Multi-task Learning with Labeled and Unlabeled Tasks Supplementary Material

Anastasia Pentina [1] **Christoph H. Lampert** [1]

1. Preliminaries

In this section we list a few results from the literature that will be utilized in the proof of Theorem 1.

Proposition 1 (Lemma 1 in (Ben-David et al., 2010)). *Let d be the VC dimension of the hypothesis set $\mathcal{H}$ and S_1, S_2 be two i.i.d. samples of size n from D_1 and D_2 respectively. Then for any $\delta > 0$ with probability at least $1-\delta$:*

$$\mathrm{disc}(D_1, D_2) \leq \mathrm{disc}(S_1, S_2) + 2\sqrt{\frac{2d\log(2n) + \log(2/\delta)}{n}}.$$

Lemma 1 (Theorem 1 in (Maurer, 2006)). *Let $X_1, \dots, X_n$ be independent random variables taking values in the set $\mathcal{X}$ and f be a function $f : \mathcal{X}^n \to \mathbb{R}$. For any $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ and $y \in \mathcal{X}$ define:*

$$x_{y,k} = (x_1, \dots, x_{k-1}, y, x_{k+1}, \dots, x_n)$$
$$(\inf_k f)(x) = \inf_{y \in \mathcal{X}} f(x_{y,k})$$
$$\Delta_{+,f} = \sum_{i=1}^{n} (f - \inf_k f)^2.$$

Then for $t > 0$:

$$\Pr\{f - \mathbb{E} f \geq t\} \leq \exp\left(\frac{-t^2}{2\|\Delta_+\|_\infty}\right). \tag{1}$$

Lemma 2 (Corollary 6.10 in (McDiarmid, 1989)). *Let W_0^n be a martingale with respect to a sequence of random variables $(B_1, \dots, B_n)$. Let $b_1^n = (b_1, \dots, b_n)$ be a vector of possible values of the random variables $B_1, \dots, B_n$. Let*

$$r_i(b_1^{i-1}) = \sup_{b_i}\{W_i : B_1^{i-1} = b_1^{i-1}, B_i = b_i\} - \inf_{b_i}\{W_i : B_1^{i-1} = b_1^{i-1}, B_i = b_i\}. \tag{2}$$

Let $r^2(b_1^n) = \sum_{i=1}^n (r_i(b_1^{i-1}))^2$ and $\widehat{R}^2 = \sup_{b_1^n} r^2(b_1^n)$. Then

$$\Pr_{B_1^n}\{W_n - W_0 > \epsilon\} < \exp\left(-\frac{2\epsilon^2}{\widehat{R}^2}\right). \tag{3}$$

Lemma 3 (Originally (Hoeffding, 1963); in this form Theorem 18 in (Tolstikhin et al., 2014)). *Let $\{U_1, \dots, U_m\}$ and $\{W_1, \dots, W_m\}$ be sampled uniformly from a finite set of d-dimensional vectors $\{v_1, \dots, v_N\} \subset \mathbb{R}^d$ with and without replacement respectively. Then for any continuous and convex function $F : \mathbb{R}^d \to \mathbb{R}$ the following holds:*

$$\mathbb{E}\left[F\left(\sum_{i=1}^m W_i\right)\right] \leq \mathbb{E}\left[F\left(\sum_{i=1}^m U_i\right)\right] \tag{4}$$

[1]IST Austria. Correspondence to: Anastasia Pentina <apentina@ist.ac.at>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Lemma 4 (Part of Lemma 19 in (Tolstikhin et al., 2014)). *Let $x = (x_1, \ldots, x_l) \in \mathbb{R}^l$. Then the following function is convex:*

$$F(x) = \sup_{i=1\ldots l} x_i. \tag{5}$$

2. Proof of Theorem 1

We start with bounding the multi-task error by the errors on the source tasks, and transition to empirical quantities while keeping the effect of random sampling controlled.

Fix a subset of labeled tasks $I = \{i_1, \ldots, i_k\}$, a task $\langle D_t, f_t\rangle$ and a weight vector $\alpha \in \Lambda^I$. Let $h_i^* \in \arg\min_{h\in\mathcal{H}}(\mathrm{er}_t(h) + \mathrm{er}_i(h))$.[1] Writing $\ell(h, h')$ as shorthand for $\ell(h(x), h'(x))$, we have

$$|\,\mathrm{er}_\alpha(h) - \mathrm{er}_t(h)| = \Big|\sum_{i\in I}\alpha_i\,\mathrm{er}_i(h) - \mathrm{er}_t(h)\Big| \le \sum_{i\in I}\alpha_i\big|\,\mathrm{er}_i(h) - \mathrm{er}_t(h)\big| \tag{6}$$

$$\le \sum_{i\in I}\alpha_i\left(\big|\,\mathrm{er}_i(h) - \mathop{\mathbb{E}}_{x\sim D_i}\ell(h, h_i^*)\big| + \big|\mathop{\mathbb{E}}_{x\sim D_i}\ell(h, h_i^*) - \mathop{\mathbb{E}}_{x\sim D_t}\ell(h, h_i^*)\big| + \big|\,\mathrm{er}_t(h) - \mathop{\mathbb{E}}_{x\sim D_t}\ell(h, h_i^*)\big|\right) = (*) \tag{7}$$

