Appendix

In this appendix, we present the proof of Theorems 1 and 2. We begin by introducing some more notation and auxiliary results.

A. Notation and tools

Issues of measurability will be ignored throughout, in particular, if $\mathcal{F}$ is a class of real valued functions on a domain $\mathcal{X}$ and X a random variable with values in $\mathcal{X}$ then we will always write $\mathbb{E}\sup_{f\in\mathcal{F}} f(X)$ to mean $\sup\{\mathbb{E}\max_{f\in\mathcal{F}_0} f(X) : \mathcal{F}_0 \subseteq \mathcal{F},\ \mathcal{F}_0 \text{ finite}\}$.

In the sequel H denotes a finite or infinite dimensional Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. If T is a bounded linear operator on H its operator norm is written $\|T\|_\infty = \sup\{\|Tx\| : \|x\| = 1\}$.

Members of H are denoted with lower case italics such as x, v, w, vectors composed of such vectors are in bold lower case, i.e. $\mathbf{x} = (x_1, \ldots, x_m)$ or $\mathbf{v} = (v_1, \ldots, v_n)$, where m or n are explained in the context.

Let B be the unit ball in H. An *example* is a pair $z = (x, y) \in B \times \mathbb{R} =: \mathcal{Z}$, a sample is a vector of such pairs $\mathbf{z} = (z_1, \ldots, z_m) = ((x_1, y_1), \ldots, (x_m, y_m))$. Here we also write $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, with $\mathbf{x} = (x_1, \ldots, x_m) \in H^m$ and $\mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$.

A multisample is a vector $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$ composed of samples. We also write $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ with $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$.

For members of $\mathbb{R}^K$ we use the greek letters γ or β. Depending on context the inner product and euclidean norm on $\mathbb{R}^K$ will also be denoted with $\langle\cdot,\cdot\rangle$ and $\|.\|$. The ℓ_1-norm $\|\cdot\|_1$ on $\mathbb{R}^K$ is defined by $\|\beta\|_1 = \sum_{k=1}^K |\gamma_k|$.

In the sequel we denote with $\mathcal{C}_\alpha$ the set $\left\{\beta \in \mathbb{R}^K : \|\beta\|_1 \le \alpha\right\}$, abbreviate $\mathcal{C}$ for the ℓ_1-unit ball $\mathcal{C}_1$. The canonical basis of $\mathbb{R}^K$ is denoted $e_1, \ldots, e_K$. Unless otherwise specified the summation over he index i will always run from 1 to m, t will run from 1 to T, and k will run from 1 to K.

A.1. Covariances

For $\mathbf{x} \in H^m$ the empirical covariance operator $\hat{\Sigma}(\mathbf{x})$ is specified by

$$\left\langle \hat{\Sigma}(\mathbf{x})\, v, w \right\rangle = \frac{1}{m}\sum_i \langle v, x_i\rangle \langle x_i, w\rangle,\ v, w \in H.$$

The definition implies the inequality

$$\sum_i \langle v, x_i\rangle^2 = m\left\langle \hat{\Sigma}(\mathbf{x})\, v, v\right\rangle \le m\left\|\hat{\Sigma}(\mathbf{x})\right\|_\infty \|v\|^2. \tag{6}$$

It also follows that $\operatorname{tr}\left(\hat{\Sigma}(\mathbf{x})\right) = (1/m)\sum_i \|x_i\|^2$.

For a multisample $\mathbf{X} \in H^{mT}$ we will consider two quantities defined in terms of the empirical covariances.

$$\begin{aligned} S_1(\mathbf{X}) &= \frac{1}{T}\sum_t \left\|\hat{\Sigma}(\mathbf{x}_t)\right\|_1 := \frac{1}{T}\sum_t \operatorname{tr}\left(\hat{\Sigma}(\mathbf{x}_t)\right) \\ S_\infty(\mathbf{X}) &= \frac{1}{T}\sum_t \left\|\hat{\Sigma}(\mathbf{x}_t)\right\|_\infty := \frac{1}{T}\sum_t \lambda_{\max}\left(\hat{\Sigma}(\mathbf{x}_t)\right) \end{aligned}$$

where $\lambda_{\max}$ is the largest eigenvalue. If all data points x_{ti} lie in the unit ball of H then $S_1(\mathbf{X}) \le 1$. Of course $S_1(\mathbf{X})$ can also be written as the trace of the total covariance $(1/T)\sum_t \hat{\Sigma}(\mathbf{x}_t)$, while $S_\infty(\mathbf{X})$ will always be at least as large as the largest eigenvalue of the total covariance. We always have $S_\infty(\mathbf{X}) \le S_1(\mathbf{X})$, with equality only if the data is one-dimensional for all tasks. The quotient $S_1(\mathbf{X})/S_\infty(\mathbf{X})$ can be regarded as a crude measure of the effective dimensionality of the data. If the data have a high dimensional distribution for each task then $S_\infty(\mathbf{X})$ can be considerably smaller than $S_1(\mathbf{X})$.

A.2. Concentration inequalities

Let $\mathcal{X}$ be any space. For $\mathbf{x} \in \mathcal{X}^n$, $1 \le k \le n$ and $y \in \mathcal{X}$ we use $\mathbf{x}_{k\leftarrow y}$ to denote the object obtained from $\mathbf{x}$ by replacing the k-th coordinate of $\mathbf{x}$ with y. That is

$$\mathbf{x}_{k\leftarrow y} = (x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n).$$

The concentration inequality in part (i) of the following theorem, known as the bounded difference inequality is given in (McDiarmid, 1998). A proof of inequality (ii) is given in (Maurer, 2006).

