## A. Proof of Lemma 3

*Proof.* We adapt the proof of Rademacher based uniform convergence for our purpose. Fix the distribution over $\mathsf{T}$ to $\mathcal{R}(\mathsf{S}, w')$ for some $w'$. Recall that $\bar{\mathsf{T}} = \{\bar{\mathsf{T}}_i\}$ with $\bar{\mathsf{T}}_i = \{y_i\} \cup \mathsf{T}_i$ and the elements of $\mathsf{T}_i$ are drawn i.i.d. from $\mathcal{R}(x_i, w')$. Since the only random part in $\bar{\mathsf{T}}_i$ is $\mathsf{T}_i$ and $y_i \in \mathsf{S}$, it suffices to show concentration of $\mathbb{E}_{\mathsf{T}}\left[L(w, \mathsf{S}, \mathsf{T})\right] - L(w, \mathsf{S}, \mathsf{T})$ for all $w$ and $\mathsf{S}$. For a fixed $\mathsf{S}$, we will consider $L(w, \mathsf{S}, \mathsf{T})$ to be a function of $\mathsf{T}$ and $w$ and denote it by $L(\mathsf{T}, w; \mathsf{S})$. In what follows, we will consider $\mathsf{T}$ to be an $mn$-dimensional vector whose elements (structured outputs) are conditionally independent (but not identically distributed) given a data set $\mathsf{S}$. Define,

$$\varphi(\mathsf{T}; \mathsf{S}) \stackrel{\text{def}}{=} \sup_{w \in \mathbb{R}^{d,s}} \mathbb{E}_{\mathsf{T} \sim \mathcal{R}(\mathsf{S}, w')}\left[L(\mathsf{T}, w; \mathsf{S})\right] - L(\mathsf{T}, w; \mathsf{S}). \tag{20}$$

$\varphi(\mathsf{T}; \mathsf{S})$ is $(1/m)$-Lipschitz and the elements of $\mathsf{T}$ are independent. Therefore, by McDiarmid's inequality, we have that:

$$\Pr_{\mathsf{T}}\left\{\mathbb{E}_{\mathsf{T}}\left[\varphi(\mathsf{T}; \mathsf{S})\right] - \varphi(\mathsf{T}; \mathsf{S}) \leq \sqrt{\frac{\ln(1/\delta)}{2m}} \mid \mathsf{S}\right\} \geq 1 - \delta. \tag{21}$$

Therefore, with probability at least $1 - \delta$ over the choice of $\mathsf{T}$:

$$\begin{aligned} &(\forall w \in \mathbb{R}^{d,s})\ \mathbb{E}_{\mathsf{T}}\left[L(\mathsf{T}, w; \mathsf{S})\right] - L(\mathsf{T}, w; \mathsf{S}) \\ &\quad \leq \sup_{w \in \mathbb{R}^{d,s}} \mathbb{E}_{\mathsf{T}}\left[L(\mathsf{T}, w; \mathsf{S})\right] - L(\mathsf{T}, w; \mathsf{S}) = \varphi(\mathsf{T}; \mathsf{S}) \\ &\quad \leq \mathbb{E}_{\mathsf{T}}\left[\varphi(\mathsf{T}; \mathsf{S})\right] + \sqrt{\frac{\ln 1/\delta}{2m}}. \end{aligned} \tag{22}$$

Next, we will use a symmetrization argument to bound $\mathbb{E}_{\mathsf{T}}\left[\varphi(\mathsf{T}; \mathsf{S})\right]$. Let $\mathsf{T}' \sim \mathcal{R}(\mathsf{S})$ be an independent copy of $\mathsf{T}$. Observe that:

$$\begin{aligned} \mathbb{E}_{\mathsf{T}'}\left[L(\mathsf{T}, w; \mathsf{S}) \mid \mathsf{T}\right] &= L(\mathsf{T}, w; \mathsf{S}) \\ \mathbb{E}_{\mathsf{T}'}\left[L(\mathsf{T}', w; \mathsf{S}) \mid \mathsf{T}\right] &= \mathbb{E}_{\mathsf{T}}\left[L(\mathsf{T}, w; \mathsf{S})\right]. \end{aligned}$$

