## A Rademacher complexity

**Definition 6** (Rademacher complexity)**.** Given a family of functions $\mathcal{F}$ and a training set $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$, *the Rademacher complexity of* $\mathcal{F}$ *conditioned on* $\mathbf{Y}'$ is given by

$$\widehat{\mathfrak{R}}_{\mathbf{Z}}(\mathcal{F}) = \mathbb{E}_{\mathbf{Z},\sigma}\left[\max_{f\in\mathcal{F}} \frac{1}{m}\sum\nolimits_{i=1}^{m} \sigma_i f(Z_i)\middle|\mathbf{Y}'\right]$$

where $\sigma_1, \ldots, \sigma_m$ are i.i.d. random variables uniform on $\{-1, +1\}$. *The Rademacher complexity* of $\mathcal{F}$ for sample size $m$ is given by

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_{\mathbf{Y}'}\left[\widehat{\mathfrak{R}}_{\mathbf{Z}}(\mathcal{F})\right].$$

The Rademacher complexity has been studied for a variety of function classes. For instance, for the linear hypothesis space $\mathcal{H} = \{x \to w^\top x, \|w\|_2 \leq \Lambda\}$, $\widehat{\mathfrak{R}}_{\mathbf{Z}}$ can be upper bounded by $\widehat{\mathfrak{R}}_{\mathbf{Z}}(\mathcal{H}) \leq \frac{\Lambda}{\sqrt{m}} \max_i \|Z_i\|_2$. As another example, the hypothesis class of ReLu feed-forward neural networks with $d$ layers and weight matrices $W_k$ such that $\prod_{k=1}^{d} \|W\|_F \leq \gamma$ verifies $\widehat{\mathfrak{R}}_{\mathbf{Z}}(\mathcal{H}) \leq \frac{2^{d-1/2}\gamma}{\sqrt{m}} \max_i \|Z_i\|_2$ (Neyshabur et al., 2015).

## B Discrepancy analysis

**Proposition 1.** *Let* $\mathcal{H}$ *be a hypothesis space and let* $L$ *be a bounded loss function which respects the triangle inequality. Let* $h' \in \mathcal{H}$*. Then,*

$$\Delta \leq \Delta_s + \mathcal{L}(h \mid \mathbf{Y}) + \mathcal{L}(h \mid \mathbf{Y}')$$

*Proof.* Let $h, h' \in \mathcal{H}$. For ease of notation, we write

$$\begin{aligned}\Delta_s(h, h', \mathbf{Y}') =& \frac{1}{m}\sum_i L(h(Y_1^T(i)), h'(Y_1^T(i))) \\ & - \frac{1}{m}\sum_i L(h(Y_1^{T-1}(i)), h'(Y_1^{T-1}(i))).\end{aligned}$$

Applying the triangle inequality to $L$,

$$\begin{aligned}\mathcal{L}(h \mid \mathbf{Y}) =& \frac{1}{m}\sum_i \mathbb{E}[L(h(Y_1^T(i)), Y_{T+1}(i)) \mid \mathbf{Y}] \\ \leq& \frac{1}{m}\sum_i L(h(Y_1^T(i)), h'(Y_1^T(i))) \\ & + \frac{1}{m}\sum_i \mathbb{E}[L(h'(Y_1^T(i)), Y_{T+1}(i)) \mid \mathbf{Y}] \\ =& \frac{1}{m}\sum_i L(h(Y_1^T(i)), h'(Y_1^T(i))) + \mathcal{L}(h' \mid \mathbf{Y}).\end{aligned}$$

Then, by definition of $\Delta_s(h, h', \mathbf{Y}')$, we have

$$\begin{aligned}\mathcal{L}(h \mid \mathbf{Y}) \leq& \frac{1}{m}\sum_i L(h(Y_1^T(i)), h'(Y_1^T(i))) \\ & - \frac{1}{m}\sum_i L(h(Y_1^{T-1}(i)), h'(Y_1^{T-1}(i))) \\ & + \frac{1}{m}\sum_i L(h(Y_1^{T-1}(i)), h'(Y_1^{T-1}(i))) \\ & + \mathcal{L}(h' \mid \mathbf{Y}) \\ \leq& \Delta_s(h, h', \mathbf{Y}') + \mathcal{L}(h' \mid \mathbf{Y}) \\ & + \frac{1}{m}\sum_i L(h(Y_1^{T-1}(i)), h'(Y_1^{T-1}(i))).\end{aligned}$$