We can bound each summand:

$$|\,\mathrm{er}_i(h) - \mathop{\mathbb{E}}_{x\sim D_i}\ell(h, h_i^*)| \le \mathrm{er}_i(h^*)$$
$$|\mathop{\mathbb{E}}_{x\sim D_i}\ell(h, h_i^*) - \mathop{\mathbb{E}}_{x\sim D_t}\ell(h, h_i^*)\big| \le \mathrm{disc}(D_i, D_t)$$
$$|\,\mathrm{er}_t(h) - \mathop{\mathbb{E}}_{x\sim D_t}\ell(h, h_i^*)\big| \le \mathrm{er}_t(h_i^*)$$

where the first and the last inequalities hold by the triangular inequality for ℓ and the second one follows from the definition of discrepancy. Therefore,

$$(*) \le \sum_{i\in I}\alpha_i(\mathrm{er}_i(h_i^*) + \mathrm{disc}(D_i, D_t) + \mathrm{er}_t(h_i^*)) = \sum_{i\in I}\alpha_i(\lambda_{it} + \mathrm{disc}(D_i, D_t)). \tag{8}$$

Consequently, assuming that every task t has its own weights α^t we obtain that:

$$\frac{1}{T}\sum_{t=1}^T \mathrm{er}_t(h) \le \frac{1}{T}\sum_{t=1}^T \mathrm{er}_{\alpha^t}(h_t) + \frac{1}{T}\sum_{t=1}^T\sum_{i\in I}\alpha_i^t\,\mathrm{disc}(D_t, D_i) + \frac{1}{T}\sum_{t=1}^T\sum_{i\in I}\alpha_i^t\lambda_{ti}. \tag{9}$$

We continue with bounding every expectation on the right hand side of (9) by its empirical counterpart.

2.1. Bound $\frac{1}{T}\sum_{t=1}^T\sum_{i\in I}\alpha_i^t\,\mathrm{disc}(D_t, D_i)$

We apply Proposition 1 to every summand and combine the results using a union bound argument. We obtain that with probability at least $1-\delta/2$ uniformly for all choices of I and $\alpha^1, \ldots, \alpha^T \in \Lambda^I$:

$$\frac{1}{T}\sum_{t=1}^T\sum_{i\in I}\alpha_i^t\,\mathrm{disc}(D_t, D_i) \le \frac{1}{T}\sum_{t=1}^T\sum_{i\in I}\alpha_i^t\,\mathrm{disc}(S_t, S_i) + 2\sqrt{\frac{2d\log(2n) + \log(4T^2/\delta)}{n}}. \tag{10}$$

2.2. Bound $\frac{1}{T}\sum_{t=1}^T \mathrm{er}_{\alpha^t}(h_t)$

Now we upper-bound the error term in two steps.

[1] If the minimum is not attained, the same inequality follows by an argument of arbitrary close approximation.

2.2.1. RELATE $\frac{1}{T}\sum_{t=1}^{T}\mathrm{er}_{\alpha^t}(h_t)$ TO $\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathrm{er}}_{\alpha^t}(h_t)$

We start with relating the multi-task error to the hypothetical empirical error, if the learner would receive labels for all examples in the selected labeled tasks:

$$\tilde{\mathrm{er}}_{\alpha}(h)=\sum_{i\in I}\alpha_i\widehat{\mathrm{er}}_{S_i^u}(h) \tag{11}$$

for

$$\widehat{\mathrm{er}}_{S_i^u}(h)=\frac{1}{n}\sum_{j=1}^{n}\ell(h(x_j^i),f_i(x_j^i)). \tag{12}$$

Clearly, if $m=n$ this part is not necessary and we can avoid the resulting complexity terms.

Because the choice of the tasks to label, I, their weights, $\boldsymbol{\alpha}=(\alpha^1,\ldots,\alpha^T)$, and the predictors, $\mathbf{h}=(h_1,\ldots,h_T)$, all depend on the unlabeled data, we aim for a bound that is holds simultaneous for all choices of these quantities, under the condition that I and $\boldsymbol{\alpha}$ depend only on the unlabeled samples, while $\mathbf{h}$ can be chosen based also on the labeled subsets.

Our main tool is a refined version of McDiarmid's inequality, due to Maurer (Maurer, 2006) (Lemma 1), which allows us to make use of the internal structure of the weights, $\boldsymbol{\alpha}$, while deriving a large deviation bound.

For any $\mathbf{S}=(S_1^u,\ldots,S_T^u)$ define:

$$\Psi(\mathbf{S})=\sup_{I=\{i_1,\ldots,i_k\}}\sup_{\alpha^1,\ldots,\alpha^n\in\Lambda^I}\sup_{h_1,\ldots,h_T}\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\alpha_i^t(\mathrm{er}_i(h_t)-\widehat{\mathrm{er}}_{S_i^u}(h_t))=\sup_{I}\sup_{\boldsymbol{\alpha}}\sup_{\mathbf{h}}\ g(\boldsymbol{\alpha},\mathbf{h},\mathbf{S}) \tag{13}$$

for

$$g(\boldsymbol{\alpha},\mathbf{h},\mathbf{S})=\sum_{i=1}^{T}\sum_{j=1}^{n}\left(\frac{1}{Tn}\sum_{t=1}^{T}\alpha_i^t(\mathrm{er}_i(h_t)-\ell(h_t(x_j^i),f_t(x_j^i)))\right). \tag{14}$$

