Rotation Report
Group Lampert

Contents

Chapter 1

Analysing Response of Retinal Neurons in Fish

1.1 Task

To understand the mechanism how retinal Neurons encode visual stimuli, researchers have tried to record individual Neurons reacting to predefined visual information in controlled experiments. In most experiments the visual stimuli were of synthetic data, mostly moving bars or dots to make the data easier to interpret. In contrast, the data for this analysis stems from real-world stimuli.

In an experiment researchers recorded the response of twenty retinal neurons to a movie showing a natural environment fifty independent experiments. The movie consists of 2141 frames of 100×100 pixels each, where each pixel is binary either white or black. Figure 1.1 shows an example frame from the movie.

For each repetition of the experiment, the neural response data is a binary 2141×20 matrix where the columns correspond to neurons and the rows to time points. As an example, the response pattern of Neuron 14 averaged over all 50 experiments is shown in Figure 1.2.

The task now is to learn the function of individual neurons by comparing the movie data with the response patterns. In order to incorporate the detection of response to movement, the actual comparison will not be performed between the expected spikes and single movie frames, but rather the expected spike and the movie frames from the last, say, 40 frames, the current frame and, for completeness, the 9 next frames. To make the data manageable by standard machine learning algorithms the 50 frames of 100×100 pixel images are reshaped in $50 \cdot 100 \cdot 100 = 500.000$ dimensional vectors $x_0, \dots x_{2091}$ (we forget about the first 40 and the last 9 frames since for those not sufficient past/future information is available). As label data we used regression labels $y_1, \dots, y_{2091} \in [-1, 1]$ resulting from averaging the binary response patterns (which were supplied in a form of ± 1).

1.2 Ridge Regression

Given the high dimensionality of the training data linear regression both has a large enough parameter space and is, at the same time, computationally feasible. Linear regression means that we are looking for weights $w \in \mathbb{R}^{500.000}$ and an offset parameter b such that

$$\langle w, x_i \rangle + b \approx y_i, \quad i \in [2091].$$

Figure 1.1: Frame 100 from the shown video

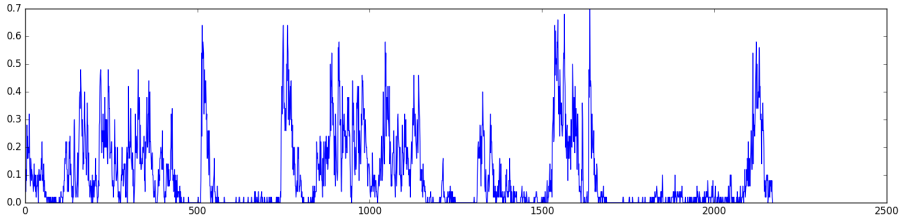

Figure 1.2: Neuron response probability for Neuron 14 as a function of the frame number

To make sure we are not over-fitting the data we introduce a penalty $\lambda > 0$ for large weights. Concretely the Ridge Regression algorithm returns

$$\begin{aligned} w_{\text{ridge}} &= \arg\min_w \left[\sum_{i=1}^{2091} (y_i - \langle x_i, w \rangle - b)^2 + \lambda \sum_{i=1}^{2091} w_i^2 \right] \\ &= \arg\min_w \left[\left\| y - Xw - \tilde{b} \right\|_2^2 + \lambda \left\| w \right\|_2^2 \right] \end{aligned}$$

where X is the 2091×500.000 sample matrix and $\tilde{b} = (b, \ldots, b)$. In general it is difficult to determine a good value for λ. An often used method is k-fold cross validation. This means that the given data set is split in k parts, trained on any selection of $k-1$ of those parts and scored against the remaining part. The average of those k scores then is the *cross-validation-score* for which we optimize. In the concrete case the cross-optimization was run

over a 5-fold split of the 50 experiment runs rather than splitting the 2091 time points in 5 parts which would have the problem of insufficient independence between training and testing data. After choosing the optimal penalty λ we can then refit the whole data set for this particular value of λ and obtain a weight vector w_{ridge} which we reshape into a $50 \times 100 \times 100$ array. High values in this weight matrix indicate a high correlation between a stimulus at the given space-time-coordinate and a Neuron spike.

The optimization objective is convex and differentiable. Thus it admits a global minimum if and only if

$$w = (\lambda + X^T X)^{-1} X (y - \tilde{b}).$$

1.3 Lasso Regression

An alternative to Ridge Regression is the so called *Lasso Regression* which is identical up to the norm of the penalty term. Explicitly,

$$w_{\text{lasso}} = \arg\min_w \left[\left\| y - Xw - \tilde{b} \right\|_2^2 + \lambda \left\| w \right\|_1 \right].$$

The practical difference is that Lasso favours sparse weights and therefore the results can be quite different.

1.4 Canonical Correlation Analysis

A learning algorithm that allows to learn the response from all Neurons simultanuously is the so called *Canonical Correlation Analysis*. The idea is that given, for simplicity centered data, $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ we want to find projections w_x, w_y which maximize the correlation between $(\langle w_x, x_k \rangle)_{k \in [n]}$ and $(\langle w_y, y_k \rangle)_{k \in [n]}$. That is, we want to maximize

$$\frac{\widehat{\mathbf{E}}[\langle w_x, x \rangle \langle w_y, y \rangle]}{\sqrt{\widehat{\mathbf{E}}[\langle w_x, x \rangle^2] \widehat{\mathbf{E}}[\langle w_y, y \rangle^2]}} = \frac{w_x^T \widehat{\mathbf{E}}[xy^T] w_y}{\sqrt{w_x^T \widehat{\mathbf{E}}[xx^T] w_x w_y^T \widehat{\mathbf{E}}[yy^T] w_y}} = \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}}$$

where $\widehat{\mathbf{E}}$ denotes the empirical expection over $(x, y) \in \{ (x_k, y_k) \mid k \in [n] \}$ and

$$C = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{E}}[xx^T] & \widehat{\mathbf{E}}[xy^T] \\ \widehat{\mathbf{E}}[yx^T] & \widehat{\mathbf{E}}[yy^T] \end{pmatrix}$$

is the empirical covariance matrix of (x, y). It turns out that this optimization sometimes yields maximal correlation, suggesting that the learning is trivial. To force nontrivial learning we may introduce penalties for large weight vectors, as in the regression case. Expicitly we want to solve the optimization problem

$$\sup_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{(w_x^T [C_{xx} + \alpha] w_x)(w_y^T [C_{yy} + \beta] w_y)}}$$

which turns out to being equivalent to the generalized eigenvalue problem

$$\begin{pmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} = \rho^2 \begin{pmatrix} C_{xx} + \alpha & 0 \\ 0 & C_{yy} + \beta \end{pmatrix}$$

where ρ is the vector of canonical correlations. As for the regression algorithm a k-fold cross validation should be performed to optimze parameters α, β via a grid search.

Applied to the task, the x data again is the vectorized time windows of pixel data of length 500.000 and the y data is the response data from all 20 neurons.

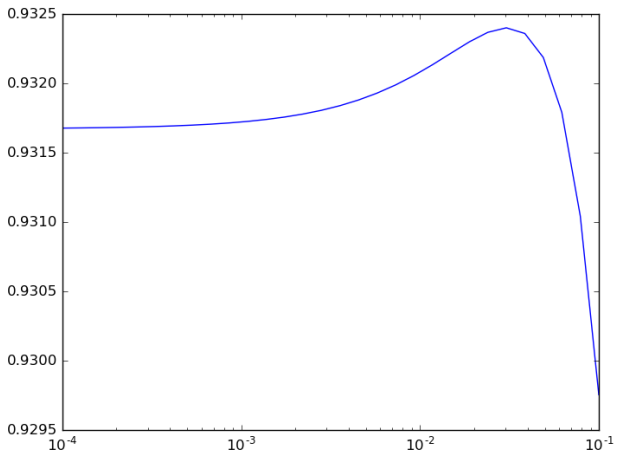

Figure 1.3: Cross validation score as a function of λ for ridge regression

1.5 Results

Restriction The cross validation for ridge regression is shown in Figure 1.5. The optimal λ turns out to be around 0.03. The decline in score is rather flat in the direction of smaller λ but drops drastically for $\lambda > 0.1$.

After performing this cross validation individually for all Neurons we learn optimal weights $w_1, \ldots, w_{20} \in \mathbb{R}^{50\times100\times100}$ with those parameters. For an overall activity plot we now perform a ℓ^2-norm in time direction, i.e.,

$$\tilde{w}_{j,k} = \sqrt{\sum_{t=1}^{50} w_{t,j,k}^2}$$

and plot the resulting matrices for all neurons, see Figure 1.5.

It turns out that the activity of Neuron 14 is espacially strong. Figure 1.5 shows a time series plot of the weights of this Neuron. It can be seen that the strong spatial correlation only shows up in the frames around the fitted frame and not, say, 15 frames in the past.

Canonical Correlation Analysis The crossvalidation has to be performed also for the CCA algorithm, but this time as a grid search since we have to optimize for two parameters. Figure 1.5 shows the crossvalidation score as a function of α and β.

Figure 1.5 shows a plot of the projections in the space of neurons for the top five eigendirections. That is, the front distribution displays the weights on the individual Neurons which is most correlated with the space time weights as displayed in Figure 1.5.