Theorem 3. *Let $F : \mathcal{X}^n \to \mathbb{R}$ and define A and B by*

$$\begin{aligned} A^2 &= \sup_{\mathbf{x}\in\mathcal{X}^n}\sum_{k=1}^n \sup_{y_1,y_2\in\mathcal{X}} \left(F(\mathbf{x}_{k\leftarrow y_1}) - F(\mathbf{x}_{k\leftarrow y_2})\right)^2 \\ B^2 &= \sup_{\mathbf{x}\in\mathcal{X}^n}\sum_{k=1}^n \left(F(\mathbf{x}) - \inf_{y\in\mathcal{X}} F(\mathbf{x}_{k\leftarrow y})\right)^2. \end{aligned}$$

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables with values in $\mathcal{X}$, and let $\mathbf{X}'$ be i.i.d. to $\mathbf{X}$. Then for any $s > 0$

(i) $\Pr\{F(\mathbf{X}) > \mathbb{E}F(\mathbf{X}') + s\} \le e^{-2s^2/A^2}$;

(ii) $\Pr\{F(\mathbf{X}) > \mathbb{E}F(\mathbf{X}') + s\} \le e^{-s^2/(2B^2)}$.

A.3. Rademacher and Gaussian averages

We will use the term *Rademacher variables* for any set of independent random variables, uniformly distributed on $\{-1,1\}$, and reserve the symbol σ for Rademacher variables. A set of random variables is called *orthogaussian* if the members are independent $\mathcal{N}(0,1)$-distributed (standard normal) variables and reserve the letter ζ for standard normal variables. Thus $\sigma_1, \sigma_2, \ldots, \sigma_i, \ldots, \sigma_{11}, \ldots, \sigma_{ij}$ etc. will always be independent Rademacher variables and $\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_{11}, \ldots, \zeta_{ij}$ will always be orthogaussian.

For $A \subseteq \mathbb{R}^n$ we define the Rademacher and Gaussian averages of A (Ledoux & Talagrand, 1991; Bartlett & Mendelson, 2002) as

$$\begin{aligned}\mathcal{R}(A) &= \mathbb{E}_\sigma \sup_{(x_1,\ldots,x_n)\in A} \frac{2}{n}\sum_{i=1}^n \sigma_i x_i,\\ \mathcal{G}(A) &= \mathbb{E}_\zeta \sup_{(x_1,\ldots,x_n)\in A} \frac{2}{n}\sum_{i=1}^n \zeta_i x_i.\end{aligned}$$

If $\mathcal{F}$ is a class of real valued functions on a space $\mathcal{X}$ and $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ we write

$$\begin{aligned}&\mathcal{F}(\mathbf{x}) = \mathcal{F}(x_1, \ldots, x_n)\\ &= \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n.\end{aligned}$$

The empirical Rademacher and Gaussian complexities of $\mathcal{F}$ on $\mathbf{x}$ are respectively $\mathcal{R}(\mathcal{F}(\mathbf{x}))$ and $\mathcal{G}(\mathcal{F}(\mathbf{x}))$.

The utility of these concepts for learning theory comes from the following key-result (see (Bartlett & Mendelson, 2002; Koltchinskii & Panchenko, 2002)), stated here in two portions for convenience in the sequel.

Theorem 4. *Let $\mathcal{F}$ be a real-valued function class on a space $\mathcal{X}$ and $\mu_1, \ldots, \mu_m$ be probability measures on $\mathcal{X}$ with product measure $\boldsymbol{\mu} = \prod_i \mu_i$ on $\mathcal{X}^m$. For $\mathbf{x} \in \mathcal{X}^m$ define*

$$\Phi(\mathbf{x}) = \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \left(\mathbb{E}_{x\sim\mu_i}[f(x)] - f(x_i)\right).$$

Then $\mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}}[\Phi(\mathbf{x})] \le \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}}\mathcal{R}(\mathcal{F}(\mathbf{x}))$.

Proof. For any realization $\sigma = \sigma_1, \ldots, \sigma_m$ of the Rademacher variables

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}}[\Phi(\mathbf{x})]\\ &= \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}} \sup_{f\in\mathcal{F}} \frac{1}{m}\mathbb{E}_{\mathbf{x}'\sim\boldsymbol{\mu}} \sum_{i=1}^m (f(x_i') - f(x_i))\\ &\le \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\boldsymbol{\mu}\times\boldsymbol{\mu}} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i (f(x_i') - f(x_i)),\end{aligned}$$

because of the symmetry of the measure $\boldsymbol{\mu}\times\boldsymbol{\mu}(\mathbf{x},\mathbf{x}') = \prod_i \mu_i \times \prod_i \mu_i(\mathbf{x},\mathbf{x}')$ under the interchange $x_i \leftrightarrow x_i'$. Taking the expectation in σ and applying the triangle inequality gives the result. □

Theorem 5. *Let $\mathcal{F}$ be a $[0,1]$-valued function class on a space $\mathcal{X}$, and $\boldsymbol{\mu}$ as above. For $\delta > 0$ we have with probability greater than $1-\delta$ in the sample $\mathbf{x} \sim \boldsymbol{\mu}$ that for all $f \in \mathcal{F}$*

$$\mathbb{E}_{x\sim\boldsymbol{\mu}}[f(x)] \le \frac{1}{m}\sum_{i=1}^m f(x_i) + \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}}\mathcal{R}(\mathcal{F}(\mathbf{x})) + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

To prove this apply the bounded-difference inequality (part (i) of Theorem 3) to the function Φ of the previous theorem (see e.g. (Bartlett & Mendelson, 2002)). Under the conditions of this result, changing one of the x_i will not change $\mathcal{R}(\mathcal{F}(\mathbf{x}))$ by more than 2, so again by the bounded difference inequality applied to $\mathcal{R}(\mathcal{F}(\mathbf{x}))$ and a union bound we obtain the data dependent version

Corollary 6. *Let $\mathcal{F}$ and $\boldsymbol{\mu}$ be as above. For $\delta > 0$ we have with probability greater than $1-\delta$ in the sample $\mathbf{x} \sim \boldsymbol{\mu}$ that for all $f \in \mathcal{F}$*

$$\mathbb{E}_{x\sim\boldsymbol{\mu}}[f(x)] \le \frac{1}{m}\sum_{i=1}^m f(x_i) + \mathcal{R}(\mathcal{F}(\mathbf{x})) + \sqrt{\frac{9\ln(2/\delta)}{2m}}.$$