For notational simplicity we will sometimes think of every S_t^u as a set of *pairs* (x_i^t,y_i^t), where $y_i^t=f_t(x_i^t)$. To apply Lemma 1 we establish a bound on $\Delta_{+,\Psi}(\mathbf{S})=\sum_i\sum_j(\Psi(\mathbf{S})-\Psi_{ij}(\mathbf{S}))^2$, with

$$\Psi_{ij}(\mathbf{S})=\inf_{(x,y)}\sup_{\boldsymbol{\alpha}}\sup_{\mathbf{h}}\ g(\boldsymbol{\alpha},\mathbf{h},\mathbf{S}\setminus\{(x_j^i,y_j^i)\}\cup\{(x,y)\}, \tag{15}$$

i.e. the possible smallest value for Ψ when changing only the data point (x_j^i,y_j^i). Let $\boldsymbol{\alpha}^*,\mathbf{h}^*$ be the point where the sup in the (13) is attained[2], i.e. $\Psi(\mathbf{S})=g(\boldsymbol{\alpha}^*,\mathbf{h}^*,\mathbf{S})$. Then:

$$\Psi_{ij}(\mathbf{S})\geq\inf_{(x,y)}g(\boldsymbol{\alpha}^*,\mathbf{h}^*,\mathbf{S}\setminus\{(x_j^i,y_j^i)\}\cup\{(x,y)\}\ ) \tag{16}$$

and therefore

$$\Psi(\mathbf{S})-\Psi_{ij}(\mathbf{S})\leq g(\boldsymbol{\alpha}^*,\mathbf{h}^*,\mathbf{S})-\inf_{(x,y)}g(\boldsymbol{\alpha}^*,\mathbf{h}^*,\mathbf{S}\setminus\{(x_j^i,y_j^i)\}\cup\{(x,y)\}) \tag{17}$$

$$\leq\sup_{(x,y)}\frac{1}{Tn}\sum_{t=1}^{T}\alpha_i^{*t}(-\ell(h_t^*(x_j^i),y_j^i)+\ell(h_t^*(x),y))\ \leq\ \frac{1}{Tn}\sum_{t=1}^{T}\alpha_i^{*t}, \tag{18}$$

where for the last inequality we use that ℓ is bounded in $[0,1]$. Because also $\Psi(\mathbf{S})-\Psi_{ij}(\mathbf{S})\geq 0$, we obtain

$$\Delta_{+,\Psi}(\mathbf{S})=\sum_{i=1}^{T}\sum_{j=1}^{n}(\Psi(\mathbf{S})-\Psi_{ij}(\mathbf{S}))^2\leq\sum_{i=1}^{T}\sum_{j=1}^{n}\frac{1}{T^2n^2}\left(\sum_{t=1}^{T}\alpha_i^{*t}\right)^2\leq\frac{1}{T^2n}\left(\sum_{i=1}^{T}\sum_{t=1}^{T}\alpha_i^{*t}\right)^2=\frac{1}{n}, \tag{19}$$

[2] If the supremum is not attained the subsequent inequality still follows from an argument of arbitrarily close approximation.

(remember that $\sum_i \alpha_i = 1$ for any $\alpha \in \Lambda^I$). Therefore, according to Lemma 1 with probability at least $1 - \delta/4$:

$$\Psi(\mathbf{S}) \leq \mathbb{E}\ \Psi(\mathbf{S}) + \sqrt{\frac{2}{n}\log\frac{4}{\delta}}. \tag{20}$$

To bound $\mathbb{E}_S\ \Psi(\mathbf{S})$ we use symmetrization and Rademacher variables, σ_{ij}:

$$\mathop{\mathbb{E}}_{\mathbf{S}} \Psi(\mathbf{S}) = \mathop{\mathbb{E}}_{\mathbf{S}} \sup_I \sup_{\alpha^1,\ldots,\alpha^T \in \Lambda^I} \sup_{h_1,\ldots,h_T} \sum_{i=1}^T \sum_{j=1}^n \left(\frac{1}{Tn} \sum_{t=1}^T \alpha_i^t (\mathrm{er}_i(h_t) - \ell(h_t(x_j^i), y_j^i)) \right) \tag{21}$$

$$\leq 2 \mathop{\mathbb{E}}_{\mathbf{S}} \mathop{\mathbb{E}}_{\sigma} \sup_I \sup_{\alpha^1,\ldots,\alpha^T \in \Lambda^I} \sup_{h_1,\ldots,h_T} \sum_{i=1}^T \sum_{j=1}^n \left(\frac{\sigma_{ij}}{Tn} \sum_{t=1}^T \alpha_i^t \ell(h_t(x_j^i), y_j^i) \right) \tag{22}$$

$$\leq 2 \mathop{\mathbb{E}}_{\mathbf{S}} \mathop{\mathbb{E}}_{\sigma} \frac{1}{T} \sum_{t=1}^T \sup_{\alpha^t \in \Lambda, h_t} \sum_{i=1}^T \sum_{j=1}^n \frac{\sigma_{ij}\alpha_i^t}{n} \sum_{t=1}^T \ell(h_t(x_j^i), y_j^i) \tag{23}$$