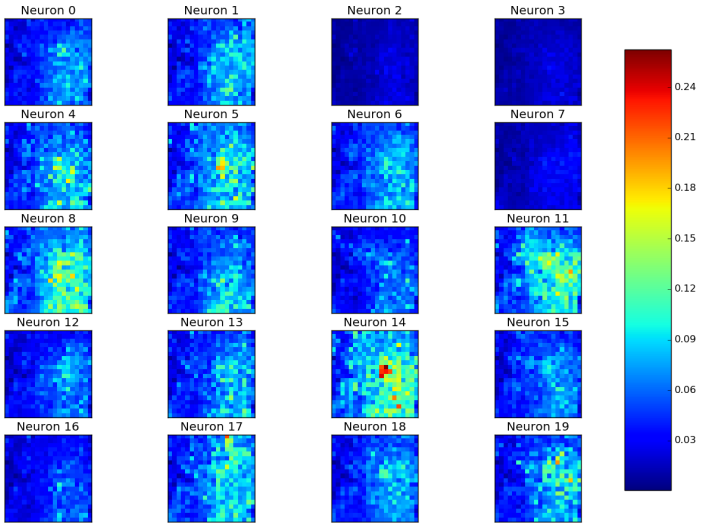

FIGURE 1.4: Spatial Neuron Activity Averaged over Time

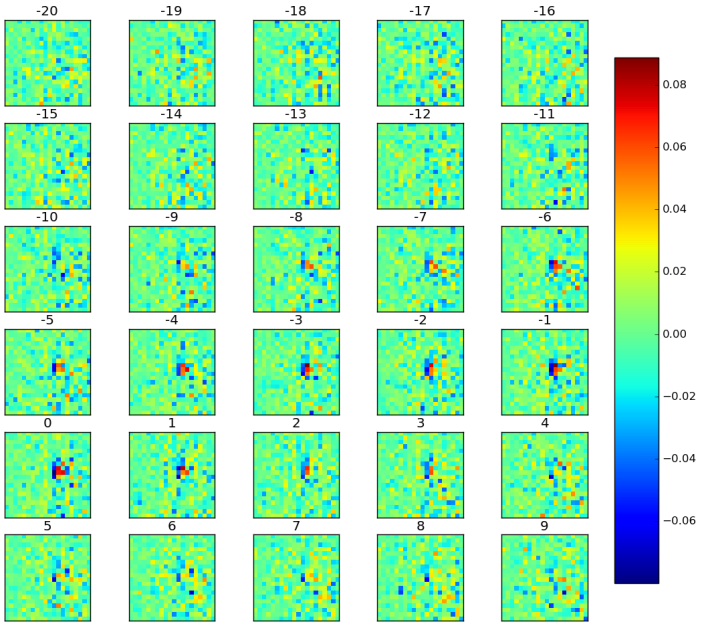

FIGURE 1.5: Spatial Neuron Activity of Neuron 14 for the last 20 and the next 10 Frames

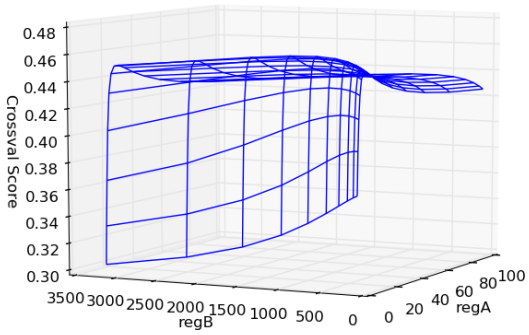

Figure 1.6: Cross validation for CCA

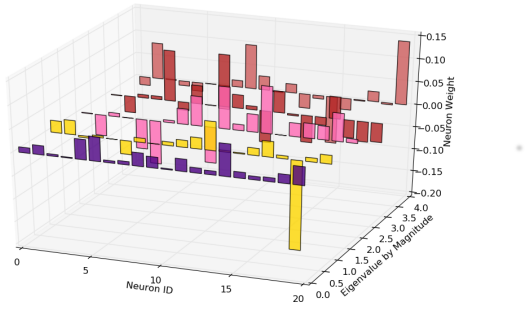

Figure 1.7: Top five neuron weights

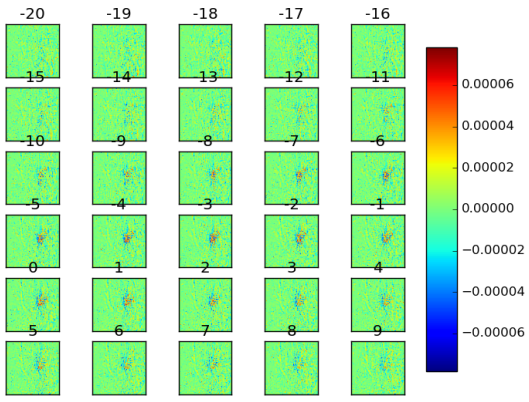

Figure 1.8: Most correlated space-time image weight

Chapter 2

Active Task Selection – Generalization Bounds

For a probability measure $\mathbf{P}$ on $\mathcal{X}$, denote the obvious product probability measure on $\mathcal{X}^m$ by $\mathbf{P}^m$. Let G be a class of functions mapping $\mathcal{X} \to [0, M]$ (endowed with the obvious metric). The goal is to get a probabilistic bound on the maximal difference between the actual expectation $\mathbf{E}_P\, g = \int g \,\mathrm{d}\mathbf{P}$ and the empirical expectation evaluated on a tuple $S = (x_1, \ldots, x_m) \in \mathcal{X}^m$, $\widehat{\mathbf{E}}_S\, g \coloneqq \frac{1}{|S|} \sum_{z \in S} g(z) \coloneqq \frac{1}{m} \sum_{i=1}^m g(z_i)$, where in a slight abuse of notation the tuple S was interpreted as a multiset. For any fixed g it should be clear that $\widehat{\mathbf{E}}_S\, g$ is very close to $\mathbf{E}_P\, g$ when $S \sim \mathbf{P}^m$ for large m. Formally, this follows, for example, from the famous Hoeffding's inequality:

Theorem 2.1 (Hoeffding's Inequality). *Let $X_1, \ldots, X_m$ be independent random variables on a probability space (Ω, P) such that $a_i \leq X_i \leq b_i$ almost surely for all i. Then*

$$P\left[\left|\sum_{i=1}^m (X_i - \mathbf{E}\, X_i)\right| \geq \epsilon\right] \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}\right).$$

We can apply this to our case with $P = \mathbf{P}^m$, $0 \leq X_i(S) \coloneqq g(x_i) \leq M$ (where x_i is the i-the component of S) and $\mathbf{E}_P\, X_i = \mathbf{E}_\mathbf{P}\, g$. Then we find

$$\mathbf{P}^m \left\{ S \in \mathcal{X}^m \;\middle|\; \left|\mathbf{E}_P\, g - \widehat{\mathbf{E}}_S\, g\right| \geq \epsilon \right\} \leq 2 \exp\left(-\frac{2m\epsilon^2}{M^2}\right).$$

Through a trivial union bound over a finite collection G this can be made uniform in $g \in G$, i.e.,

$$\mathbf{P}^m \left\{ S \in \mathcal{X}^m \;\middle|\; \left|\mathbf{E}_\mathbf{P}\, g - \widehat{\mathbf{E}}_S\, g\right| \geq \epsilon \text{ for some } g \in G \right\} \leq 2\,|G| \exp\left(-\frac{2m\epsilon^2}{M^2}\right) \qquad (1)$$

or equivalently with a probability of at least $1 - \delta$ it holds that

$$\left|\mathbf{E}_\mathbf{P}\, g - \widehat{\mathbf{E}}_S\, g\right| \leq M\sqrt{\frac{\log|G| + \log 2/\delta}{2m}} \qquad \text{for all } g \in G.$$

In the following section two slightly different approaches for generalizing eq. (1) to certain infinite families G shall be reviewed and then applied to multi-task learning in the subsequent sections.

2.1 Generalization Bounds by Covering Numbers

This section on covering numbers follows [1] closely.

Covering Numbers Given a set A, a metric d on A and $\epsilon > 0$, we say that $B \subset A$ is an ϵ-cover for A with respect to d if for any $a \in A$, there exists some $b \in B$ with $d(a, b) < \epsilon$. The minimum of cardinality of such sets B we then call the d-covering number and denote it by $\mathcal{N}(\epsilon, A, d)$.

We can then also define covering numbers for classes of functions. For the family of functions $G \subset \mathcal{Y}^{\mathcal{X}}$ (where $(\mathcal{Y}, d_{\mathcal{Y}})$ is a metric space) and a tuple $S \in \mathcal{X}^m$, we denote the range of G on $S = (x_1, \ldots, x_m)$ by $G|_S := \{ (g(x_1), \ldots, g(x_m)) \mid g \in G \} \subset \mathcal{Y}^m$. Now we define the metric $d_m \colon \mathcal{Y}^m \times \mathcal{Y}^m \to \mathbb{R}_+$ by

$$d_m(y, y') := \frac{1}{m} \sum_{i=1}^{m} d_{\mathcal{Y}}(y_i, y_i')$$

and the covering number

$$\mathcal{N}_1(\epsilon, G, m) := \sup_{S \in \mathcal{X}^m} \mathcal{N}(\epsilon, G|_S, d_m).$$

For function collections G with finite covering number, we can generalize eq. (1) to infinite classes:

Theorem 2.2. *If* $\mathbf{P}$ *is a probability measure on* $\mathcal{X}$*,* G *is a set of functions mapping* $\mathcal{X} \to [0, M]$*,* $m \in \mathbb{N}$ *and* $\epsilon > 0$*, then*

$$\mathbf{P}^m \left\{ S \in \mathcal{X}^m \;\middle|\; \left| \widehat{\mathbf{E}}_S\, g - \mathbf{E}_{\mathbf{P}}\, g \right| \geq \epsilon \text{ for some } g \in G \right\} \leq 4 \exp\left(\frac{-m\epsilon^2}{32M^2} \right) \mathcal{N}_1(\epsilon/8, G, 2m).$$

The proof is split into several Lemmata.