By an application of the triangle inequality to $L$,

$$\begin{aligned}\mathcal{L}(h, \mathcal{D}) \leq& \Delta_s(h, h', \mathbf{Y}') + \mathcal{L}(h' \mid \mathbf{Y}) \\ & + \frac{1}{m}\sum_i \mathbb{E}[L(h(Y_1^{T-1}(i)), Y_T(i)) \mid \mathbf{Y}'] \\ & + \frac{1}{m}\sum_i \mathbb{E}[L(h'(Y_1^{T-1}(i)), Y_T(i)) \mid \mathbf{Y}'] \\ =& \Delta_s(h, h', \mathbf{Y}') + \mathcal{L}(h' \mid \mathbf{Y}) + \mathcal{L}(h \mid \mathbf{Y}') \\ & + \mathcal{L}(h' \mid \mathbf{Y}').\end{aligned}$$

Finally, we obtain

$$\begin{aligned}\mathcal{L}(h \mid \mathbf{Y}) - \mathcal{L}(h \mid \mathbf{Y}') \leq& \Delta_s(h, h', \mathbf{Y}') + \mathcal{L}(h' \mid \mathbf{Y}) \\ & + \mathcal{L}(h' \mid \mathbf{Y}')\end{aligned}$$

and the result announced in the theorem follows by taking the supremum over $\mathcal{H}$ on both sides. □

**Proposition 2.** *Let* $I_1, \cdots, I_k$ *be a partition of* $\{1, \ldots, m\}$*, and* $C_1, \ldots, C_k$ *be the corresponding partition of* $\mathbf{Y}$*. Write* $c = \min_j |C_j|$*. Then we have with probability* $1 - \delta$*,*

$$\begin{aligned}\Delta_s \leq& \Delta_e + \max\left(\max_j \mathfrak{R}_{|C_j|}(\widetilde{C}_j'), \max_j \mathfrak{R}_{|C_j|}(\widetilde{I}_j)\right) \\ & + \sqrt{\frac{1}{2c}\log\frac{2k}{\delta - \sum_j (|I_j|-1)[\bar{\beta}(I_j) + \bar{\beta}'(I_j)]}}.\end{aligned}$$

*Proof.* By definition of $\Delta_s$,

$$\begin{aligned}\Delta_s = \sup_{h,h'\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \Big[L(h(Y_1^T(i)), h'(Y_1^T(i))) \\ - L(h(Y_1^{T-1}(i)), h'(Y_1^{T-1}(i)))\Big] \\ \le \sup_{h,h'\in\mathcal{H}} \Big[\frac{1}{m}\sum_{i=1}^m L(h(Y_1^T(i)), h'(Y_1^T(i))) \\ - \mathbb{E}_Y[L(h(Y_1^T), h'(Y_1^T))]\Big] \\ + \sup_{h,h'\in\mathcal{H}} \Big[\mathbb{E}_Y[L(h(Y_1^T), h'(Y_1^T))] \\ - \mathbb{E}_Y[L(h(Y_1^{T-1}), h'(Y_1^{T-1}))]\Big] \\ + \sup_{h,h'\in\mathcal{H}} \Big[\mathbb{E}_Y[L(h(Y_1^{T-1}), h'(Y_1^{T-1}))] \\ - \frac{1}{m}\sum_{i=1}^m L(h(Y_1^{T-1}(i)), h'(Y_1^{T-1}(i)))\Big]\end{aligned}$$

by sub-additivity of the supremum. Now, define

$$\begin{aligned}\phi(\mathbf{Y}) \triangleq \sup_{h,h'\in\mathcal{H}} \Big[\frac{1}{m}\sum_{i=1}^m L(h(Y_1^T(i)), h'(Y_1^T(i))) \\ - \mathbb{E}_Y[L(h(Y_1^T), h'(Y_1^T))]\Big]\end{aligned}$$

$$\begin{aligned}\psi(\mathbf{Y}') \triangleq \sup_{h,h'\in\mathcal{H}} \Big[\mathbb{E}_Y[L(h(Y_1^{T-1}), h'(Y_1^{T-1})) \\ - \frac{1}{m}\sum_{i=1}^m L(h(Y_1^{T-1}(i)), h'(Y_1^{T-1}(i)))]\Big].\end{aligned}$$