To bound Rademacher averages the following result is very useful (Bartlett & Mendelson, 2002; Ando & Zhang, 2005; Ledoux & Talagrand, 1991)

Lemma 7. *Let $A \subseteq \mathbb{R}^n$, and let $\psi_1, \ldots, \psi_n$ be real functions such that $\psi_i(s) - \psi_i(t) \le L|s-t|, \forall i$, and $s,t \in \mathbb{R}$. Define $\boldsymbol{\psi}(A) = \{\psi_1(x_1), \ldots, \psi_n(x_n) : (x_1, \ldots, x_n) \in A\}$. Then*

$$\mathcal{R}(\boldsymbol{\psi}(A)) \le L\mathcal{R}(A).$$

Sometimes it is more convenient to work with gaussian averages which can be used instead, by virtue of the next lemma. For a proof see e.g. (Ledoux & Talagrand, 1991)

Lemma 8. *For $A \subseteq \mathbb{R}^k$ we have $\mathcal{R}(A) \le \sqrt{\pi/2}\,\mathcal{G}(A)$.*

The next result is known as Slepian's lemma ((Slepian, 1962), (Ledoux & Talagrand, 1991)).

Theorem 9. *Let Ω and Ξ be mean zero, separable Gaussian processes indexed by a common set $\mathcal{S}$, such that*

$$\mathbb{E}(\Omega_{s_1} - \Omega_{s_2})^2 \le \mathbb{E}(\Xi_{s_1} - \Xi_{s_2})^2 \text{ for all } s_1, s_2 \in \mathcal{S}.$$

Then

$$\mathbb{E}\sup_{s\in\mathcal{S}} \Omega_s \le \mathbb{E}\sup_{s\in\mathcal{S}} \Xi_s.$$

B. Proofs

B.1. Multitask learning

In this section we prove Theorem 1. It is an immediate consequence of Hoeffding's inequality and the following uniform bound on the estimation error.

Theorem 10. *Let $\delta > 0$, fix K and let $\mu_1, \ldots, \mu_T$ be probability measures on $H \times \mathbb{R}$. With probability at least $1-\delta$ in the draw of $\mathbf{Z} \sim \prod_{t=1}^{T} \mu_t$ we have for all $D \in \mathcal{D}_K$ and all $\boldsymbol{\gamma} \in \mathcal{C}_\alpha^T$ that*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(x,y)\sim\mu_t}\left[\ell\left(\langle D\gamma_t, x\rangle, y\right)\right] \\
&\qquad - \frac{1}{mT}\sum_{t=1}^{T}\sum_{i=1}^{m} \ell\left(\langle D\gamma_t, x_{ti}\rangle, y_{ti}\right) \\
&\leq L\alpha\sqrt{\frac{2S_1(\mathbf{X})(K+12)}{mT}} \\
&\quad + \quad L\alpha\sqrt{\frac{8S_\infty(\mathbf{X})\ln(2K)}{m}} \quad + \quad \sqrt{\frac{9\ln 2/\delta}{2mT}}.
\end{aligned}$$

The proof of this theorem requires auxiliary results. Fix $\mathbf{X} \in H^{mT}$ and for $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T) \in \left(\mathbb{R}^K\right)^T$ define the random variable

$$F_{\boldsymbol{\gamma}} = F_{\boldsymbol{\gamma}}(\boldsymbol{\sigma}) = \sup_{D\in\mathcal{D}_K} \sum_{t,i} \sigma_{ti}\langle D\gamma_t, x_{ti}\rangle. \tag{7}$$

Lemma 11. *(i) If $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$ satisfies $\|\gamma_t\| \leq 1$ for all t, then*

$$\mathbb{E}F_{\boldsymbol{\gamma}} \leq \sqrt{mTK\ S_1(\mathbf{X})}.$$

(ii) If $\boldsymbol{\gamma}$ satisfies $\|\gamma_t\|_1 \leq 1$ for all t, then for any $s \geq 0$

$$\Pr\left\{F_{\boldsymbol{\gamma}} \geq \mathbb{E}\left[F_{\boldsymbol{\gamma}}\right] + s\right\} \leq \exp\left(\frac{-s^2}{8mT\ S_\infty(\mathbf{X})}\right).$$

Proof. (i) We observe that

$$\begin{aligned}
\mathbb{E}F_{\boldsymbol{\gamma}} &= \mathbb{E}\sup_D \sum_k \left\langle De_k, \sum_{t,i}\sigma_{ti}\gamma_{tk}x_{ti}\right\rangle \\
&\leq \sup_D\left(\sum_k \|De_k\|^2\right)^{1/2} \mathbb{E}\left(\sum_k\left\|\sum_{t,i}\sigma_{ti}\gamma_{tk}x_{ti}\right\|^2\right)^{1/2} \\
&\leq \sqrt{K}\left(\sum_k \mathbb{E}\left\|\sum_{t,i}\sigma_{ti}\gamma_{tk}x_{ti}\right\|^2\right)^{1/2} \\
&= \sqrt{K}\left(\sum_{k,t,i}|\gamma_{tk}|^2\|x_{ti}\|^2\right)^{1/2} \\
&= \sqrt{K}\left(\sum_t\left(\sum_k|\gamma_{tk}|^2\right)\sum_i\|x_{ti}\|^2\right)^{1/2} \\
&\leq \sqrt{K\sum_{t,i}\|x_{ti}\|^2} = \sqrt{mTK\ S_1(\mathbf{X})}.
\end{aligned}$$