Lemma 2.3. *For any* $\epsilon > 0$ *and* $m \geq 2M^2 \log 4/\epsilon^2$*, we have*

$$\begin{aligned} \mathbf{P}^m \left\{ S \in \mathcal{X}^m \;\middle|\; \left| \mathbf{E}_{\mathbf{P}}\, g - \widehat{\mathbf{E}}_S\, g \right| \geq \epsilon \text{ for some } g \in G \right\} \\ \leq 2\mathbf{P}^{2m} \left\{ (S, T) \in \mathcal{X}^m \times \mathcal{X}^m \;\middle|\; \left| \widehat{\mathbf{E}}_S\, g - \widehat{\mathbf{E}}_T\, g \right| \geq \frac{\epsilon}{2} \text{ for some } g \in G \right\}. \end{aligned}$$

Proof. For any fixed $g \in G, \epsilon > 0$ and $m \geq \frac{2M^2 \log 4}{\epsilon^2}$ Hoeffding's inequality states that

$$\mathbf{P}^m \left\{ S \in \mathcal{X}^m \;\middle|\; \left| \mathbf{E}_P\, g - \widehat{\mathbf{E}}_S\, g \right| < \frac{\epsilon}{2} \right\} > 1 - 2\exp\left(-\frac{\epsilon^2 m}{2M^2} \right) \geq \frac{1}{2}.$$

If some some $(S, T) \in \mathcal{X}^m \times \mathcal{X}^m$ there exists $g \in G$ such that $\left| \mathbf{E}_{\mathbf{P}}\, g - \widehat{\mathbf{E}}_T\, g \right| \geq \epsilon$ and $\left| \mathbf{E}_{\mathbf{P}}\, g - \widehat{\mathbf{E}}_S\, g \right| < \epsilon/2$ then this (S, T) occurs on the rhs of the claimed inequality. By independence, given the existence of g with $\left| \mathbf{E}_{\mathbf{P}}\, g - \widehat{\mathbf{E}}_T\, g \right| \geq \epsilon$, this event occurs with probability of at least $\frac{1}{2}$ due to Hoeffding's inequality from above. Thus the claim follows. □

Next, we further bound the probability on the right hand side of the above Lemma in terms of permutations of labels. For $m \in \mathbb{N}$, denote the set of those permutations σ of $[2m]$ which satisfy that for all $i \in [m]$ either $\sigma(i) = i, \sigma(m + i) = m + i$ or $\sigma(i) = m + i, \sigma(m + i) = i$, by Γ_m. Denote the uniform probability distribution on Γ_m by P_{Γ_m}.

Lemma 2.4. *For any measurable subset $R \subset \mathcal{X}^{2m}$, it holds that*

$$\mathbf{P}^{2m}(R) = \mathbf{E}_{S\sim P^{2m}} P_{\Gamma_m}\{\,\sigma \mid \sigma S \in R\,\} \leq \sup_{S\in\mathcal{X}^{2m}} P_{\Gamma_m}\{\,\sigma \mid \sigma S \in R\,\}.$$

Proof. Since σ is measure preserving with respect to $\mathbf{P}^{2m}$, we have

$$\begin{aligned}\mathbf{P}^{2m}(R) &= \frac{1}{|\Gamma_m|}\sum_{\sigma\in\Gamma_m} \mathbf{E}_{S\sim\mathbf{P}^{2m}} 1_R(\sigma S) = \mathbf{E}_{S\sim\mathbf{P}^{2m}}\left[\frac{1}{|\Gamma_m|}\sum_{\sigma\in\Gamma_m} 1_R(\sigma S)\right] \\ &= \mathbf{E}_{S\sim P^{2m}} P_{\Gamma_m}\{\,\sigma \mid \sigma S \in R\,\}.\end{aligned}$$

□

Lemma 2.5. *Let $(S,T) \in \mathcal{X}^m \times \mathcal{X}^m$, $\epsilon > 0$ and $F \subset G$. Suppose that $F|_{(S,T)}$ is an $\epsilon/8$ cover of $G|_{(S,T)}$ with respect to d_1. Then, if $\left|\widehat{\mathbf{E}}_S\, g - \widehat{\mathbf{E}}_T\, g\right| \geq \epsilon/2$ for some $g \in G$, there is some $f \in F$ with $\left|\widehat{\mathbf{E}}_S\, f - \widehat{\mathbf{E}}_T\, f\right| \geq \epsilon/4$.*

Proof. Fix $(S,T) \in \mathcal{X}^m \times \mathcal{X}^m$ and $g \in G$ with $\left|\widehat{\mathbf{E}}_S\, g - \widehat{\mathbf{E}}_T\, g\right| \geq \epsilon/2$. We can then find $f \in F$ with $d_1(f,g) = \widehat{\mathbf{E}}_{(S,T)}\,|g-f| < \epsilon/8$. Then

$$\begin{aligned}\left|\widehat{\mathbf{E}}_S\, f - \widehat{\mathbf{E}}_T\, f\right| &= \left|\widehat{\mathbf{E}}_S\, g - \widehat{\mathbf{E}}_T\, g + \widehat{\mathbf{E}}_S(f-g) - \widehat{\mathbf{E}}_T(f-g)\right| \geq \frac{\epsilon}{2} - \left|\widehat{\mathbf{E}}_S(f-g) - \widehat{\mathbf{E}}_T(f-g)\right| \\ &\geq \frac{\epsilon}{2} - (\widehat{\mathbf{E}}_S\,|f-g| + \widehat{\mathbf{E}}_T\,|f-g|) = \frac{\epsilon}{2} - 2\,\widehat{\mathbf{E}}_{(S,T)}\,|f-g| \geq \frac{\epsilon}{4}.\end{aligned}$$

□

Now we are ready to give the proof of the Theorem.

Proof of Theorem 2.2. For $m \leq 2M^2 \log 4/\epsilon^2$ the statement is trivial since probabilities are at most 1. Else, using Lemmata 2.3 and 2.4 we find that the probability from the Theorem is at most

$$2\sup_{S\in\mathcal{X}^{2m}} P_{\Gamma_m}\left\{\,\sigma \;\middle|\; \left|\widehat{\mathbf{E}}_{(\sigma S)_{1:m}}\, g - \widehat{\mathbf{E}}_{(\sigma S)_{m+1:2m}}\, g\right| \geq \frac{\epsilon}{2} \text{ for some } g \in G\,\right\}$$

where for $T = (x_1,\ldots,x_{2m}) \in \mathcal{X}^{2m}$ we use the shorthand notations $T_{1:m} = (x_1,\ldots,x_m)$ and $T_{m+1:2m} = (x_{m+1},\ldots,x_{2m})$. Now fix some $S = (x_1,\ldots,x_{2m}) \in \mathcal{X}^{2m}$ and let $F \subset G$ be minimal such that $F|_S$ is an $\epsilon/8$ cover of $G|_S$. Then, using Lemma 2.5 we can further bound the above probability by

$$\begin{aligned}&P_{\Gamma_m}\left\{\,\sigma \;\middle|\; \left|\widehat{\mathbf{E}}_{(\sigma S)_{1:m}}\, f - \widehat{\mathbf{E}}_{(\sigma S)_{m+1:2m}}\, f\right| \geq \frac{\epsilon}{4} \text{ for some } f \in F\,\right\} \\ &\quad\leq |F| \max_{f\in F} P_{\Gamma_m}\left\{\,\sigma \;\middle|\; \left|\widehat{\mathbf{E}}_{(\sigma S)_{1:m}}\, f - \widehat{\mathbf{E}}_{(\sigma S)_{m+1:2m}}\, f\right| \geq \frac{\epsilon}{4}\,\right\} \\ &\quad= |F| \max_{f\in F} P_{\Gamma_m}\left\{\,\sigma \;\middle|\; \left|\frac{1}{m}\sum_{i=1}^{m}\left(f(x_{\sigma(i)}) - f(x_{\sigma(m+i)})\right)\right| \geq \frac{\epsilon}{4}\,\right\} \\ &\quad= |F| \max_{f\in F} P_{\lambda\in\{-1,1\}^m}\left\{\,\left|\frac{1}{m}\sum_{i=1}^{m}\lambda_i\left|f(x_{\sigma(i)}) - f(x_{\sigma(m+i)})\right|\right| \geq \frac{\epsilon}{4}\,\right\}\end{aligned}$$

where in the last line λ is drawn uniformly. Since the absolute value in the sum is at most $2M$, we can apply Hoeffding's inequality to bound the probability by $2\exp\left(\frac{-m\epsilon^2}{32M^2}\right)$. This concludes the proof. □

2.2 Generalization Bounds by Rademacher Complexity and VC-dimension

This section follows [2] closely. While in the preceding section importantly relied on Hoeffding's inequality, we now need a more general result called the bounded differences inequality, or sometimes (mainly for the case of constant c_k) the McDiarmid inequality. Firstly, some notation. For a vector $x \in \mathcal{X}^m$, an element $\tilde{x} \in \mathcal{X}$ and an index $k \in [m]$ define the shorthand notation $x_{k,\tilde{x}}$ to denote the vector obtained by replacing x_k in x by $\tilde{x}$ and $x^{(k)}$ to denote the vector of length $m-1$ with dropped component x_k.

