1 - \binom{v}{b}^{-1} \sum_{i=0}^{b-1} \frac{(v-b)^i}{i!} && \text{(20.b)} \\
&= 1 - \binom{v}{b}^{-1} \frac{e^{v-b} \int_{v-b}^{+\infty} t^{b-1} e^{-t} dt}{(b-1)!} \\
&= 1 - \binom{v}{\lfloor \alpha v \rfloor}^{-1} \frac{e^{v-\lfloor \alpha v \rfloor} \int_{v-\lfloor \alpha v \rfloor}^{+\infty} t^{\lfloor \alpha v \rfloor - 1} e^{-t} dt}{(\lfloor \alpha v \rfloor - 1)!} && \text{(20.c)} \\
&\geq 1 - 1/2 && \text{(20.d)}
\end{aligned}
$$

where the step in eq.(20.a) follows from the fact that for a fixed set y of b elements, if the set $\hat{y}$ has $b - i$ common elements with y, then there are $\binom{v-b}{i}$ possible ways of choosing the remaining i non-common elements in y' from out of $v - b$ possible elements. The step in eq.(20.b) follows from well-known inequalities for the binomial coefficient. The step in eq.(20.c) follows from making $b = \lfloor \alpha v \rfloor$. The step in eq.(20.d) follows for any $\alpha \in [0, 1/2]$. □

A.8 Proof of Claim vi

Proof. Let $\Delta \equiv \phi(x, y) - \phi(x, y')$. We also introduce a superindex p for the partitions. That is, for all $p \in \mathcal{P}(x)$, let $\Delta^p \equiv \phi(x, y) - \phi(x, y')$ for some $y' \in \mathcal{Y}_p(x)$. By assumption, since $y' \in \mathcal{Y}_p(x)$ then $|\Delta_p^p| = b$ and $(\forall q \neq p)\ \Delta_q^p = 0$. Note that $\|\Delta^p\|_1 = \sum_{q \in \mathcal{P}(x)} |\Delta_q^p| = |\Delta_p^p| = b$. Thus $|\Delta_p^p| / \|\Delta^p\|_1 = 1$ and $(\forall q \neq p)\ \Delta_q^p / \|\Delta^p\|_1 = 0$. Therefore:

$$
\begin{aligned}
\left\| \mathop{\mathbb{E}}_{y' \sim R(x)} [\mu(\Delta)] \right\|_2 &= \sqrt{\sum_{q \in \mathcal{P}(x)} \mathop{\mathbb{E}}_{y' \sim R(x)} \left[\frac{\Delta_q}{\|\Delta\|_1} \right]^2} \\
&\leq \sqrt{\sum_{q \in \mathcal{P}(x)} \mathop{\mathbb{E}}_{y' \sim R(x)} \left[\frac{|\Delta_q|}{\|\Delta\|_1} \right]^2} \\
&= \sqrt{\sum_{q \in \mathcal{P}(x)} \left(\sum_{p \in \mathcal{P}(x)} \mathop{\mathbb{P}}_{y' \sim R(x)} [y' \in \mathcal{Y}_p(x)] \frac{|\Delta_q^p|}{\|\Delta^p\|_1} \right)^2} \\
&= \sqrt{\sum_{q \in \mathcal{P}(x)} \left(\mathop{\mathbb{P}}_{y' \sim R(x)} [y' \in \mathcal{Y}_q(x)] \frac{|\Delta_q^q|}{\|\Delta^q\|_1} \right)^2} \\
&= \sqrt{|\mathcal{P}(x)| \left(\frac{1}{|\mathcal{P}(x)|} \right)^2} \\
&= 1/\sqrt{|\mathcal{P}(x)|}
\end{aligned}
$$

where we used the fact that for a uniform proposal distribution $R(x)$, we have $\mathbb{P}_{y' \sim R(w,x)}[y' \in \mathcal{Y}_q(x)] = 1/|\mathcal{P}(x)|$. Finally, since we assume that $n \leq |\mathcal{P}(x)|/4$, we have $1/\sqrt{|\mathcal{P}(x)|} \leq 1/(2\sqrt{n})$ and we prove our claim. □

A.9 Proof of Claim vii

Proof. Let $\Delta \equiv \phi(x,y) - \phi(x,y')$. By assumption $|\Delta_p| = b$ for all $p \in \mathcal{P}(x)$. Note that $\|\Delta\|_1 = \sum_{p\in\mathcal{P}(x)} |\Delta_p| = |\mathcal{P}(x)|\, b$. Thus $|\Delta_p|/\|\Delta\|_1 = 1/|\mathcal{P}(x)|$ for all $p \in \mathcal{P}(x)$. Therefore:

$$\begin{aligned}
\left\| \mathop{\mathbb{E}}_{y'\sim R(w,x)} [\mu(\Delta)] \right\|_2 &= \sqrt{\sum_{p\in\mathcal{P}(x)} \mathop{\mathbb{E}}_{y'\sim R(w,x)} \left[\frac{\Delta_p}{\|\Delta\|_1}\right]^2} \\
&\leq \sqrt{\sum_{p\in\mathcal{P}(x)} \mathop{\mathbb{E}}_{y'\sim R(w,x)} \left[\frac{|\Delta_p|}{\|\Delta\|_1}\right]^2} \\
&= \sqrt{|\mathcal{P}(x)| \left(\frac{1}{|\mathcal{P}(x)|}\right)^2} \\
&= 1/\sqrt{|\mathcal{P}(x)|}
\end{aligned}$$

Finally, since we assume that $n \leq |\mathcal{P}(x)|/4$, we have $1/\sqrt{|\mathcal{P}(x)|} \leq 1/(2\sqrt{n})$ and we prove our claim. □

A.10 Proof of Claim viii

Proof. Algorithm 1 depends solely on the linear ordering induced by the parameter w and the mapping $\phi(x,\cdot)$. That is, at any point in time, Algorithm 1 executes comparisons of the form $\phi(x,y)\cdot w > \phi(x,\hat{y})\cdot w$ for any two structured outputs y and $\hat{y}$. □

A.11 Proof of Claim ix

Proof. Algorithm 2 depends solely on the linear ordering induced by the parameter w and the mapping $\phi(x,\cdot)$. That is, at any point in time, Algorithm 2 executes comparisons of the form $\phi(x,y)\cdot w > \phi(x,\hat{y})\cdot w$ for any two structured outputs y and $\hat{y}$. □