Now,

$$\begin{aligned} &\mathbb{E}_{\mathsf{T}}\left[\varphi(\mathsf{T})\right] \\ &= \mathbb{E}_{\mathsf{T}}\left[\sup_{w \in \mathbb{R}^{d,s}} \mathbb{E}_{\mathsf{T}}\left[L(\mathsf{T}, w; \mathsf{S})\right] - L(\mathsf{T}, w; \mathsf{S})\right] \\ &= \mathbb{E}_{\mathsf{T}}\left[\sup_{w \in \mathbb{R}^{d,s}} \mathbb{E}_{\mathsf{T}'}\left[L(\mathsf{T}', w; \mathsf{S}) \mid \mathsf{T}\right] - \mathbb{E}_{\mathsf{T}'}\left[L(\mathsf{T}, w; \mathsf{S}) \mid \mathsf{T}\right]\right] \\ &\leq \mathbb{E}_{\mathsf{T}, \mathsf{T}'}\left[\sup_{w \in \mathbb{R}^{d,s}} \frac{1}{m} \sum_{i=1}^{m} z_i' - z_i\right], \end{aligned}$$

where $z_i' = \Pr_\gamma\left\{f_{w,\gamma,\mathsf{T}'}(x_i) \neq y_i\right\}$ and $z_i = \Pr_\gamma\left\{f_{w,\gamma,\mathsf{T}}(x_i) \neq y_i\right\}$. Since $z_i' - z_i$ has a distribution that is symmetric around zero, $z_i' - z_i$ and $\sigma_i(z_i' - z_i)$ have the same distribution, where $\sigma_i$'s are independent Rademacher variables. Continuing the above derivation,

$$\begin{aligned} &\mathbb{E}_{\mathsf{T}}\left[\varphi(\mathsf{T})\right] \\ &\quad \leq \mathbb{E}_{\mathsf{T}, \mathsf{T}', \sigma}\left[\sup_{w \in \mathbb{R}^{d,s}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (z_i' - z_i)\right] \\ &\quad = \frac{2}{m} \mathbb{E}_{\mathsf{T}, \sigma}\left[\sup_{w \in \mathbb{R}^{d,s}} \sum_{i=1}^{m} \sigma_i \Pr_\gamma\left\{f_{w,\gamma,\mathsf{T}}(x_i) \neq y_i\right\}\right] \\ &\quad = 2\mathbb{E}_{\mathsf{T}}\left[\widehat{\mathfrak{R}}_{\mathsf{T}}(\mathcal{G})\right], \end{aligned}$$

where $\widehat{\mathfrak{R}}_{\mathsf{T}}(\mathcal{G})$ is the empirical Rademacher complexity of the function class $\mathcal{G} = \{g_w \mid w \in \mathbb{R}^{d,s}\}$ with respect to $\mathsf{T}$, with $g_w(x,y) = \Pr_\gamma \{f_{w,\gamma,\mathsf{T}}(x) \neq y\}$. Next, using the same argument as in the proof of Theorem 1, we can bound $\widehat{\mathfrak{R}}_{\mathsf{T}}(\mathcal{G})$ for any set $\mathsf{T}$, and get the following bound:

$$\mathbb{E}_{\mathsf{T}}\left[\varphi(\mathsf{T})\right] \leq 2\sqrt{\frac{s(\log d + 2\log(nr))}{m}} \tag{23}$$

Note that the above differs from the bound in Theorem 1 in the log factor since we need to consider linear orderings of $nr$ structured outputs. Therefore from (22) and (23) we have that:

$$\begin{aligned}\Pr_{\mathsf{T}}\{(\forall w \in \mathbb{R}^{d,s})\ \mathbb{E}_{\mathsf{T}}\left[L(\mathsf{T}, w; \mathsf{S})\right] - L(\mathsf{T}, w; \mathsf{S})& \\ \leq \varepsilon_2(d, s, n, r, m, \delta) \mid \mathsf{S}\} \geq 1 - \delta.&\end{aligned} \tag{24}$$

By Definition 1 and from the results by (Bennett, 1956; Bennett & Hays, 1960; Cover, 1967), there are at most $\binom{d}{s}(mr)^{2s}$ effective (equivalence classes) proposal distributions $\mathcal{R}(.)$ Taking a union bound over all such proposal distributions we prove our claim. $\square$