Theorem 2.6 (Bounded Differences Inequality). *Suppose $f\colon \mathcal{X}^m \to \mathbb{R}$ and $c_k\colon \mathcal{X}^{m-1} \to \mathbb{R}$ for $k \in [m]$ are functions such that*

$$\sup_{x',x'' \in \mathcal{X}} |f(x_{k,x'}) - f(x_{k,x''})| \leq c_k(x^{(k)})$$

for all $x \in \mathcal{X}^m$ and $k \in [m]$. If $v \coloneqq \sup_{x \in \mathcal{X}^m} \sum_{k \in [m]} c_k^2(x^{(k)}) < \infty$ it then holds for $\epsilon > 0$ and a random vector $X = (X_1, \ldots, X_m)$ with independent entries that

$$\Pr\left[f(X) \geq \mathbf{E}_X\, f(X) + \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{v}\right).$$

While the version with constant bounds c_k is usually proved with martingale techniques, the general case can be proved using the following entropy method. Recall that the entropy of a non-negative random variable is defined to be

$$\mathrm{Ent}(Y) = \mathbf{E}[\Phi(Y)] - \Phi(\mathbf{E}\, Y)$$

where $\Phi(x) = x \log x$ for $x > 0$ and $\Phi(0) = 0$. For a random vector $X = (X_1, \ldots, X_m)$ and a random variable Y introduce the shorthand notation

$$\mathbf{E}^{(k)}\, Y \coloneqq \mathbf{E}[Y | X^{(k)}]$$

to denote the conditional expectation of Y given the random vector $X^{(k)}$. Similarly we denote the entropy of Y condition on $X^{(k)}$ by

$$\mathrm{Ent}^{(k)}(Y) = \mathbf{E}^{(k)}[\Phi(Y)] - \Phi(\mathbf{E}^{(k)}\, Y).$$

It is a standard fact about relative entropies that for $Z = f(X_1, \ldots, X_m)$ we have the following subadditivity of entropies

$$\mathrm{Ent}(Z) \leq \mathbf{E} \sum_{k \in [m]} \mathrm{Ent}^{(k)}(Z). \tag{2}$$

Proof of Theorem 2.6. Write $Z = f(X)$ and let $\lambda > 0$. Then by eq. (2) we find

$$\mathrm{Ent}(e^{\lambda Z}) \leq \mathbf{E} \sum_{k \in [m]} \mathrm{Ent}^{(k)}(e^{\lambda Z}).$$

The key observation now is that Z given $X^{(k)}$ is a random variable with a range of length at most $c_k(X^{(k)})$. At this point we need a Lemma

Lemma 2.7. *If Y is a random variable taking values in $[a, b]$ and $\lambda \in \mathbb{R}$, then*

$$\mathrm{Ent}(e^{\lambda Y}) \leq \frac{(b-a)^2 \lambda^2}{8}\, \mathbf{E}\, e^{\lambda Y}$$

Proof. First note that we can assume wlog. $\mathbf{E}\, Y = 0$ since the general case then follows from the statement for $Y - \mathbf{E}\, Y$. For $\psi(\lambda) = \log \mathbf{E}\, e^{\lambda Y}$ we find $\psi'(\lambda) = \frac{\mathbf{E}\, Y e^{\lambda Y}}{\mathbf{E}\, e^{\lambda Y}}$ and

$$\begin{aligned}\psi''(\lambda) &= \frac{\mathbf{E}\, Y^2 e^{\lambda Y}}{\mathbf{E}\, e^{\lambda Y}} - \left(\frac{\mathbf{E}\, Y e^{\lambda Y}}{\mathbf{E}\, e^{\lambda Y}}\right)^2 = \mathbf{E}\left[\left(Y - \frac{\mathbf{E}\, Y e^{\lambda Y}}{\mathbf{E}\, e^{\lambda Y}}\right)^2 \frac{e^{\lambda Y}}{\mathbf{E}\, e^{\lambda Y}}\right] \\ &= \mathbf{E}\left[\left(Y - \frac{b+a}{2}\right)^2 \frac{e^{\lambda Y}}{\mathbf{E}\, e^{\lambda Y}}\right] - \left(\frac{\mathbf{E}\, Y e^{\lambda Y}}{\mathbf{E}\, e^{\lambda Y}} - \frac{a+b}{2}\right)^2 \leq \frac{(a-b)^2}{4}\end{aligned}$$

since $\left|Y - \frac{a+b}{2}\right| \leq \frac{a-b}{2}$ almost surely. Now it follows that

$$\frac{\mathrm{Ent}(e^{\lambda Y})}{\mathbf{E}\, e^{\lambda Y}} = \lambda \frac{\mathbf{E}\, Y e^{\lambda Y}}{\mathbf{E}\, e^{\lambda Y}} - \log \mathbf{E}\, e^{\lambda Y} = \lambda \psi'(\lambda) - \psi(\lambda) = \int_0^\lambda t\psi''(t)\, \mathrm{d}t \leq \frac{\lambda^2 (b-a)^2}{8},$$

just as claimed. □

Using the Lemma we then find

$$\mathrm{Ent}(e^{\lambda Z}) \leq \mathbf{E} \sum_{k \in [m]} \frac{\lambda^2 c_k^2(X^{(k)})}{8} \mathbf{E}^{(k)}\, e^{\lambda Z} \leq \frac{\lambda^2 v}{8} \mathbf{E}\, e^{\lambda Z}.$$

and consequently

$$\begin{aligned}\log \mathbf{E} e^{\lambda(Z - \mathbf{E}\, Z)} &= \lambda \frac{\log \mathbf{E}\, e^{\lambda(Z - \mathbf{E}\, Z)}}{\lambda} = \lambda \int_0^\lambda \frac{\mathbf{E}(Z - \mathbf{E}\, Z) e^{t(Z - \mathbf{E}\, Z)}}{t\, \mathbf{E}\, e^{t(Z - \mathbf{E}\, Z)}} - \frac{\log \mathbf{E}\, e^{\lambda(Z - \mathbf{E}\, Z)}}{t^2}\, \mathrm{d}t \\ &= \lambda \int_0^\lambda \frac{\mathrm{Ent}(e^{t(Z - \mathbf{E}\, Z)})}{t^2\, \mathbf{E}\, e^{t(Z - \mathbf{E}\, Z)}}\, \mathrm{d}t = \lambda \int_0^\lambda \frac{\mathrm{Ent}(e^{tZ})}{t^2\, \mathbf{E}\, e^{tZ}}\, \mathrm{d}t \leq \lambda \int_0^\lambda \frac{v}{8}\, \mathrm{d}t = \frac{\lambda^2 v}{8}. \qquad (3)\end{aligned}$$

Now the claim follows from Markov's inequality by noticing

$$\Pr\left[Z - \mathbf{E}\, Z \geq \epsilon\right] = \Pr\left[e^{\lambda(Z - \mathbf{E}\, Z)} \geq e^{\lambda\epsilon}\right] \leq \frac{\mathbf{E}\, e^{\lambda(Z - \mathbf{E}\, Z)}}{e^{\lambda\epsilon}} \leq e^{\lambda^2 v/8 - \lambda\epsilon}$$

and choosing the optimal $\lambda = \frac{4\epsilon}{v}$. □

At the price of a worse constant in the bound we can relax the assumption of bounded differences considerably:

Theorem 2.8. *Suppose $f\colon \mathcal{X}^m \to \mathbb{R}$ is a function such that*

$$v := \sup_{x \in \mathcal{X}^m} \sum_{k \in [m]} \left(f(x) - \inf_{x' \in \mathcal{X}} f(x_{k,x'})\right)^2 < \infty.$$

Then for $\epsilon > 0$ and a random vector $X = (X_1, \ldots, X_m)$ with independent entries we have

$$\Pr\left[f(X) \geq \mathbf{E}_X\, f(X) + \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2v}\right).$$

Proof. Write $Z = f(X)$ and $Z_k = \inf_{x' \in \mathcal{X}} f(X_{k,x'})$ and let $\lambda > 0$. For any non-negative random variable Y and any number $u > 0$ it follows from $\log x \leq x - 1$ that

$$\mathrm{Ent}(Y) = \mathbf{E}\, Y \log \frac{u}{\mathbf{E}\, Y} + \mathbf{E}[Y(\log Y - \log u)] \leq \mathbf{E}[Y(\log Y - \log u) - (Y - u)],$$

which in fact becomes an equality when taking the infimum over u. Then again by the subadditivity of the entropy and using this entropy inequality for the conditional expectations we find

$$\begin{aligned}\operatorname{Ent}(e^{\lambda Z}) &\leq \mathbf{E} \sum_{k\in[m]} \operatorname{Ent}^{(k)}(e^{\lambda Z}) \leq \mathbf{E} \sum_{k\in[m]} \mathbf{E}^{(k)} \left[e^{\lambda Z}(\lambda Z - \lambda Z_k) - (e^{\lambda Z} - e^{\lambda Z_k})\right] \\ &= \mathbf{E} \sum_{k\in[m]} \mathbf{E}^{(k)} \left[e^{\lambda Z}\left(e^{-\lambda(Z-Z_k)} + \lambda(Z-Z_k) - 1\right)\right] \\ &\leq \mathbf{E} \sum_{k\in[m]} \mathbf{E}^{(k)} \left[e^{\lambda Z}\frac{\lambda^2(Z-Z_k)^2}{2}\right] \leq \frac{\lambda^2 v}{2}\,\mathbf{E}\left[e^{\lambda Z}\right]\end{aligned}$$

where it was used that $\lambda(Z - Z_k) \geq 0$. Now proceeding as in eq. (3) this implies $\mathbf{E}\, e^{\lambda(Z-\mathbf{E}\,Z)} \leq e^{\lambda^2 v/2}$ and thereby by Markov's inequality

$$\Pr\left[Z - \mathbf{E}\,Z \geq \epsilon\right] \leq e^{\lambda^2 v/2 - \lambda\epsilon}.$$

Minimizing this expression at $\lambda = \frac{\epsilon}{v}$ completes the proof. □