$$\leq 2 \mathop{\mathbb{E}}_{\mathbf{S}} \mathop{\mathbb{E}}_{\sigma} \sup_{\alpha, h} \sum_{i=1}^T \sum_{j=1}^n \frac{\sigma_{ij}\alpha_i}{n} \ell(h(x_j^i), y^i), \tag{24}$$

where line (23) is obtained from line (22) by dropping the assumption of a common sparsity pattern between the α-s. Note that the function inside the last sup is linear in $\alpha \in \Lambda$, therefore $\sup_\alpha$ can be reduced to the sup over the corners of the simplex, $\{(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)\}$. At the same time, by Sauer's lemma, the number of different choices of h on $\mathbf{S}$ is bounded by $\left(\frac{eTn}{d}\right)^d$. Therefore, the total number of different choices in (24) is bounded by $T\left(\frac{enT}{d}\right)^d$. Furthermore, for any choice of α and h, the norm of the Tn-vector formed by the summands of (24) is bounded by $1/\sqrt{n}$, because

$$\sum_{i=1}^T \sum_{j=1}^n \left(\frac{\sigma_{ij}\alpha_i}{n} \ell(h(x_j^i), y^i) \right)^2 = \frac{1}{n^2} \sum_{i=1}^T \sum_{j=1}^n \left(\alpha_i \ell(h(x_j^i), y^i) \right)^2 \leq \frac{1}{n^2} \sum_{j=1}^n \left(\sum_{i=1}^T \alpha_i \right)^2 = \frac{1}{n}. \tag{25}$$

Therefore, by Massart's lemma:

$$\mathop{\mathbb{E}}_{\sigma} \sup_{\alpha, h} \sum_{i=1}^T \sum_{j=1}^n \frac{\sigma_{il}\alpha_i}{n} \ell(h(x_l^i), y_l^i) \leq \frac{\sqrt{2(\log T + d\log(enT/d))}}{\sqrt{n}}. \tag{26}$$

Combining (20) and (26) we obtain that with probability at least $1 - \delta/4$ simultaneously for all choices of tasks to be labeled, I, weights $\boldsymbol{\alpha}$ and hypotheses $\mathbf{h}$:

$$\frac{1}{T} \sum_{t=1}^T \mathrm{er}_{\alpha^t}(h_t) \leq \frac{1}{T} \sum_{t=1}^T \tilde{\mathrm{er}}_{\alpha^t}(h_t) + \sqrt{\frac{8(\log T + d\log(enT/d))}{n}} + \sqrt{\frac{2}{n}\log\frac{4}{\delta}}. \tag{27}$$

2.2.2. Relate $\frac{1}{T}\sum_{t=1}^T \widehat{\mathrm{er}}_{\alpha^t}(h_t)$ to $\frac{1}{T}\sum_{t=1}^T \tilde{\mathrm{er}}_{\alpha^t}(h_t)$

Fix the unlabeled samples $S_1^u, \ldots, S_T^u$. This uniquely determines the chosen tasks I and the weights $\alpha^1, \ldots, \alpha^T \in \Lambda^I$, so the only remaining source of randomness is the uncertainty which subsets of the selected tasks are labeled.

For notational simplicity we pretend that exactly the first k tasks were selected, i.e. $I = \{1, \ldots, k\}$. The general case can be obtained by changing the indices in the proof from $1, \ldots, k$ to $i_1, \ldots, i_k$.

To deal with the dependencies between the labeled data points we first note that any random labeled subset $S_i^l = (\bar{s}_1^i, \ldots, \bar{s}_m^i)$ can be described as the first m elements of a random permutation $Z_i = (z_1^i, \ldots, z_n^i)$ over n elements that correspond to the unlabeled sample S_i^u, i.e. $\bar{s}_j^i = (\bar{x}_j^i, \bar{y}_j^i) = (x_{z_j^i}^i, y_{z_j^i}^i)$. With this notation and writing $\mathbf{Z} = (Z_1, \ldots, Z_k)$ and $\ell(h, z_j^i) = \ell(h(\bar{x}_j^i), \bar{y}_j^i)$ we define the following function

$$\Phi(\mathbf{Z}) = \sup_{h_1,\ldots,h_T} \frac{1}{T} \sum_{t=1}^T \tilde{\mathrm{er}}_{\alpha^t}(h_t) - \widehat{\mathrm{er}}_{\alpha^t}(h_t) = \sup_{h_1,\ldots,h_T} \sum_{i=1}^k \frac{1}{T} \sum_{t=1}^T \alpha_i^t \Big(\frac{1}{n} \sum_{j=1}^n \ell(h_t, z_j^i) - \frac{1}{m} \sum_{j=1}^m \ell(h_t, z_j^i) \Big). \tag{28}$$

Our main tool is McDiarmid's inequality (Lemma 2) for martingales.