(ii) For any configuration $\boldsymbol{\sigma}$ of the Rademacher variables let $D(\boldsymbol{\sigma})$ be the maximizer in the definition of $F_{\boldsymbol{\gamma}}(\boldsymbol{\sigma})$. Then for any $s \in \{1, \ldots, T\}$, $j \in \{1, \ldots, m\}$ and any $\sigma' \in \{-1, 1\}$ to replace σ_{sj} we have

$$F_{\boldsymbol{\gamma}}(\boldsymbol{\sigma}) - F_{\boldsymbol{\gamma}}\left(\boldsymbol{\sigma}_{(sj)\leftarrow\sigma'}\right) \leq 2\left|\langle D(\boldsymbol{\sigma})\gamma_s, x_{sj}\rangle\right|.$$

Using the inequality (6) we then obtain

$$\begin{aligned}
&\textstyle\sum_{sj}\left(F_\gamma(\boldsymbol{\sigma}) - \inf_{\sigma'\in\{-1,1\}} F_{\boldsymbol{\gamma}}\left(\boldsymbol{\sigma}_{(sj)\leftarrow\sigma'}\right)\right)^2 \\
&\qquad \leq \quad 4\sum_{t,i}\langle D(\boldsymbol{\sigma})\gamma_t, x_{ti}\rangle^2 \\
&\qquad \leq \quad 4m\sum_t\left\|\hat{\Sigma}(\mathbf{x}_t)\right\|_\infty \|D(\boldsymbol{\sigma})\gamma_t\|^2 \\
&\qquad \leq \quad 4m\sum_t\left\|\hat{\Sigma}(\mathbf{x}_t)\right\|_\infty.
\end{aligned}$$

In the last inequality we used the fact that for any $D \in \mathcal{D}_K$ we have $\|D\gamma_t\| \leq \sum_k|\gamma_{tk}|\,\|De_k\| \leq \|\gamma_t\|_1 \leq 1$. The conclusion now follows from part (ii) of Theorem 3. □

Proposition 12. *We have for every fixed $\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \in (H \times \mathbb{R})^{mT}$ we have*

$$\begin{aligned}
&\mathbb{E}_\sigma \sup_{D\in D, \boldsymbol{\gamma}\in(\mathcal{C}_\alpha)^T} \textstyle\sum_{t,i}\sigma_{it}\ell\left(\langle D\gamma_t, x_{ti}\rangle, y_{ti}\right) \\
&\leq L\alpha\sqrt{2mTS_1(\mathbf{X})(K+12)} + L\alpha T\sqrt{8mS_\infty(\mathbf{X})\ln(2K)}.
\end{aligned}$$

Proof. It suffices to prove the result for $\alpha = 1$, the general result being a consequence of rescaling. By Lemma 7 and the Lipschitz properties of the loss function ℓ we have

$$\mathbb{E}_\sigma \sup_{D\in\mathcal{D}_K,\boldsymbol{\gamma}\in(\mathcal{C})^T,} \sum_{t,i}\sigma_{it}\ell\left(\langle D\gamma_t, x_{ti}\rangle, y_{ti}\right)$$

$$\leq L\mathbb{E}_\sigma \sup_{D\in\mathcal{D}_K,\boldsymbol{\gamma}\in(\mathcal{C})^T,} \sum_{t,i}\sigma_{it}\langle D\gamma_t, x_{ti}\rangle . \qquad (8)$$

Since linear functions on a compact convex set attain their maxima at the extreme points, we have

$$\mathbb{E}\sup_{D\in\mathcal{D}_K,\boldsymbol{\gamma}\in(\mathcal{C})^T,} \sum_{t=1}^{T}\sum_{i=1}^{m}\sigma_{it}\langle D\gamma_t, x_{ti}\rangle = \mathbb{E}\max_{\boldsymbol{\gamma}\in\mathrm{ext}(\mathcal{C})^T} F_{\boldsymbol{\gamma}}, \qquad (9)$$

where $F_{\boldsymbol{\gamma}}$ is defined as in (7). Let $c=\sqrt{mKTS_1(\mathbf{X})}$. Now for any $\delta\geq 0$ we have, since $F_{\boldsymbol{\gamma}}\geq 0$,

$$\begin{aligned}
\mathbb{E}\max_{\boldsymbol{\gamma}\in\mathrm{ext}(\mathcal{C})^T} F_{\boldsymbol{\gamma}} &= \int_0^\infty \Pr\left\{\max_{\boldsymbol{\gamma}\in\mathrm{ext}(\mathcal{C})^T} F_{\boldsymbol{\gamma}} > s\right\} ds \\
&\leq c+\delta+\sum_{\boldsymbol{\gamma}\in(\mathrm{ext}(\mathcal{C}))^T}\int_{\sqrt{mKTS_1(\mathbf{X})}+\delta}^\infty \Pr\left\{F_{\boldsymbol{\gamma}}>s\right\} ds \\
&\leq c+\delta+\sum_{\boldsymbol{\gamma}\in(\mathrm{ext}(\mathcal{C}))^T}\int_{\delta}^\infty \Pr\left\{F_{\boldsymbol{\gamma}}>\mathbb{E}F_{\boldsymbol{\gamma}}+s\right\} ds \\
&\leq c+\delta+(2K)^T\int_\delta^\infty \exp\left(\frac{-s^2}{8mTS_\infty(\mathbf{X})}\right) ds \\
&\leq c+\delta+\frac{4mTS_\infty(\mathbf{X})(2K)^T}{\delta}\exp\left(\frac{-\delta^2}{8mTS_\infty(\mathbf{X})}\right).
\end{aligned}$$

Here the first inequality follows from the fact that probabilities never exceed 1 and a union bound. The second inequality follows from Lemma 11, part (i), since $\mathbb{E}F_{\mathbf{k}}\leq\sqrt{mKTS_1(\mathbf{X})}$. The third inequality follows from Lemma 11, part (ii), and the fact that the cardinality of $\mathrm{ext}(\mathcal{C})$ is $2K$, and the last inequality follows from a well known estimate on Gaussian random variables. Setting $\delta=\sqrt{8mTS_\infty(\mathbf{X})\ln\left(e(2K)^T\right)}$ we obtain with some easy simplifying estimates

$$\mathbb{E}\max_{\boldsymbol{\gamma}\in\mathrm{ext}(\mathcal{C})^T} F_{\boldsymbol{\gamma}} \leq \sqrt{2mT(K+12)S_1(\mathbf{X})} + T\sqrt{8mS_\infty(\mathbf{X})\ln(2K)},$$

which together with (8) and (9) gives the result. □

Theorem 10 now follows from Corollary 6.