We now apply Theorem 2.6 to derive a generalization bound in the fashion of Theorem 2.2. Define $\phi\colon \mathcal{X}^m \to \mathbb{R}$ by

$$f(S) \coloneqq \sup_{g\in G}\left[\mathbf{E}\, g - \widehat{\mathbf{E}}_S\, g\right] = \sup_{g\in G}\frac{1}{m}\sum_{k\in[m]}\left[\mathbf{E}\, g - g(x_k)\right].$$

Then $|f(S_{k,x'}) - f(S_{k,x''})| = |g(x'') - g(x')|/m \leq M/m$ and thereby

$$\Pr[f(S) \geq \mathbf{E}\, f + \epsilon] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{k\in[m]} M^2/m^2}\right) = \exp\left(-\frac{2m\epsilon^2}{M^2}\right).$$

To bound $\mathbf{E}\, f$ we firstly symmetrize the expression by noticing that

$$\mathbf{E}\, g = \mathbf{E}_{\tilde{S}}\, \widehat{\mathbf{E}}_{\tilde{S}}\, g = \mathbf{E}_{\tilde{S}}\, \frac{1}{m}\sum_{k\in[m]} g(\tilde{x}_k)$$

for $\tilde{S} = (\tilde{x}_1, \ldots, \tilde{x}_m)$ being distributed according to $\mathbf{P}^m$. Therefore by convexity of the supremum function and Jensen's inequality we find

$$\mathbf{E}\, f \leq \mathbf{E}_{S,\tilde{S}} \sup_{g\in G}\frac{1}{m}\sum_{k\in[m]}(g(\tilde{x}_k) - g(x_k)).$$

By exchanging x_k with $\tilde{x}_k$ we only change the order of summation and thereby not the expectation. On the other hand this exchange switches the sign of the k-th summand and therefore the result is invariant under sign flips of any summands. Independent random variables $\sigma_1, \ldots, \sigma_m$ taking the values ± 1 independently with equal probabilities are often called *Rademacher variables*. By the above argument we can insert them into the sum without changing the result and thereby

$$\mathbf{E}\, f \leq \mathbf{E}_{S,\tilde{S}}\, \mathbf{E}_\sigma \sup_{g\in G}\frac{1}{m}\sum_{k\in[m]}\sigma_k(g(\tilde{x}_k) - g(x_k)) \leq 2\,\mathbf{E}_S\, \mathbf{E}_\sigma \sup_{g\in G}\frac{1}{m}\sum_{k\in[m]}\sigma_k g(x_k) =: 2\,\mathbf{E}_S\, \widehat{\mathcal{R}}_S(G)$$

which is often called the Rademacher complexity. An important tool to bound Rademacher complexities is the following well known Lemma

Lemma 2.9 (Massart's Lemma). *For any set $A \subset \mathbb{R}^m$ the empirical* Rademacher complexity *can be bounded by*

$$\mathbf{E}_\sigma \frac{1}{m} \sup_{a \in A} \sum_{k \in [m]} \sigma_k a_k \leq \frac{\sqrt{2 \log |A|}}{m} \sup_{a \in A} \|a\|_2 .$$

For any fixed S applying this Lemma to the above complexity term gives

$$\widehat{\mathcal{R}}_S(G) = \mathbf{E}_\sigma \sup_{g \in G} \frac{1}{m} \sum_{k \in [m]} \sigma_k g(x_k) \leq M \sqrt{\frac{2 \log \left| G|_S \right|}{m}}$$

where $G|_S \coloneqq \{ (g(x_1), \ldots, g(x_m)) \mid g \in G \}$. When we define the *growth function* of G evaluated in m to be the number of distinct ways G can map any m points from $\mathcal{X}$, i.e.

$$\Pi_G(m) \coloneqq \sup_{S \in \mathcal{X}^m} \left| G|_S \right| ,$$

this yields

$$\mathbf{E}\, f \leq 2M \sqrt{\frac{2 \log \Pi_G(m)}{m}}.$$

The above argument proves the following Theorem:

Theorem 2.10. *If $\mathbf{P}$ is a probability measure on $\mathcal{X}$, G is a set of functions mapping $\mathcal{X} \to [0, M]$, $m \in \mathbb{N}$ and $\epsilon > 0$, then*

$$\begin{aligned} &\mathbf{P}^m \left\{ S \in \mathcal{X}^m \;\middle|\; \mathbf{E}_\mathbf{P}\, g \geq \widehat{\mathbf{E}}_S\, g + 2\, \mathbf{E}_S\, \widehat{\mathcal{R}}_S(G) + \epsilon \text{ for some } g \in G \right\} \\ &\leq \mathbf{P}^m \left\{ S \in \mathcal{X}^m \;\middle|\; \mathbf{E}_\mathbf{P}\, g \geq \widehat{\mathbf{E}}_S\, g + 2M \sqrt{\frac{2 \log \Pi_G(m)}{m}} + \epsilon \text{ for some } g \in G \right\} \leq \exp\left(-\frac{2m\epsilon^2}{M^2} \right). \end{aligned}$$

VC dimension The growth function is rather difficult to handle in general but for concrete classes of functions it can be simplified considerably. From now on we shall assume that G maps $\mathcal{X}$ into $\{-1, 1\}$. We say that a $S = \{x_1, \ldots, x_m\}$ is shattered by G if G is of full range on S, i.e., if $\left| G|_S \right| = 2^m$. The *VC-dimension* (Vapnik-Chervonenkis dimension) d of G is defined to be the size of the largest tuple shattered by G, i.e.,

$$d \coloneqq \max \{ m \in \mathbb{N} \mid \Pi_m(G) = 2^m \} .$$

The following well known Lemma relates the growth function for all m to the VC-dimension:

Lemma 2.11 (Sauer's Lemma). *If G maps $\mathcal{X} \to \{-1, 1\}$ and is of VC-dimension d, then*

$$\Pi_m(G) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d} \right)^d .$$

2.3 Active Task Selection

Let $\mathcal{X}$ be an input space and $(\mathcal{Y}, d_\mathcal{Y})$ a corresponding (metric) label space. We shall assume that we want to learn a set of T tasks, where each task is formally represented by a distribution $D_1, \ldots, D_T$ over $\mathcal{X}$ and a deterministic labelling function $f_1, \ldots, f_T \colon \mathcal{X} \to \mathcal{Y}$. We

further assume a bounded loss function $l\colon \mathcal{Y} \times \mathcal{Y} \to [0, M]$ satisfying the triangle inequality $l(y_0, y_2) \leq l(y_0, y_1) + l(y_1, y_2)$ for all $y_0, y_1, y_2 \in \mathcal{Y}$. We define the expected risk of some hypothesis $h\colon \mathcal{X} \to \mathcal{Y}$ for task (D_t, f_t) as

$$\mathrm{er}_t(h) \coloneqq \mathbf{E}_{x\sim D_t}\, l(h(x), f_t(x))$$

and the empirical risk evaluated on some sample $S_t = (x_1, \ldots, x_m) \in \mathcal{X}^m$ independently drawn according to D_t as

$$\widehat{\mathrm{er}}_{S_t}(h) \coloneqq \frac{1}{|S_t|} \sum_{x\in S_t} l(h(x), f_t(x))$$

(where in a slight abuse of notation the tuple S_t was identified as a multiset). The goal is to find predictors $h_1, \ldots, h_T$ from a given hypothesis set $H \subset \mathcal{Y}^{\mathcal{X}}$ minimizing the averaged expected risk

$$\mathrm{er}(h_1, \ldots, h_T) \coloneqq \frac{1}{T} \sum_{t=1}^{T} \mathrm{er}_{D_t}(h_t).$$

In the specific setting we shall assume that labelling information can only be requested for a subset of the tasks and will only be provided for a random subset of the samples representing the task. Formally, the learner is given independently drawn samples $S_1 \sim D_1^n, \ldots, S_T \sim D_T^n$ of size n for all tasks, may then ask for labels for tasks $\{i_1, \ldots, i_k\} \subset [T]$ and will be provided with labels $f_{i_j}(x)$ for $x \in \overline{S_{i_j}}$ and $j \in [k]$ for randomly drawn sub-samples $\overline{S_{i_j}} \subset S_{i_j}$ of sizes $\left|S_{i_j}\right| = m$.

2.4 Single Task Transfer

The most straight forward learning strategy would be assigning some labelled task to each of the unlabelled tasks. Such an assignment shall be encoded by a vector $c = (c_1, \ldots, c_T) \in [T]^T$ having at most k different values. The unlabelled tasks are then solved by using the hypothesis trained on the corresponding labelled task, i.e., $h_t = h_{c_t}$. We shall now derive a bound on the averaged expected risk of this strategy in terms of individual risks and the similarity of the assigned tasks.