By definition of $\Delta_e$, we have from the previous inequality

$$\Delta_s \le \Delta_e + \phi(\mathbf{Y}_1^T) + \psi(\mathbf{Y}_1^{T-1}).$$

We now proceed to give a high-probability bound for $\phi$; the same reasoning will yield a bound for $\psi$. By sub-additivity of the max,

$$\begin{aligned}\phi(\mathbf{Y}) \le \sum_j \frac{|C_j|}{m} \sup_{h\in\mathcal{H}} \Big[\mathbb{E}_Y[f(h, Y_1^T)] \\ - \frac{1}{|C_j|}\sum_{Y\in C_j} f(h, Y_1^T)\Big] \\ \le \sum_j \frac{|C_j|}{m}\phi(C_j)\end{aligned}$$

and so by union bound, for $\epsilon > 0$

$$\Pr(\phi(\mathbf{Y}) > \epsilon) \le \sum_j \Pr(\phi(C_j) > \epsilon).$$

Let $\epsilon > \max_j \mathbb{E}[\phi(\widetilde{C}_j)]$ and set $\epsilon_j = \epsilon - \mathbb{E}[\phi(\widetilde{C}_j)]$.

Define for time series $Y(i), Y(j)$ the mixing coefficient

$$\bar\beta(i,j) = \|\Pr(Y_1^T(i), Y_1^T(j)) - \Pr(Y_1^T(i))\Pr(Y_1^T(j))\|_{TV}$$

where we also extend the usual notation to $\bar\beta(C_j)$.

$$\begin{aligned}\Pr\left(\phi(C_j) > \epsilon\right) = \Pr\left(\phi(C_j) - \mathbb{E}[\phi(\widetilde{C}_j)] > \epsilon_j\right) \\ \overset{(a)}{\le} \Pr\Big(\phi(\widetilde{C}_j) \\ - \mathbb{E}[\phi(\widetilde{C}_j)] > \epsilon_j\Big) + (|I_j|-1)\bar\beta(I_j) \\ \overset{(b)}{\le} e^{-2c\epsilon_j^2} + (|I_j|-1)\bar\beta(I_j),\end{aligned}$$

where (a) follows by applying Prop. 6 to the indicator function of the event $\Pr(\phi(C_j) - \mathbb{E}[\phi(\widetilde{C}_j)] \ge \epsilon)$, and (b) is a direct application of McDiarmid's inequality to $\phi(\widetilde{C}_j) - \mathbb{E}[\phi(\widetilde{C}_j)]$.

Hence, by summing over $j$ we obtain

$$\begin{aligned}\Pr\left(\phi(\mathbf{Y}) > \epsilon\right) \le k e^{-2\min_j |C_j|(\epsilon - \max_j \mathbb{E}[\phi(\widetilde{C}_j)])^2} \\ + \sum_j (|I_j|-1)\bar\beta(I_j)\end{aligned}$$

and similarly

$$\begin{aligned}\Pr\left(\psi(\mathbf{Y}') > \epsilon\right) \le k e^{-2\min_j |C'_j|(\epsilon - \max_j \mathbb{E}[\psi(\widetilde{C}'_j)])^2} \\ + \sum_j (|I_j|-1)\bar\beta'(I_j),\end{aligned}$$

which finally yields

$$\begin{aligned}\Pr(\Delta_s - \Delta_e > \epsilon) \le \Pr(\phi(\mathbf{Y}) > \epsilon) + \Pr(\psi(\mathbf{Y}') > \epsilon) \\ \le 2k\exp(-2c(\epsilon - \max(\max_j \mathbb{E}[\phi(\widetilde{C}_j)], \max_j \mathbb{E}[\psi(\widetilde{C}'_j)]))^2) \\ + \sum_j (|I_j|-1)[\bar\beta(I_j) + \bar\beta'(I'_j)],\end{aligned}$$

where we recall that we write $c = \min_j |C_j|$. We invert the previous equation by setting

$$\begin{aligned}\epsilon = \max(\max_j \mathbb{E}[\phi(\widetilde{C}_j)], \max_j \mathbb{E}[\psi(\widetilde{C}'_j)]) \\ + \sqrt{\frac{1}{2c}\log\frac{2k}{\delta - \sum_j (|I_j|-1)[\beta(I_j) + \beta(I'_j)]}},\end{aligned}$$

yielding with probability $1-\delta$,

$$\begin{aligned}\Delta_s \le \Delta_e + \max(\max_j \mathbb{E}[\phi(\widetilde{C}_j)], \max_j \mathbb{E}[\psi(\widetilde{C}'_j)]) \\ + \sqrt{\frac{1}{2c}\log\frac{2k}{\delta - \sum_j (|I_j|-1)[\bar\beta(I_j) + \bar\beta'(I_j)]}}.\end{aligned}$$