Construct a martingale sequence

For this, we interpret $\mathbf{Z} = (z_1^1, z_2^1, \ldots, z_n^k)$ as a sequence of kn dependent variables, $z_{11}, \ldots, z_{kn}$. For the sake of notational consistency we will keep using double indices, with the convention that the sample index, $j = 1, \ldots, n$, runs faster than the task index, $i = 1, \ldots, k$. Segments of a sequence will be denoted by upper and lower double indices, $z_{ij}^{\overline{i}\overline{j}} = (z_{ij}, z_{i(j+1)}, \ldots, z_{\overline{i}\overline{j}})$ for $ij \leq \overline{i}\overline{j}$ and $z_{ij}^{\overline{i}\overline{j}} = \emptyset$ otherwise. We now create a martingale sequence using Doob's construction (Doob, 1940):

$$W_{ij} = \mathop{\mathbb{E}}_{\mathbf{Z}}\{\Phi(\mathbf{Z})|\ z_{11}^{ij}\ \}. \tag{29}$$

where here and in the following when taking expectations over $\mathbf{Z}$ it is silently assumed that the expectation is taken only with respect to variables that are not conditioned on. Note that because of this convention, the expectations in (29) is only with respect to $z_{i(j+1)}, \ldots, z_{kn}$, so each W_{ij} is a random variable of $z_{11}, \ldots, z_{ij}$. In particular, $W_{00} = \mathbb{E}_{\mathbf{Z}}\,\Phi(\mathbf{Z})$ and $W_{kn} = \Phi(\mathbf{Z})$, and the in between sequence is a martingale with respect to $z_{11}, \ldots, z_{kn}$:

$$\mathop{\mathbb{E}}_{\mathbf{Z}}\{\ W_{ij}|z_{11}^{i(j-1)}\ \} = \mathop{\mathbb{E}}_{\mathbf{Z}}\{\ \mathop{\mathbb{E}}_{\mathbf{Z}}\{\Phi(\mathbf{Z})|\ z_{11}^{ij}\}\big|\ z_{11}^{i(j-1)}\ \} = \mathop{\mathbb{E}}_{\mathbf{Z}}\{\Phi(\mathbf{Z})|z_{11}^{i(j-1)}\ \} = W_{i(j-1)}. \tag{30}$$

Upper-bound $\widehat{R}^2$

In order to apply Lemma 2 we need an upper bound on the coefficient $\widehat{R}^2$ defined there.

Let $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, n\}$ be fixed and let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ be specific permutations of n elements for which we use the same index conventions as for $\mathbf{Z}$. By σ and τ will denote elements in $\pi_{i(j+1)}^{in}$, i.e. σ and τ do not occur in any of the first j positions of the permutation π_i. Then

$$\begin{aligned} r_{ij}(\pi_{11}^{i(j-1)}) &= \sup_{\sigma \in \pi_{i(j+1)}^{in}} \{\ W_{ij} : z_{11}^{i(j-1)} = \pi_{11}^{i(j-1)},\ z_{ij} = \sigma\} - \inf_{\sigma \in \pi_{i(j+1)}^{in}} \{\ W_{ij} : z_{11}^{i(j-1)} = \pi_{11}^{i(j-1)},\ z_{ij} = \sigma\} \\ &= \sup_{\sigma \in \pi_{i(j+1)}^{in}} \sup_{\tau \in \pi_{i(j+1)}^{in}} \Big[\mathop{\mathbb{E}}_{z_{i(j+1)}^{kn}} \{\Phi(\pi_{11}^{i(j-1)}, \sigma, z_{i(j+1)}^{kn})\} - \mathop{\mathbb{E}}_{z_{i(j+1)}^{kn}} \{\Phi(\pi_{11}^{i(j-1)}, \tau, z_{i(j+1)}^{kn})\} \Big]. \end{aligned} \tag{31}$$

To analyze (31) further, recall that:

$$\mathop{\mathbb{E}}_{z_{i(j+1)}^{kn}} \{\Phi(\pi_{11}^{i(j-1)}, \sigma, z_{i(j+1)}^{kn})\} = \sum_{\pi_{i(j+1)}^{kn}} \Phi(\pi_{11}^{i(j-1)}, \sigma, \pi_{i(j+1)}^{kn}) \times \Pr(\ z_{i(j+1)}^{kn} = \pi_{i(j+1)}^{kn}\ |z_{11}^{i(j-1)} = \pi_{11}^{i(j-1)} \wedge z_{ij} = \sigma\ ),$$

where here and in the following we use the convention that sums over parts of $\boldsymbol{\pi}$ run only over values that lead to valid permutations. Because the permutations of different task are independent, this is equal to

$$= \sum_{\pi_{i(j+1)}^{kn}} \Phi(\pi_{11}^{i(j-1)}, \sigma, \pi_{i(j+1)}^{kn}) \Pr(\ z_{i(j+1)}^{in} = \pi_{i(j+1)}^{in}\ |z_{i1}^{i(j-1)} = \pi_{i1}^{i(j-1)} \wedge z_{ij} = \sigma\ ) \Pr(z_{(i+1)1}^{kn} = \pi_{(i+1)1}^{kn}) \tag{32}$$