Definition 2.12 (Discrepancy). *The discrepancy between two distributions D_1, D_2 over $\mathcal{X}$ with respect to some hypothesis set H is defined as*

$$\mathrm{disc}(D_1, D_2) \coloneqq \sup_{h,h'\in H} \left|\mathrm{er}_{D_1}(h, h') - \mathrm{er}_{D_2}(h, h')\right|.$$

Lemma 2.13. *For two tasks $(D_1, f_1), (D_2, f_2)$ and any hypothesis $h \in H$ it holds that*

$$\mathrm{er}_2(h) \leq \mathrm{er}_1(h) + \mathrm{disc}(D_1, D_2) + \inf_{h^*\in H}(\mathrm{er}_1(h^*) + \mathrm{er}_2(h^*)).$$

Proof. For any $h, h^* \in H$ it holds by the triangle inequality that

$$\begin{aligned}
\mathrm{er}_2(h) &= \mathrm{er}_2(h, f_2) \leq \mathrm{er}_2(h^*, f_2) + \mathrm{er}_2(h^*, h) \\
&= \mathrm{er}_2(h^*) + \mathrm{er}_1(h^*, h) + (\mathrm{er}_2(h^*, h) - \mathrm{er}_1(h^*, h)) \\
&\leq \mathrm{er}_2(h^*) + \mathrm{er}_1(h^*, f_1) + \mathrm{er}_1(h, f_1) + \mathrm{disc}(D_1, D_2) \\
&= \mathrm{er}_1(h) + \mathrm{disc}(D_1, D_2) + (\mathrm{er}_1(h^*) + \mathrm{er}_2(h^*))
\end{aligned}$$

from which the claim follows after taking the infimum over all $h^* \in H$. □

Using the shorthand notation

$$\lambda_{ij} \coloneqq \inf_{h^* \in H} (\mathrm{er}_i(h^*) + \mathrm{er}_j(h^*))$$

we can use the above Lemma to bound the average expected risk by

$$\mathrm{er}(h_1, \dots, h_T) \leq \frac{1}{T} \sum_{t=1}^{T} \mathrm{er}_{D_{c_t}}(h_{c_t}) + \frac{1}{T} \sum_{t=1}^{T} \mathrm{disc}(D_t, D_{c_t}) + \frac{1}{T} \sum_{t=1}^{T} \lambda_{tc_t}. \qquad (4)$$

The next step is bounding these expectation expression by their empirical counterparts, and thereby making the bound independent of the (in general unknown) distributions D_t.

Now we want to apply our general Theorems 2.2 and 2.10 to the empirical error estimate. Given $H, H' \in \mathcal{Y}^{\mathcal{X}}$, define

$$L_{H,H'} \coloneqq \left\{ x \mapsto l(h(x), h'(x)) \;\middle|\; h \in H, h' \in H' \right\} \subset [0, M]^{\mathcal{X}}.$$

Firstly, a Lemma for estimating covering numbers of Lipschitz-loss functions:

Lemma 2.14. *Assume that the bounded loss function $l\colon \mathcal{Y}^2 \to [0, M]$ is L-Lipschitz, i.e., $|l(y_0, y_1) - l(y_0, y_2)| \leq L d(y_1, y_2)$. Then for any $\epsilon_1 + \epsilon_2 = \epsilon > 0$ and $m \in \mathbb{N}$ it holds that*

$$\mathcal{N}_1(\epsilon, L_{H,H'}, m) \leq \mathcal{N}_1(\epsilon_1/L, H, m) \cdot \mathcal{N}_1(\epsilon_2/L, H', m).$$

In particular, if $|H'| = 1$, then $\mathcal{N}_1(\epsilon, L_{H,H'}, m) \leq \mathcal{N}_1(\epsilon/L, H, m)$.

Proof. Fix $S \in \mathcal{X}^m$ and let $F \subset H, F' \subset H'$ generate ϵ_1/L and ϵ_2/L coverings of $H|_S$ and $H'|_S$, respectively. Then, for any $h \in H, h' \in H'$ we find f, f' from the coverings such that $\widehat{\mathbf{E}}_S |h - f| < \epsilon_1/L$ and $\widehat{\mathbf{E}}_S |h' - f'| < \epsilon_2/L$. Then

$$\begin{aligned} d_1(l(h(\cdot), h'(\cdot)), l(f(\cdot), f'(\cdot))) &= \mathbf{E}_S \left| l(h(\cdot), h'(\cdot)) - l(f(\cdot), f'(\cdot)) \right| \\ &\leq \mathbf{E}_S \left| l(h(\cdot), h'(\cdot)) - l(h(\cdot), f'(\cdot)) \right| + \mathbf{E}_S \left| l(h(\cdot), f'(\cdot)) - l(f(\cdot), f'(\cdot)) \right| \\ &\leq \widehat{\mathbf{E}}_S \, L d(h'(\cdot), f'(\cdot)) + \widehat{\mathbf{E}}_S \, L d(h(\cdot), f(\cdot)) < \epsilon. \end{aligned}$$

This shows that $L_{F,F'}$ is a ϵ-covering of $L_{H,H'}$. The second claim follows since a family consisting of one element has covering number 1 for all ϵ. □

Secondly, a Lemma for Rademacher complexities in the case of binary classification $Y = \{-1, 1\}$:

Lemma 2.15. *Assume $H \subset \{-1, 1\}^{\mathcal{X}}$, a binary loss function l given by $l(y, y') = 1_{y \neq y'} = (1 - y \cdot y')/2$ and a given target function $f\colon \mathcal{X} \to \{-1, 1\}$. Then*

$$\widehat{\mathcal{R}}_S(L_{H,\{f\}}) = \frac{1}{2} \widehat{\mathcal{R}}_S(H) \quad \textit{and} \quad \widehat{\mathcal{R}}_S(L_{H,H}) \leq \widehat{\mathcal{R}}_S(H).$$

Proof. Since the σ_k have zero mean we find

$$\widehat{\mathcal{R}}_S(L_{H,\{f\}}) = \mathbf{E}_\sigma \sup_{h \in H} \frac{1}{m} \sum_{k \in [m]} \sigma_k \frac{1 - f(x_k) h(x_k)}{2} = \frac{1}{2} \mathbf{E}_\sigma \sup_{h \in H} \frac{1}{m} \sum_{k \in [m]} \sigma_k h(x_k) = \frac{1}{2} \widehat{\mathcal{R}}_S(H)$$

and therefore also

$$\widehat{\mathcal{R}}_S(L_{H,H}) = \mathbf{E}_\sigma \sup_{h, h' \in H} \frac{1}{m} \sum_{k \in [m]} \sigma_k l(h(x_k), h'(x_k)) \leq 2 \, \mathbf{E}_\sigma \sup_{h \in H} \frac{1}{m} \sum_{k \in [m]} \sigma_k l(h(x_k), f(x_k)) = \widehat{\mathcal{R}}_S(H).$$

□

This enables us to prove the following Theorem

Theorem 2.16. *Let $\mathbf{P}_1, \ldots, \mathbf{P}_m$ be measures implementing the distributions $D_1, \ldots, D_m$. With the definitions from the beginning of the chapter the following assertions holds true.*

(a) If H is finite, then

$$\mathbf{P}_t^m \left\{ S \in \mathcal{X}^m \;\middle|\; \sup_{h \in H} [\mathrm{er}_t(h) - \widehat{\mathrm{er}}_S(h)] < \epsilon \right\} > 1 - |H| \exp\left(-\frac{2m\epsilon^2}{M^2}\right)$$

and

$$\mathbf{P}_t^n \left\{ S_t \in \mathcal{X}^n \mid \mathrm{disc}(D_t, S_t) < \epsilon \right\} > 1 - 2\,|H|^2 \exp\left(\frac{-2n\epsilon^2}{M^2}\right).$$

(b) If the loss function l is L-Lipschitz, then

$$\mathbf{P}_t^m \left\{ S \in \mathcal{X}^m \;\middle|\; \sup_{h \in H} [\mathrm{er}_t(h) - \widehat{\mathrm{er}}_S(h)] < \epsilon \right\} > 1 - 2\exp\left(\frac{-m\epsilon^2}{32M^2}\right) \mathcal{N}_1\left(\frac{\epsilon}{8L}, H, 2m\right)$$

and

$$\mathbf{P}_t^n \left\{ S_t \in \mathcal{X}^n \mid \mathrm{disc}(D_t, S_t) < \epsilon \right\} > 1 - 4\exp\left(\frac{-n\epsilon^2}{32M^2}\right) \mathcal{N}_1\left(\frac{\epsilon}{8L}, H, 2n\right)^2.$$

(c) In the case of binary classification, 0-1-loss and H of VC-dimension d we have

$$\mathbf{P}_t^m \left\{ S \in \mathcal{X}^m \;\middle|\; \sup_{h \in H} [\mathrm{er}_t(h) - \widehat{\mathrm{er}}_S(h)] < \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \epsilon \right\} > 1 - \exp\left(-2m\epsilon^2\right)$$

and

$$\mathbf{P}_t^n \left\{ S_t \in \mathcal{X}^n \;\middle|\; \mathrm{disc}(D_t, S_t) < 2\sqrt{\frac{2d \log \frac{en}{d}}{n}} + \epsilon \right\} > 1 - 2\exp\left(-2n\epsilon^2\right).$$

Moreover, the deviation bounds in the first claim of each statement can be made two-sided at the expense of doubling the error probability.