If the set $\mathcal{C}_\alpha$ is replaced by any other subset $\mathcal{C}'$ of the ℓ_2-ball of radius α, a similar proof strategy can be employed. The denominator in the exponent of Lemma 11-(ii) then obtains another factor of $\sqrt{K}$. The union bound over the extreme points in $\mathrm{ext}(\mathcal{C})$ in the previous proposition can be replaced by a union bound over a cover $\mathcal{C}'$. This leads to the alternative result mentioned in Remark 5 following the statement of Theorem 1.

Another modification leads to a bound for the method presented in (Kumar & Daumé III, 2012), where the constraint $\|De_k\|\leq 1$ is replaced by $\|D\|_2\leq\sqrt{K}$ (here $\|\cdot\|_2$ is the Frobenius or Hilbert Schmidt norm) and the constraint $\|\gamma_t\|_1\leq\alpha,\forall t$ is replaced by $\sum\|\gamma_t\|_1\leq\alpha T$. To explain the modification we set $\alpha=1$. Part (i) of Lemma 11 is easily verified. The union bound over $(\mathrm{ext}(\mathcal{C}))^T$ in the previous proposition is replaced by a union bound over the $2TK$ extreme points of the ℓ_1-Ball of radius T in $\mathbb{R}^{TK}$. For part (ii) we use the fact that the concentration result is only needed for $\boldsymbol{\gamma}$ being an extremepoint (so that it involves only a single task) and obtain the bound $\sum_t\left\|\hat{\Sigma}(\mathbf{x}_t)\right\|_\infty\|D\gamma_t\|^2\leq TKS'_\infty(\mathbf{X})$, leading to

$$\Pr\left\{F_{\boldsymbol{\gamma}}\geq E[F_{\boldsymbol{\gamma}}]+s\right\}\leq\exp\left(\frac{-s^2}{8mTK\ S'_\infty(\mathbf{X})}\right).$$

Proceeding as above we obtain the excess risk bound

$$L\alpha\sqrt{\tfrac{2S_1(\mathbf{X})(K+12)}{mT}}+L\alpha\sqrt{\tfrac{8KS'_\infty(\mathbf{X})\ln(2KT)}{m}}+\sqrt{\frac{8\ln 4/\delta}{mT}},$$

to replace the bound in Theorem 1. The factor $\sqrt{K}$ in the second term seems quite weak, but it must be borne in mind that the constraint $\|D\|_2\leq\sqrt{K}$ is much weaker than $\|De_k\|\leq 1$, and allows for a smaller approximation error. If we retain $\|De_k\|\leq 1$ and only modify the γ-constraint to $\sum\|\gamma_t\|_1\leq\alpha T$ the $\sqrt{K}$ in the second term disappears and by comparison to Theorem 1 there is only and additional $\ln T$ and the switch from $S_\infty(\mathbf{X})$ to $S'_\infty(\mathbf{X})$, reflecting the fact that $\sum\|\gamma_t\|_1\leq\alpha T$ is a much weaker constraint than $\|\gamma_t\|_1\leq\alpha,\forall t$, so that, again, a smaller minimum in (1) is possible for the modified method.

B.2. Learning to learn

In this section we prove Theorem 2. The basic strategy is as follows. Recall the definition (4) of the measure $\rho_{\mathcal{E}}$, which governs the generation of a training sample in the environment $\mathcal{E}$. On a given training sample $\mathbf{z}\sim\rho_{\mathcal{E}}$ the algorithm A_D as defined in (3) incurs the empirical risk

$$\hat{R}_D(\mathbf{z})=\min_{\gamma\in\mathcal{C}_\alpha}\frac{1}{m}\sum_{i=1}^{m}\ell\left(\langle D\gamma, x_i\rangle, y_i\right).$$

The algorithm A_D, essentially being the Lasso, has very good estimation properties, so $\hat{R}_D(\mathbf{z})$ will be close to the true risk of A_D in the corresponding task. This means that we only really need to estimate the expected empirical risk $\mathbb{E}_{\mathbf{z}\sim\rho_{\mathcal{E}}}\hat{R}_D(\mathbf{z})$ of A_D on future

tasks. On the other hand the minimization problem (1) can be written as

$$\min_{D\in\mathcal{D}_K}\frac{1}{T}\sum_{t=1}^{T}\hat{R}_D\left(\mathbf{z}_t\right)\text{ with }\mathbf{Z}=\left(z_1,\ldots,z_T\right)\sim\left(\rho_{\mathcal{E}}\right)^T,$$

with dictionary $D\left(\mathbf{Z}\right)$ being the minimizer. If $\mathcal{D}_K$ is not too large this should be similar to $\mathbb{E}_{\mathbf{z}\sim\rho_{\mathcal{E}}}\hat{R}_{D(\mathbf{Z})}\left(\mathbf{z}\right)$. In the sequel we make this precise.