We make the following observation: for any fixed π_{i1}^{ij} and any $\tau \notin \pi_{i1}^{ij}$, we can rephrase a summation over $\pi_{i(j+1)}^{in}$ into a sum over all positions where τ can occur, and a sum over all configuration for the entries that are not τ:

$$\sum_{\pi_{i(j+1)}^{in}} F(\pi_{i(j+1)}^{in}) = \sum_{l=j+1}^{n} \sum_{\pi_{i(j+1)}^{i(l-1)}} \sum_{\pi_{i(l+1)}^{in}} F(\pi_{i(j+1)}^{i(l-1)}, \tau, \pi_{i(l+1)}^{in}) \tag{33}$$

for any function F. Applying this to the summation in (32), we obtain

$$\begin{aligned} &\sum_{\pi_{i(j+1)}^{kn}} \Phi(\pi_{11}^{i(j-1)}, \sigma, \pi_{i(j+1)}^{kn}) \Pr(\ z_{i(j+1)}^{in} = \pi_{i(j+1)}^{in}\ |z_{i1}^{i(j-1)} = \pi_{i1}^{i(j-1)} \wedge z_{ij} = \sigma\ ) \\ &\quad \times \Pr(z_{(i+1)1}^{kn} = \pi_{(i+1)1}^{kn}) = \sum_{l=j+1}^{n} \sum_{\pi_{i(j+1)}^{i(l-1)}} \sum_{\pi_{i(l+1)}^{kn}} \Phi(\pi_{11}^{i(j-1)}, \sigma, \pi_{i(j+1)}^{i(l-1)}, \tau, \pi_{i(l+1)}^{kn}) \end{aligned}$$

$$\times \Pr(\, z^{i(l-1)}_{i(j+1)} = \pi^{i(l-1)}_{i(j+1)} \wedge z^{kn}_{i(l+1)} = \pi^{kn}_{i(l+1)} | z^{i(j-1)}_{11} = \pi^{i(j-1)}_{11} \wedge z_{ij} = \sigma \wedge z_{il} = \tau)$$
$$\times \Pr(z^{kn}_{(i+1)1} = \pi^{kn}_{(i+1)1}) = \mathop{\mathbb{E}}_{l \sim U^n_{j+1}} \mathop{\mathbb{E}}_{\mathbf{Z}} \Phi(\mathbf{Z} | z^{i(j-1)}_{11} = \pi^{i(j-1)}_{11} \wedge z_{ij} = \sigma \wedge z_{il} = \tau),$$

where U^n_{j+1} denotes the uniform distribution over the set $\{j+1, \dots, n\}$. The analogue derivation can be applied to the quantity in line (31) with σ and τ exchanged.

For any $\mathbf{Z}$ denote by $\mathbf{Z}^{ij \leftrightarrow il}$ the permutation obtained by switching z_{ij} and z_{il}. Then, due to the linearity of the expectation:

$$r_{ij}(\pi^{i(j-1)}_{11}) = \sup_{\sigma,\tau} \{ \mathop{\mathbb{E}}_{l \sim U^n_{j+1}} \mathop{\mathbb{E}}_{\mathbf{Z}} \{ \Phi(\mathbf{Z}) - \Phi(\mathbf{Z}^{ij \leftrightarrow il}) | z^{i(j-1)}_{11} = \pi^{i(j-1)}_{11}, z_{ij} = \sigma, z_{il} = \tau). \tag{34}$$

From the definition of Φ we see that $\Phi(\mathbf{Z}) - \Phi(\mathbf{Z}^{ij \leftrightarrow il}) = 0$ when $j, l \in \{1, \dots, m\}$ or $j, l \in \{m+1 \dots, n\}$. Since $l > j$ in (34) this implies $r_{ij}(\pi^{i(j-1)}_{11}) = 0$ for $j \in \{m+1, \dots, n\}$. The only remaining cases are $j \in \{1, \dots, m\}$ and $l \in \{m+1, \dots, n\}$, for which we obtain

$$\Phi(\mathbf{Z}) - \Phi(\mathbf{Z}^{ij \leftrightarrow il}) \le \sup_{h_1, \dots, h_T} \frac{1}{T} \sum_{t=1}^{T} \alpha^t_i \frac{1}{m} (-\ell(h_t, z^i_j) + \ell(h_t, z^i_l)) \le \frac{1}{Tm} \sum_{t=1}^{T} \alpha^t_i.$$

where for the first inequality we used that $\sup F - \sup G \le \sup(F - G)$ for any F, G, and for the second inequality we used that ℓ is bounded by $[0, 1]$. Consequently, $r_{ij}(\pi^{i(j-1)}_{11}) \le \frac{n-m}{n-j} \frac{1}{Tm} \sum_{t=1}^{T} \alpha^t_i$ in this case. Therefore[3]

$$\widehat{R}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(r_{ij}(\pi^{i(j-1)}_{11}) \right)^2 \le \frac{1}{T^2 m^2} \sum_{j=1}^{m} \left(\frac{n-m}{n-j} \right)^2 \sum_{i=1}^{k} \left(\sum_{t=1}^{T} \alpha^t_i \right)^2 \le \frac{1}{T^2 m} \sum_{i=1}^{k} \left(\sum_{t=1}^{T} \alpha^t_i \right)^2. \tag{35}$$

Upper-bound $\mathbb{E}_{\mathbf{Z}} \Phi(\mathbf{Z})$

The main tool here is Lemma 3. First we rewrite $\Phi(\mathbf{Z})$ in the following way:

$$\Phi(\mathbf{Z}) = \frac{1}{T} \sum_{t=1}^{T} \sup_h \sum_{i=1}^{k} \alpha^t_i (\widehat{\mathrm{er}}_{S^u_i}(h) - \widehat{\mathrm{er}}_{S^l_i}(h)) = \frac{1}{Tm} \sum_{t=1}^{T} \Phi_t(\mathbf{Z})$$
$$\Phi_t(\mathbf{Z}) = \sup_h \sum_{i=1}^{k} m\, \alpha^t_i (\widehat{\mathrm{er}}_{S^u_i}(h) - \widehat{\mathrm{er}}_{S^l_i}(h)).$$

Note that even though $\mathcal{H}$ can be infinitely large, we can identify a finite subset that represents all possible predictions of hypothesis in $\mathcal{H}$ on $S^u_1 \cup \dots \cup S^u_k$. We denote their number by $L \le 2^{kn}$ and the corresponding hypotheses by $h^1, \dots, h^L$.