Proof. The first claims of all three cases follow directly from Theorems 2.1, 2.2 and 2.10 and the Lemmata relating the covering numbers and complexities of $L_{H,H'}$ to those of H. For the second claim note that if

$$\Pr\left[\sup_{g \in L_{H,H}} [\mathbf{E}_{\mathbf{P}_t}\, g - \widehat{\mathbf{E}}_{S_t}\, g] > \epsilon\right] \leq \delta \quad \text{and} \quad \Pr\left[\sup_{g \in L_{H,H}} [\widehat{\mathbf{E}}_{S_t}\, g - \mathbf{E}_{\mathbf{P}_t}\, g] > \epsilon\right] \leq \delta$$

then also

$$\Pr\left[\sup_{g \in L_{H,H}} \left|\mathbf{E}_{\mathbf{P}_t}\, g - \widehat{\mathbf{E}}_{S_t}\, g\right| > \epsilon\right] = \Pr[\mathrm{disc}(D_t, S_t) > \epsilon] \leq 2\delta.$$

The second claims then follow just as the first ones. □

The bound in terms of covering numbers is difficult to handle in general. For the other situation we can, given some $\delta > 0$ solve the bounds for ϵ so that every single bound from the Theorem holds with a probability of at least $1 - \delta/2T$. By the triangle inequality for discrepancies

$$\operatorname{disc}(D_t, D_s) - \operatorname{disc}(S_t, S_s) \leq \operatorname{disc}(D_t, S_t) + \operatorname{disc}(D_s, S_s)$$

the bounds on distribution-sample-discrepancies $\operatorname{disc}(D_t, S_t)$ suffice to bound the expected discrepancies by their empirical counterparts. According to this, we find that with a probability of at least $1 - \delta$ over $S_1 \sim D_1^n, \ldots, S_T^n \sim D_T^n$ and random m-subsets of those it holds that

$$\begin{aligned}\operatorname{er}(h_1, \ldots, h_T) \leq & \frac{1}{T} \sum_{t=1}^{T} \widehat{\operatorname{er}}_{\overline{S_{c_t}}}(h_{c_t}) + \frac{1}{T} \sum_{t=1}^{T} \operatorname{disc}(S_t, S_{c_t}) + \frac{1}{T} \sum_{t=1}^{T} \lambda_{tc_t} \\ & + M \sqrt{\frac{\log |H| + \log T + \log 2/\delta}{2m}} + M \sqrt{\frac{4 \log |H| + 2 \log T + 2 \log 4/\delta}{n}}.\end{aligned}$$

Similarly from the result for binary classification we find

$$\begin{aligned}\operatorname{er}(h_1, \ldots, h_T) \leq & \frac{1}{T} \sum_{t=1}^{T} \widehat{\operatorname{er}}_{\overline{S_{c_t}}}(h_{c_t}) + \frac{1}{T} \sum_{t=1}^{T} \operatorname{disc}(S_t, S_{c_t}) + \frac{1}{T} \sum_{t=1}^{T} \lambda_{tc_t} \\ & + \sqrt{\frac{\log T + \log 2/\delta}{2m}} + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + 2\sqrt{\frac{\log T + \log 4/\delta}{2n}} + 4\sqrt{\frac{2d \log \frac{en}{d}}{n}}.\end{aligned}$$

This bound is not optimal and we shall see in the subsequent section that a joint bound on the mean error by using Theorem 2.8 instead of 2.6 gets rid of the $\log T$ in the first part of the bound.

2.5 Multi-task Learning

We now turn to a more general learning strategy. Denote the weight simplex by

$$\Lambda \coloneqq \left\{ \alpha \in \mathbb{R}_+^T \;\middle|\; \sum_{i \in [T]} \alpha_i = 1 \right\}$$

and for a set of selected tasks $I \subset [T]$ the weight simplex with sparsity pattern I by $\Lambda^I \coloneqq \{ \alpha \in \Lambda \mid \operatorname{supp} \alpha \subset I \}$. Then we can define the α-weighted (empirical error) by

$$\operatorname{er}_\alpha(h) \coloneqq \sum_{i \in I} \alpha_i \operatorname{er}_i(h), \qquad \widehat{\operatorname{er}}_\alpha(h) \coloneqq \sum_{i \in I} \alpha_i \widehat{\operatorname{er}}_{\overline{S_i}}(h).$$

Through Lemma 2.13 we can relate a single task error in terms of the weighted error. Explicitly,

$$\operatorname{er}_t(h) - \operatorname{er}_\alpha(h) = \sum_{i \in I} \alpha_i (\operatorname{er}_t(h) - \operatorname{er}_i(\alpha)) \leq \sum_{i \in I} \alpha_i \left(\operatorname{disc}(D_i, D_t) + \lambda_{it} \right).$$

If now every task (D_t, f_t) has its own weight vector $\alpha^t \in \Lambda^I$ we can apply this to the averaged error to find for any selected hypotheses $h_1, \ldots, h_T$

$$\operatorname{er}(h_1, \ldots, h_T) \leq \frac{1}{T} \sum_{t \in [T]} \operatorname{er}_{\alpha^t}(h_t) + \frac{1}{T} \sum_{t \in [T]} \sum_{i \in I} \alpha_i^t \operatorname{disc}(D_i, D_t) + \frac{1}{T} \sum_{t \in [T]} \sum_{i \in I} \alpha_i^t \lambda_{it} \tag{5}$$

Our goal is the find a bound on the probability

$$\begin{aligned}
&\Pr\left[\frac{1}{T}\sum_{t\in[T]}\sum_{i\in I}\alpha_i^t\left(\mathrm{er}_i(h_t)-\widehat{\mathrm{er}}_{S_i}(h_t)\right)\le\epsilon \quad \text{for all } I\in\binom{[T]}{K},\alpha\in(\Lambda^I)^T,h\in H^T\right]\\
&\quad=\Pr\left[\sup_{I\in\binom{[T]}{k}}\sup_{\alpha\in(\Lambda^I)^T}\sup_{h\in H^T}\frac{1}{mT}\sum_{t\in[T]}\sum_{i\in I}\sum_{k\in[m]}\alpha_i^t\left[\mathrm{er}_i(h_t)-l(h_t(x_k^i),f_t(x_k^i))\right]\le\epsilon\right]\\
&\quad=\Pr\left[\sup_{I\in\binom{[T]}{k}}\sup_{h\in H^T}\sup_{i\in I^T}\frac{1}{mT}\sum_{t\in[T]}\sum_{k\in[m]}\left[\mathrm{er}_{i_t}(h_t)-l(h_t(x_k^{i_t}),f_t(x_k^{i_t}))\right]\le\epsilon\right]
\end{aligned}$$

where the probability is taken over

$$S=(S_1,\ldots,S_T)=\begin{pmatrix}x_1^1 & \cdots & x_1^T\\ \vdots & & \vdots\\ x_m^1 & \cdots & x_m^T\end{pmatrix}$$

with $x^t\sim D_t^m$ independent from each other. Define

$$\begin{aligned}
g(h,i,S)&:=\frac{1}{mT}\sum_{t\in[T]}\sum_{k\in[m]}\left[\mathrm{er}_{i_t}(h_t)-l(h_t(x_k^{i_t}),f_t(x_k^{i_t}))\right],\\
f(S)&=\sup_{I\in\binom{[T]}{K}}\sup_{i\in I^T}\sup_{h\in H^T}g(h,i,S)=\sup_{h\in H^T}g(h,i^*(S),S)
\end{aligned}$$

where $i^*(S)\in(I^*)^T$, $I^*(S)\in\binom{[T]}{K}$ is the location of the maximum of $i\mapsto\sup_h g(h,i,S)$ and fix $s\in[T],j\in[m]$. Then

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left[f(S)-f(S_{(j,s),x})\right]\le\sup_{x,I,i,h}\left[g(h,i,S)-g(h,i,S_{(j,s),x})\right]\\
&\le\sup_{x,h}\left[g(h,i^*(S),S)-g(h,i^*(S),S_{(j,s),x})\right]\\
&=\sup_{x,h}\frac{1}{mT}\sum_{t\in[T],i^*(S)_t=s}\left[l(h_t(x),f_t(x))-l(h_t(x_j^s),f_t(x_j^s))\right]\le\frac{M\left|\{\,t\in[T]\mid i^*(S)_t=s\,\}\right|}{mT}
\end{aligned}$$

for all S and consequently we can apply Theorem 2.8 with $v=\frac{M^2}{m}$ and find that with a probability of at least $1-\delta/2$ over S,

$$f(S)\le\mathbf{E}\,f+M\sqrt{\frac{2\log 2/\delta}{m}}.$$

Next, we consider how to get a bound on $\mathbf{E}\,f$. For the case of binary classification following the standard Rademacher technique we find

$$\mathbf{E}\,f\le\mathbf{E}_{S,\sigma}\sup_{h\in H^T}\sup_{i\in[T]^T}\frac{1}{mT}\sum_{t,k}\sigma_{t,k}h_t(x_k^{i_t})$$

on which we apply Massart's Lemma on

$$A|_S=\left\{\,(h_t(x_k^{i_t}))_{t\in[T],k\in[m]}\;\middle|\;h\in H^T,i\in[T]^T\,\right\}.$$

Using Sauer's Lemma

$$\left|A|_S\right| \leq \Pi_m(H)^T \leq \left(\frac{em}{d}\right)^{dT}.$$

we find that with a probability of at least $1-\delta/2$

$$\frac{1}{T}\sum_{t\in[T]} \mathrm{er}_{\alpha^t}(h_t) \leq \frac{1}{T}\sum_{t\in[T]} \widehat{\mathrm{er}}_{\alpha^t}(h_t) + \sqrt{\frac{2d\log\frac{em}{d}}{m}} + \sqrt{\frac{2\log\frac{2}{\delta}}{m}}$$

simultaneously for all h and α. Similarly, in the case of an arbitrary but finite hypotheses set we find

$$\frac{1}{T}\sum_{t\in[T]} \mathrm{er}_{\alpha^t}(h_t) \leq \frac{1}{T}\sum_{t\in[T]} \widehat{\mathrm{er}}_{\alpha^t}(h_t) + 2M\sqrt{\frac{2\log|H|}{m}} + M\sqrt{\frac{2\log\frac{2}{\delta}}{m}},$$

again simultaneously with a probability of at least $1-\delta/2$. The discrepancies can be bounded as is in single-task learning case and we have completed the proof of the following Theorem:

Theorem 2.17. *Let $\delta > 0$ be given. Then with a probability of at least $1-\delta$ over samples $S_1 \sim D_1^n, \ldots, S_T \sim D_T^n$ and random sub-samples $\overline{S_1}, \ldots, \overline{S_T}$ of size m the following assertions hold true uniformly in $h_1, \ldots, h_T \in H$, $I \in \binom{[T]}{k}$ and $\alpha^1, \ldots, \alpha^T \in \Lambda^I$.*