Lemma 13. *For $v\in H$ with $\|v\|\leq 1$ and $\mathbf{x}\in H^m$ let F be the random variable*

$$F=\left|\left\langle v,\sum_i\sigma_i x_i\right\rangle\right|.$$

Then (i) $\mathbb{E}F\leq\sqrt{m}\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty^{1/2}$ and (ii) for $t\geq 0$

$$\Pr\left\{F>\mathbb{E}F+s\right\}\leq\exp\left(\frac{-s^2}{2m\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty}\right).$$

Proof. (i). Using Jensen's inequality and (6) we get

$$\begin{aligned}\mathbb{E}F &\leq \left(\mathbb{E}\left\langle v,\sum_i\sigma_i x_i\right\rangle^2\right)^{1/2}\\ &= \left(\sum_i\left\langle v,x_i\right\rangle^2\right)^{1/2}\leq m\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty.\end{aligned}$$

(ii) Let $\boldsymbol{\sigma}$ be any configuration of the Rademacher variables. For any $\sigma',\sigma''\in\{-1,1\}$ to replace σ_{sj} we have

$$F\left(\boldsymbol{\sigma}_{(sj)\leftarrow\sigma'}\right)-F\left(\boldsymbol{\sigma}_{(sj)\leftarrow\sigma''}\right)\leq 2\left|\left\langle v,x_j\right\rangle\right|,$$

so the conclusion follows from the bounded difference inequality, Theorem 3 (i). □

Lemma 14. *For $v_1,\ldots,v_K\in H$ satisfying $\|v_k\|\leq 1$, $\mathbf{x}\in H^m$ we have*

$$\mathbb{E}\max_k\left|\left\langle v_k,\sum_i\sigma_i x_i\right\rangle\right|\leq\sqrt{2m\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty}\left(2+\sqrt{\ln K}\right).$$

Proof. Let $F_k=\left|\left\langle v_k,\sum_i\sigma_i x_i\right\rangle\right|$. Setting $c=\sqrt{m\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty}$ and using integration by parts we have for $\delta\geq 0$

$$\begin{aligned}&\mathbb{E}\max_k F_k\\ &\leq c+\delta+\int_{\sqrt{m\left\|\hat{\Sigma}(\mathbf{x})\right\|_\infty}+\delta}^{\infty}\max_k\Pr\left\{F_k\geq s\right\}ds\\ &\leq c+\delta+\sum_k\int_\delta^\infty\Pr\left\{F_k\geq\mathbb{E}F_k+s\right\}ds\\ &\leq c+\delta+\sum_k\int_\delta^\infty\exp\left(\frac{-s^2}{2m\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty}\right)ds\\ &\leq c+\delta+\frac{mK\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty}{\delta}\exp\left(\frac{-\delta^2}{2m\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty}\right).\end{aligned}$$

Above the first inequality is trivial, the second follows from Lemma 13 (i) and a union bound, the third inequality follows from Lemma 13 (ii) and the last from a well known approximation. The conclusion follows from substitution of $\delta=\sqrt{2m\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty\ln\left(eK\right)}$. □

Proposition 15. *Let $S_{\mathcal{E}}:=\mathbb{E}_{\tau\sim\mathcal{E}}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu_\tau^m}\left\|\hat{\Sigma}\left(\mathbf{x}\right)\right\|_\infty$. With probability at least $1-\delta$ in the multisample $\mathbf{Z}\sim\rho_{\mathcal{E}}^T$*

$$\begin{aligned}&\sup_{D\in\mathcal{D}_K}R_{\mathcal{E}}\left(A_D\right)-\frac{1}{T}\sum_{t=1}^{T}\hat{R}_D\left(\mathbf{z}_t\right) &(10)\\ \leq\;& L\alpha K\sqrt{\frac{2\pi S_1\left(\mathbf{X}\right)}{T}} &(11)\\ +\;& 4L\alpha\sqrt{\frac{S_\infty\left(\mathcal{E}\right)\left(2+\ln K\right)}{m}}+\sqrt{\frac{\ln 1/\delta}{2T}}.\end{aligned}$$

Proof. Following our strategy we write (abbreviating $\rho=\rho_{\mathcal{E}}$)

$$\begin{aligned}&\sup_{D\in\mathcal{D}_K}R_{\mathcal{E}}\left(A_D\right)-\frac{1}{T}\sum_{t=1}^{T}\hat{R}_D\left(\mathbf{z}_t\right)\\ \leq\;&\sup_{D\in\mathcal{D}_K}\mathbb{E}_{\tau\sim\mathcal{E}}\mathbb{E}_{\mathbf{z}\sim\mu_\tau^m} &(12)\\ &\left[\mathbb{E}_{(x,y)\sim\mu_\tau}\left[\ell\left(\left\langle A_D\left(\mathbf{z}\right),x\right\rangle,y\right)\right]-\hat{R}_D\left(\mathbf{z}\right)\right]\\ +\;&\sup_{D\in\mathcal{D}_K}\mathbb{E}_{\mathbf{z}\sim\rho}\left[\hat{R}_D\left(\mathbf{z}\right)\right]-\frac{1}{T}\sum_{t=1}^{T}\hat{R}_D\left(\mathbf{z}_t\right)\end{aligned}$$

and proceed by bounding each of the two terms in turn.