Let $t \in \{1, \dots, T\}$ be fixed. For every $i \in \{1, \dots, k\}$ define a set of n L-dimensional vectors, $V^t_i = \{v^t_{i1}, \dots, v^t_{in}\}$, where for every $j \in \{1, \dots, n\}$:

$$v^t_{ij} = \Big[\alpha^t_i \big(\tilde{\mathrm{er}}_i(h^1) - \ell(h^1(x^i_j), y^i_j) \big),\ \dots,\ \alpha^t_i \big(\tilde{\mathrm{er}}_i(h^L) - \ell(h^L(x^i_j), y^i_j) \big) \Big]. \tag{36}$$

With this notation, for every $i \in \{1, \dots, k\}$ choosing a random subset $S^l_i \subset S^u_i$ corresponds to sampling m vectors from V^t_i uniformly without replacement.

For every $i \in \{1, \dots, k\}$, let $U_i = \{u_{i1}, \dots, u_{im}\}$ be sampled from V^t_i in that way. Then

$$\Phi_t(\mathbf{Z}) = F \left(\sum_{i=1}^{k} \sum_{j=1}^{m} u_{ij} \right), \tag{37}$$

[3] We generously bound $\frac{n-m}{n-j} \le 1$ in this step. By keeping the corresponding factor in the analysis one obtains that the constant B in the theorem can be improved at least by a factor of $\frac{(n-m)^2}{(n-0.5)(n-m-0.5)}$.

where the function F takes as input an L-dimensional vector and returns the value of its maximum component. We now bound $\mathbb{E}_Z \Phi_t(\mathbf{Z})$ by applying Lemma 3 k times:

$$\mathop{\mathbb{E}}_{Z} \Phi_t(\mathbf{Z}) = \mathop{\mathbb{E}}_{U_1,\dots,U_k} F\Big(\sum_{i=1}^{k}\sum_{j=1}^{m} u_{ij}\Big) \tag{38}$$

$$= \mathop{\mathbb{E}}_{U_1,\dots,U_{k-1}} \left[\mathop{\mathbb{E}}_{U_k}\left[F\Big(\sum_{i=1}^{k-1}\sum_{j=1}^{m} u_{ij} + \sum_{j=1}^{m} u_{kj}\Big)\Big| U_1,\dots,U_{k-1}\right]\right] \tag{39}$$

By Lemma 4 $F(x)$ is a convex function. Thus $F(const + x)$ is also convex and we can apply Lemma 3 with respect to U_k.

$$\leq \mathop{\mathbb{E}}_{U_1,\dots,U_{k-1}} \left[\mathop{\mathbb{E}}_{\hat{U}_k}\left[F\left(\sum_{i=1}^{k-1}\sum_{j=1}^{m} u_{ij} + \sum_{j=1}^{m} \hat{u}_{kj}\right)\Big| U_1,\dots,U_{k-1}\right]\right] \tag{40}$$

where $\hat{U}_k = \{u_{ki},\dots,u_{km}\}$ is a set of m vectors sampled from V_k^t *with replacement*.

$$= \mathop{\mathbb{E}}_{U_1,\dots,U_{k-1},\hat{U}_k} \left[F\left(\sum_{i=1}^{k-1}\sum_{j=1}^{m} u_{ij} + \sum_{j=1}^{m} \hat{u}_{kj}\right)\right]. \tag{41}$$

Repeating the process k times, we obtain

$$\leq \cdots \leq \mathop{\mathbb{E}}_{\hat{U}_1,\dots,\hat{U}_k} \left[F\left(\sum_{i=1}^{k}\sum_{j=1}^{m} \hat{u}_{ij}\right)\right]. \tag{42}$$

Note that writing the conditioning in the above expressions is just for clarity of presentation, since the $U_1,\dots,U_k$ are actually independent of each other.

Switching from the U sets by the $\hat{U}$ sets in Φ corresponds to switching from random subsets S_i^l to random sets $\tilde{S}_i$ consisting of m points sampled from S_i^u uniformly *with* replacement. Therefore we obtain

$$\mathop{\mathbb{E}}_{Z} \Phi_t(Z) = \mathop{\mathbb{E}}_{S_1^l,\dots,S_k^l} \Phi_t(S_1^l,\dots,S_k^l) \leq \mathop{\mathbb{E}}_{\tilde{S}_1,\dots,\tilde{S}_k} \Phi_t(\tilde{S}_1,\dots,\tilde{S}_k), \tag{43}$$

which allows us to continue analyzing $\mathbb{E}_Z \Phi_t(\mathbf{Z})$ in the standard way using Rademacher complexities and independent samples. Applying the common symmetrization trick and introducing Rademacher random variables σ_{ij} we obtain

$$\Phi_t(\tilde{S}_1,\dots,\tilde{S}_k) \leq 2\mathop{\mathbb{E}}_{\sigma}\sup_{h}\sum_{i=1}^{k}\sum_{j=1}^{m}\sigma_{ij}\alpha_i^t\ell(h(x_j^i),y_j^i).$$