(a) If H is finite, then

$$\mathrm{er}(h_1,\ldots,h_T) \leq \frac{1}{T}\sum_{t\in[T]} \widehat{\mathrm{er}}_{\alpha^t}(h_t) + \frac{1}{T}\sum_{t\in[T]}\sum_{i\in I} \alpha_i^t \,\mathrm{disc}(S_i,S_t) + \frac{1}{T}\sum_{t\in[T]}\sum_{i\in I}\alpha_i^t\lambda_{it}$$
$$2M\sqrt{\frac{2\log|H|}{m}} + M\sqrt{\frac{2\log\frac{2}{\delta}}{m}} + 4M\sqrt{\frac{2\log|H|}{n}} + 2M\sqrt{\frac{\log T + \log 4/\delta}{2n}}.$$

(b) If H is a binary classifier of VC-dimension d, then

$$\mathrm{er}(h_1,\ldots,h_T) \leq \frac{1}{T}\sum_{t\in[T]} \widehat{\mathrm{er}}_{\alpha^t}(h_t) + \frac{1}{T}\sum_{t\in[T]}\sum_{i\in I} \alpha_i^t \,\mathrm{disc}(S_i,S_t) + \frac{1}{T}\sum_{t\in[T]}\sum_{i\in I}\alpha_i^t\lambda_{it}$$
$$\sqrt{\frac{2d\log\frac{em}{d}}{m}} + \sqrt{\frac{2\log\frac{2}{\delta}}{m}} + 4\sqrt{\frac{2d\log\frac{en}{d}}{n}} + 2\sqrt{\frac{\log T + \log 4/\delta}{2n}}.$$

2.6 Weight dependent convergence rate

If one fixes some weight $\alpha \in \mathbb{R}_+^k$ such that $\sum_{i\in[k]}\alpha_i = 1$ and $h \in H$ Hoeffding's inequality gives the bound

$$\Pr\left[|\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h)| \geq \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{\sum_{i\in[k]}\alpha_i^2}\right), \tag{6}$$

or equivalently that with a probability of at least $1-\delta$,

$$|\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h)| \leq \|\alpha\|_2\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{7}$$

Indeed,

$$|\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h)| = \left| \sum_{i\in[k]} \sum_{j\in[m]} \frac{\alpha_i}{m} \left[\mathrm{er}_i(h) - l(h(x_j^i), f_i(x_j^i)) \right] \right|$$

where the (i,j)-th term in the sum has zero mean and lies in a range of at most $\frac{\alpha_i}{m}$. Thus it follows from Hoeffding's inequality that the claimed probability is bounded by

$$2\exp\left(\frac{-2\epsilon^2}{\sum_{i\in[k]} \sum_{j\in[m]} \frac{\alpha_i^2}{m^2}} \right),$$

just as claimed. This bound is intuitively pleasing in the sense that it gets better as α becomes more uniform. For example, for a uniform α the bound becomes $2\exp(2km\epsilon^2)$ which would be the same as in the unweighted case with km samples.

We now want to investigate whether a bound as in eq. (7) can be achieved uniformly in α and h. To that end define

$$\begin{aligned} f(S) &:= \sup_{\alpha\in\Delta} \sup_{h\in H} \frac{1}{\|\alpha\|_2} \left[\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h) \right] = \sup_{\alpha\in\Delta} \sup_{h\in H} \sum_{i\in[k]} \frac{\alpha_i}{\|\alpha\|_2} \left[\mathrm{er}_i(h) - \widehat{\mathrm{er}}_{S_i}(h) \right] \\ &= \sup_{\alpha\in\Delta} \sup_{h\in H} \sum_{i\in[k]} \sum_{j\in[m]} \frac{\alpha_i}{m\|\alpha\|_2} \left[\mathrm{er}_i(h) - l(h(x_j^i), f_i(x_j^i)) \right] \\ &= \sup_{h\in H} \sum_{i\in[k]} \sum_{j\in[m]} \frac{\alpha^*(S)_i}{m\|\alpha^*(S)\|_2} \left[\mathrm{er}_i(h) - l(h(x_j^i), f_i(x_j^i)) \right] \end{aligned}$$

where $\alpha^*(S)$ is a maximizer, the existence of which is guaranteed by compactness. Now,

$$\left| f(S) - \inf_{x\in\mathcal{X}} f(S_{(j,i),x}) \right| \leq \frac{\alpha^*(S)_i}{m\|\alpha^*(S)\|}$$

and thereby

$$\sum_{(i,j)\in[k]\times[m]} \left(f(S) - \inf_{x\in\mathcal{X}} f(S_{(j,i),x}) \right)^2 \leq \sum_{i\in[k]} \frac{\alpha^*(S)_i^2}{m\|\alpha^*(S)\|_2^2} = \frac{1}{m}.$$

It now follows from Theorem 2.8 that with a probability of at least $1-\delta/2$

$$\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h) \leq \|\alpha\|_2 \left(\mathbf{E}\, f + \sqrt{\frac{2\log 2/\delta}{m}} \right),$$

uniformly for all α, h. It remains to bound $\mathbf{E}\, f$. Firstly since $\|\alpha\|_2 \geq \frac{1}{\sqrt{k}}$,

$$\mathbf{E}_S\, f(S) \leq \sqrt{k}\, \mathbf{E}_S \sup_{h\in H} \max_{i\in[k]} \left(\mathrm{er}_i(h) - \widehat{\mathrm{er}}_{S_i}(h) \right)$$

and in terms of Rademacher variables σ this can be symmetrically bounded by

$$\sqrt{k}\, \mathbf{E}_S\, \mathbf{E}_\sigma \sup_{h,i} \frac{1}{m} \sum_{j\in[m]} \sigma_j h(x_j^i) \leq \sqrt{\frac{2kd\log\frac{em}{d}}{m}}$$

where Massart's and Sauer's Lemmata were used. The exact same analysis can be applied to achieve the identical uniform bound on $\widehat{\mathrm{er}}_\alpha(h) - \mathrm{er}_\alpha(h)$ and therefore after a union bound we find that with a probability of at least $1-\delta$

$$|\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h)| \leq \|\alpha\|_2 \left(\sqrt{\frac{2kd \log \frac{em}{d}}{m}} + \sqrt{\frac{2 \log 2/\delta}{m}} \right) \tag{8}$$

uniformly for all α, h. This bound is unsatisfactory since the $\sqrt{k}$ factor just about cancels the rate improving weight factor, even in the optimal case of uniform weights.

So can we improve on that $\sqrt{k}$ factor? To make things simpler, we consider the case of a single hypothesis h. After defining the vector $\mathrm{er} = (|\mathrm{er}_i(h) - \widehat{\mathrm{er}}_{S_i}(h)|)_{i=1}^k$ the target probability simplifies to

$$\Pr\left[\frac{1}{\|\alpha\|_2} |\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h)| \geq \epsilon \quad \text{for some } \alpha \in \Delta \right] = \Pr\left[\sup_{\alpha \in \Delta} \left| \langle \frac{\alpha}{\|\alpha\|_2}, \mathrm{er} \rangle \right| \geq \epsilon \right].$$

On the one hand side by Cauchy–Schwarz $\left| \langle \frac{\alpha}{\|\alpha\|_2}, \mathrm{er} \rangle \right| \leq \|\mathrm{er}\|_2$ this probability is at most $\Pr[\|\mathrm{er}\|_2 \geq \epsilon]$. In fact, this would be an equality if we didn't have the constraint of $\alpha_i \geq 0$ for all $i \in [k]$. But it turns out that this constraint only changes the supremum by a constant factor. Explicitly, if $\|\mathrm{er}\|_2 \geq \sqrt{2}\epsilon$, then either either $\sum_{i \in [k], \mathrm{er}_i > 0} \mathrm{er}_i^2 \geq \epsilon^2$ or $\sum_{i \in [k], \mathrm{er}_i < 0} \mathrm{er}_i^2 \geq \epsilon^2$. In, say, the former case we could then define

$$\alpha_i := \begin{cases} \mathrm{er}_i \,/ \sum_{j \in [k], \mathrm{er}_j > 0} \mathrm{er}_j & \text{if } \mathrm{er}_i > 0, \\ 0 & \text{else} \end{cases}$$

satisfying $\alpha \in \Delta$. But then

$$\left| \langle \frac{\alpha}{\|\alpha\|_2}, \mathrm{er} \rangle \right| = \sum_{i \in [k], \mathrm{er}_i > 0} \mathrm{er}_i^2 \Big/ \sqrt{\sum_{i \in [k], \mathrm{er}_i > 0} \mathrm{er}_i^2} \geq \epsilon$$

and thus

$$\Pr\left[\|\mathrm{er}\|_2 \geq \epsilon \right] \geq \Pr\left[\sup_{\alpha \in \Delta} \left| \langle \frac{\alpha}{\|\alpha\|_2}, \mathrm{er} \rangle \right| \geq \epsilon \right] \geq \Pr\left[\|\mathrm{er}\|_2 \geq \sqrt{2}\epsilon \right].$$

This can be written out as

$$\Pr\left[\frac{1}{\|\alpha\|_2} |\mathrm{er}_\alpha(h) - \widehat{\mathrm{er}}_\alpha(h)| \geq \epsilon \quad \text{for some } \alpha \in \Delta \right] \geq \Pr\left[\sum_{i \in [k]} (\mathrm{er}_i(h) - \widehat{\mathrm{er}}_{S_i}(h))^2 \geq 2\epsilon^2 \right]$$

which means that by independence of the terms in the sum, any successful generalization bound ϵ has to grow with k as $\sqrt{k}$. Therefore the $\sqrt{k}$ factor from eq. (8) can't be improved on, in general.

Bibliography

[1] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

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