For any fixed dictionary D and any measure μ on $\mathcal{Z}$

we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}\sim\mu^m}\left[\mathbb{E}_{(x,y)\sim\mu}\left[\ell\left(\left\langle A_D\left(\mathbf{z}\right),x\right\rangle,y\right)\right]-\hat{R}_D\left(\mathbf{z}\right)\right]\\
\leq\ &\mathbb{E}_{\mathbf{z}\sim\mu^m}\sup_{\gamma\in\mathcal{C}_\alpha}\left[\mathbb{E}_{(x,y)\sim\mu}\left[\ell\left(\left\langle D\gamma,x\right\rangle,y\right)\right]\right.\\
&\left.-\frac{1}{m}\sum_{i=1}^m\ell\left(\left\langle D\gamma,x_i\right\rangle,y_i\right)\right]\\
\leq\ &\frac{2}{m}\mathbb{E}_{\mathbf{z}\sim\mu^m}\mathbb{E}_\sigma\sup_{\gamma\in\mathcal{C}_\alpha}\sum_{i=1}^m\sigma_i\ell\left(\left\langle D\gamma,x_i\right\rangle,y_i\right)\quad\text{[Theorem 4]}\\
\leq\ &\frac{2L}{m}\mathbb{E}_{\mathbf{z}\sim\mu^m}\mathbb{E}_\sigma\sup_{\gamma\in\mathcal{C}_\alpha}\sum_k\gamma_k\left\langle De_k,\sum_{i=1}^m\sigma_ix_i\right\rangle\quad\text{[Lemma 7]}\\
\leq\ &\frac{2L\alpha}{m}\mathbb{E}_{\mathbf{z}\sim\mu^m}\mathbb{E}_\sigma\max_k\left|\left\langle De_k,\sum_{i=1}^m\sigma_ix_i\right\rangle\right|\quad\text{[Hölder's ineq.]}\\
\leq\ &\frac{2L\alpha}{m}\mathbb{E}_{\mathbf{z}\sim\mu^m}\sqrt{2m\lambda_{\max}\left(\hat{\Sigma}\left(\mathbf{x}\right)\right)}\left(2+\sqrt{\ln K}\right)\quad\text{[Lemma 13 (i)]}\\
\leq\ &2L\alpha\sqrt{\frac{4\mathbb{E}_{\mathbf{z}\sim\mu^m}\lambda_{\max}\left(\hat{\Sigma}\left(\mathbf{x}\right)\right)\left(2+\ln K\right)}{m}}\quad\text{[Jensen's ineq.].}
\end{aligned}$$

This gives the bound

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}\sim\mu^m}\left[\mathbb{E}_{(x,y)\sim\mu}\left[\ell\left(\left\langle A_D\left(\mathbf{z}\right),x\right\rangle,y\right)\right]-\hat{R}_D\left(\mathbf{z}\right)\right]\\
&\qquad\leq 4L\alpha\sqrt{\frac{\mathbb{E}_{\mathbf{z}\sim\mu^m}\lambda_{\max}\left(\hat{\Sigma}\left(\mathbf{x}\right)\right)\left(2+\ln K\right)}{m}}\qquad(13)
\end{aligned}$$

valid for every measure μ on $H\times\mathbb{R}$ and every $D\in\mathcal{D}_K$. Replacing μ by μ_τ, taking the expectation as $\tau\sim\mathcal{E}$ and using Jensen's inequality bounds the first term on the right hand side of (12) by the second term on the right hand side of (10).

We proceed to bound the second term. From Corollary 6 and Lemma 8 we get that with probability at least $1-\delta$ in $\mathbf{Z}\sim\left(\rho_{\mathcal{E}}\right)^T$

$$\begin{aligned}
&\sup_{D\in\mathcal{D}_K}\mathbb{E}_{\mathbf{z}\sim\rho}\left[\hat{R}_D\left(\mathbf{z}\right)\right]-\tfrac{1}{T}\textstyle\sum_{t=1}^T\hat{R}_D\left(\mathbf{z}_t\right)\\
&\qquad\leq\frac{\sqrt{2\pi}}{T}\mathbb{E}_\zeta\sup_{D\in\mathcal{D}_K}\sum_{t=1}^T\zeta_t\hat{R}_D\left(\mathbf{z}_t\right)+\sqrt{\frac{\ln 1/\delta}{2T}},
\end{aligned}$$

where ζ_t is an orthogaussian sequence. Define two Gaussian processes Ω and Ξ indexed by $\mathcal{D}_K$ as

$$\Omega_D=\textstyle\sum_{t=1}^T\zeta_t\hat{R}_D\left(\mathbf{z}_t\right)$$

and

$$\Xi_D=\tfrac{L\alpha}{\sqrt{m}}\textstyle\sum_{t=1}^T\sum_{i=1}^m\sum_{k=1}^K\zeta_{kij}\left\langle De_k,x_{ti}\right\rangle,$$

where the ζ_{ijk} are also orthogaussian. Then for $D_1,D_2\in\mathcal{D}_K$

$$\begin{aligned}
&\mathbb{E}\left(\Omega_{D_1}-\Omega_{D_2}\right)^2=\\
&=\sum_{t=1}^T\left(\hat{R}_{D_1}\left(\mathbf{z}_t\right)-\hat{R}_{D_2}\left(\mathbf{z}_t\right)\right)^2\\
&\leq\sum_{t=1}^T\left(\sup_{\gamma\in\mathcal{C}_\alpha}\frac{1}{m}\sum_{i=1}^m\ell\left(\left\langle D_1\gamma,x_{ti}\right\rangle,y_{ti}\right)\right.\\
&\quad\left.-\ell\left(\left\langle D_2\gamma,x_{ti}\right\rangle,y_{ti}\right)\right)^2\\
&\leq L^2\sum_{t=1}^T\sup_{\gamma\in\mathcal{C}_\alpha}\left(\frac{1}{m}\sum_{i=1}^m\left\langle\gamma,\left(D_1^*-D_2^*\right)x_{ti}\right\rangle\right)^2\ \text{Lipschitz}\\
&\leq\frac{L^2}{m}\sum_{t=1}^T\sup_{\gamma\in\mathcal{C}_\alpha}\sum_{i=1}^m\left\langle\gamma,\left(D_1^*-D_2^*\right)x_{ti}\right\rangle^2\ \text{Jensen}\\
&\leq\frac{L^2\alpha^2}{m}\sum_{t=1}^T\sum_{i=1}^m\sum_{k=1}^K\left\|\left(D_1^*-D_2^*\right)x_{ti}\right\|^2\ \text{(Cauchy-Schwarz)}\\
&=\frac{L^2\alpha^2}{m}\sum_{t=1}^T\sum_{i=1}^m\sum_{k=1}^K\left(\left\langle D_1e_k,x_{ti}\right\rangle-\left\langle D_2e_k,x_{ti}\right\rangle\right)^2\\
&=\mathbb{E}\left(\Xi_{D_1}-\Xi_{D_2}\right)^2.
\end{aligned}$$