We can rewrite this using the fact that $\ell(y,y') = [\![y \neq y']\!] = \frac{1-yy'}{2}$:

$$\mathop{\mathbb{E}}_{\sigma}\sup_{h}\sum_{i=1}^{k}\sum_{j=1}^{m}\sigma_{ij}\alpha_i^t\ell(h(x_j^i),y_j^i) = \mathop{\mathbb{E}}_{\sigma}\sup_{h}\sum_{i=1}^{k}\sum_{j=1}^{m}\sigma_{ij}\alpha_i^t\frac{1-h(x_j^i)y_j^i}{2} = \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\sup_{h}\sum_{i=1}^{k}\sum_{j=1}^{m}-\sigma_{ij}y_j^i\alpha_i^t h(x_j^i)$$

Since $-\sigma_{ij}y_j^i$ has the same distribution as σ_{ij}:

$$= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\sup_{a(h)\in A}\sum_{i=1}^{k}\sum_{j=1}^{m}\sigma_{ij}a_{ij}(h),$$

where $a_{ij}(h) = \alpha_i^t h(x_j^i)$ and $A = \{a(h) : \ h \in \mathcal{H}\}$. According to Sauer's lemma (Corollary 3.3 in (Mohri et al., 2012)):

$$|A| \leq \left(\frac{ekm}{d}\right)^d. \tag{44}$$

At the same time:

$$\|a\|_2 = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m}(\alpha_i^t h(x_j^i))^2} = \sqrt{m}\sqrt{\sum_{i=1}^{k}(\alpha_i^t)^2}. \tag{45}$$

Therefore, by Massart's lemma (Theorem 3.3 in (Mohri et al., 2012)):

$$\mathop{\mathbb{E}}_{\sigma}\sup_{h}\sum_{i=1}^{k}\sum_{j=1}^{m}\sigma_{ij}\alpha_i^t\ell(h(x_j^i), y_j^i) \leq \frac{1}{2}\sqrt{\sum_{i=1}^{k}(\alpha_i^t)^2}\cdot\sqrt{2dm\log(ekm/d)}. \tag{46}$$

By applying this result for all t we obtain:

$$\mathop{\mathbb{E}}_{\mathbf{Z}}\Phi(\mathbf{Z}) = \frac{1}{Tm}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathbf{Z}}\Phi_t(\mathbf{Z}) \leq \frac{1}{Tm}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\tilde{S}}\Phi_t(\tilde{S}) \leq \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{i=1}^{k}(\alpha_i^t)^2}\cdot\sqrt{\frac{2d\log(ekm/d)}{m}}. \tag{47}$$

Combining (35) and (47) with Lemma 2 we obtain that for fixed unlabeled samples $S_1^u, \ldots, S_T^u$ with probability at least $1 - \delta/4$ for all choices of $h_1, \ldots, h_T$:

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathrm{er}}_{\alpha^t}(h_t) \leq \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{er}}_{\alpha^t}(h_t) + \frac{1}{T}\|\alpha\|_{2,1}\sqrt{\frac{2d\log(ekm/d)}{m}} + \frac{1}{T}\|\alpha\|_{1,2}\sqrt{\frac{\log(4/\delta)}{2m}}.$$

By further combining it with (27) we obtain that the following inequality holds uniformly in $h_1, \ldots, h_T \in \mathcal{H}$ with probability at least $1 - \delta/2$ over the sampling of the unlabeled training sets, $S_1^u, \ldots, S_T^u$, and labeled training sets, $(S_i^l)_{i\in I}$, provided that the subset of labeled tasks, $I \subset \{1, \ldots, T\}$, and the task weights, $\alpha^1, \ldots, \alpha^T \in \Lambda^I$, depend deterministically on the unlabeled training only.

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathrm{er}_{\alpha^t}(h_t) \leq \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{er}}_{\alpha^t}(h_t) + \frac{1}{T}\|\alpha\|_{2,1}\sqrt{\frac{2d\log(ekm/d)}{m}} + \frac{1}{T}\|\alpha\|_{1,2}\sqrt{\frac{\log(4/\delta)}{2m}} \\ + \sqrt{\frac{8(\log T + d\log(enT/d))}{n}} + \sqrt{\frac{2}{n}\log\frac{4}{\delta}}.\end{aligned} \tag{48}$$

The statement of Theorem 1 follows by combining (9) with (10) and (48).

References

Ben-David, Shai, Blitzer, John, Crammer, Koby, Kulesza, Alex, Pereira, Fernando, and Vaughan, Jennifer Wortman. A theory of learning from different domains. *Machine Learning*, 2010.

Doob, Joseph L. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3), 1940.

Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 1963.

Maurer, Andreas. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 2006.

McDiarmid, Colin. On the method of bounded differences. In *Surveys in Combinatorics*, 1989.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet. *Foundations of Machine Learning*. The MIT Press, 2012.

Tolstikhin, I., Blanchard, G., and Kloft, M. Localized complexities for transductive learning. In *Workshop on Computational Learning Theory (COLT)*, 2014.