So by Slepian's Lemma

$$\begin{aligned}
&\mathbb{E}\sup_{D\in\mathcal{D}_K}\textstyle\sum_{t=1}^T\zeta_j\hat{R}_D\left(\mathbf{z}_t\right)\\
&=\mathbb{E}\sup_{D\in\mathcal{D}_K}\Omega_D\leq\mathbb{E}\sup_{D\in\mathcal{D}}\Xi_D\\
&=\frac{2\pi}{T}\frac{L\alpha}{\sqrt{m}}\mathbb{E}\sup_{D\in\mathcal{D}_K}\sum_{t=1}^T\sum_{i=1}^m\sum_{k=1}^K\zeta_{kij}\left\langle De_k,x_{ti}\right\rangle\\
&=\frac{L\alpha}{\sqrt{m}}\mathbb{E}\sup_{D\in\mathcal{D}_K}\sum_{k=1}^K\left\langle De_k,\sum_{t=1}^T\sum_{i=1}^m\zeta_{kij}x_{ti}\right\rangle\\
&\leq\frac{L\alpha}{\sqrt{m}}\sup_{D\in\mathcal{D}_K}\left(\sum_k\left\|De_k\right\|^2\right)^{1/2}\\
&\quad\mathbb{E}_\zeta\left(\sum_k\left\|\sum_{t,i}\zeta_{tki}x_{ti}\right\|^2\right)^{1/2}\\
&\leq\frac{L\alpha\sqrt{K}}{\sqrt{m}}\left(\sum_k\mathbb{E}_\zeta\left\|\sum_{t,i}\zeta_{tki}x_{ti}\right\|^2\right)^{1/2}\\
&\leq\frac{L\alpha\sqrt{K}}{\sqrt{m}}\left(\sum_k\sum_{t,i}\left\|x_{ti}\right\|^2\right)^{1/2}\leq L\alpha K\sqrt{mTS_1\left(\mathbf{X}\right)}.
\end{aligned}$$

We therefore have that with probability at least $1-\delta$ in the draw of the multi sample $\mathbf{Z}\sim\rho^T$

$\sup_{D\in\mathcal{D}_K} \mathbb{E}_{\mathbf{z}\sim\rho}\left[\hat{R}_D\left(\mathbf{z}\right)\right] - \frac{1}{T}\sum_{i=1}^{T}\hat{R}_D\left(\mathbf{Z}_{t_j}\right)$

$$\leq L\alpha K\sqrt{\frac{2\pi S_1\left(\mathbf{X}\right)}{\sqrt{T}}} + \sqrt{\frac{9\ln 2/\delta}{2T}}. \tag{14}$$

which in (12) combines with (13) to give the conclusion. □

Proof of Theorem 2. Let D_{opt} and γ_τ the minimizers in the definition of R_{opt}, so that

$$R_{\mathrm{opt}} = \mathbb{E}_{\tau\sim\mathcal{E}}\mathbb{E}_{(x,y)\sim\mu_\tau}\ell\left[\left(\left\langle D_{\mathrm{opt}}\gamma_\tau, x\right\rangle, y\right)\right].$$

$R_{\mathcal{E}}\left(A_{D(\mathbf{Z})}\right) - R_{\mathrm{opt}}$ can be decomposed as the sum of four terms,

$$\left(R_{\mathcal{E}}\left(A_{D(\mathbf{Z})}\right) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D(\mathbf{Z})}\left(\mathbf{z}_t\right)\right) \tag{15}$$

$$+\left(\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D(\mathbf{Z})}\left(\mathbf{z}_t\right) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}_t\right)\right) \tag{16}$$

$$+\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}_t\right) - \mathbb{E}_{\mathbf{z}\sim\boldsymbol{\rho}}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}\right) \tag{17}$$

$$+\mathbb{E}_{\tau\sim\mathcal{E}}\left[\mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}\right)\right.$$

$$\left.-\mathbb{E}_{(x,y)\sim\mu_\tau}\left[\ell\left(\left\langle D_{\mathrm{opt}}\gamma_\tau, x\right\rangle, y\right)\right]\right]. \tag{18}$$

By definition of $\hat{R}$ we have for every τ that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}\right)\\
&= \mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\min_{\gamma\in\mathcal{C}_\alpha}\frac{1}{m}\sum_{i=1}^{m}\ell\left[\left(\left\langle D_{\mathrm{opt}}\gamma, x_i\right\rangle, y_i\right)\right]\\
&\leq \mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\frac{1}{m}\sum_{i=1}^{m}\ell\left[\left(\left\langle D_{\mathrm{opt}}\gamma_\tau, x_i\right\rangle, y_i\right)\right]\\
&= \mathbb{E}_{(x,y)\sim\mu_\tau}\ell\left[\left(\left\langle D_{\mathrm{opt}}\gamma_\tau, x\right\rangle, y\right)\right].
\end{aligned}$$

The term (18) above is therefore non-positive. By Hoeffding's inequality the term (17) is less than $\sqrt{\ln\left(2/\delta\right)/2T}$ with probability at least $1-\delta/2$. The term (16) is non-positive by the definition of $D\left(\mathbf{Z}\right)$. Finally we use Proposition 15 to obtain with probability at least $1-\delta/2$ that

$$\begin{aligned}
&R_{\mathcal{E}}\left(A_{D(\mathbf{Z})}\right) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D(\mathbf{Z})}\left(\mathbf{z}_t\right)\\
&\leq \sup_{D\in\mathcal{D}_K} R_{\mathcal{E}}\left(A_D\right) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_D\left(\mathbf{z}_t\right)\\
&\leq L\alpha K\sqrt{\frac{2\pi S_1\left(\mathbf{X}\right)}{T}}\\
&+ 4L\alpha\sqrt{\frac{S_\infty\left(\mathcal{E}\right)\left(2+\ln K\right)}{m}} + \sqrt{\frac{9\ln 4/\delta}{2T}}.
\end{aligned}$$

Combining these estimates on (15), (16), (17) and (18) in a union bound gives the conclusion. □