We now bound $\mathbb{E}[\phi(\widetilde{C}_j)]$ by $\mathfrak{R}_{|C_j|}(\widetilde{C}_j)$. A similar argument yields the bound for $\psi$. By definition, we have

$$\begin{aligned}\mathbb{E}[\phi(\widetilde{C}_j)] &= \mathbb{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{|C_j|}\sum_{Z\in\widetilde{C}_j} f(h, Y_1^T(i)) - \mathbb{E}_Y[f(h, Y_1^T)]\Big] \\ &= \frac{1}{|C_j|}\mathbb{E}\Big[\sup_{h\in\mathcal{H}}\sum_{Z\in\widetilde{C}_j} \underbrace{f(h, Y_1^T(i)) - \mathbb{E}_Y[f(h, Y_1^T)]}_{g(h, Y_1^T(i))}\Big] \\ &= \frac{1}{|C_j|}\mathbb{E}\Big[\sup_{h\in\mathcal{H}}\sum_{Z\in\widetilde{C}_j} g(h, Y_1^T(i))\Big]\end{aligned}$$

Standard symmetrization arguments as those used for the proof of the famous result by Koltchinskii and Panchenko (2002), which hold also when data is drawn independently but not identically at random, yield

$$\mathbb{E}[\phi(\widetilde{C}_j)] \leq \mathfrak{R}_{|C_j|}(\widetilde{C}_j).$$

The same argument yields for $\psi$

$$\mathbb{E}[\psi(\widetilde{C}'_j)] \leq \mathfrak{R}_{|C_j|}(\widetilde{C}'_j).$$

To conclude our proof, it only remains to prove the bound

$$\begin{aligned}\bar{\beta}(i,j) \leq& \beta_{\mathrm{s2s}}(i,j) \\ &+ \mathbb{E}_{\mathbf{Y}'}\Big[\mathrm{Cov}\Big(\Pr(Y_T(i) \mid \mathbf{Y}'), \Pr(Y_T(j) \mid \mathbf{Y}')\Big)\Big]\end{aligned}$$

Let $Y(i), Y(j)$ be two time series, and write $X_i = \mathbb{E}[\Pr(Y_1^T(i)) \mid \mathbf{Y}']$. Then the following bound holds

$$\begin{aligned}\bar{\beta}(i,j) =& \|\Pr(Y_1^T(i), Y_1^T(j)) - \Pr(Y_1^T(i))\Pr(Y_1^T(j))\|_{TV} \\ =& \|\mathbb{E}[\Pr(Y_1^T(i), Y_1^T(j)) \mid \mathbf{Y}'] - \mathbb{E}[X_i]\mathbb{E}[X_j]\|_{TV} \\ =& \|\mathbb{E}[\Pr(Y_1^T(i), Y_1^T(j)) \mid \mathbf{Y}_1^{T-1}] - \mathbb{E}[X_i, X_j] \\ & - \mathbb{E}[\mathrm{Cov}(X_i, X_j)]\|_{TV} \\ \leq& \beta_{\mathrm{s2s}}(i,j) + \mathbb{E}_{\mathbf{Y}'}[\mathrm{Cov}(X_i, X_j)],\end{aligned}$$

which is the desired inequality. □

We now show two useful lemmas for various specific cases of time series and hypothesis spaces.

**Proposition 3.** *If $Y(i)$ is stationary for all $1 \leq i \leq m$, and $\mathcal{H}$ is a hypothesis space such that $h \in \mathcal{H} : \mathcal{Y}^{T-1} \to \mathcal{Y}$ (i.e. the hypotheses only consider the last $T-1$ values of $Y$), then $\Delta_e = 0$.*

*Proof.* Let $h, h' \in \mathcal{H}$. For stationary $Y(i)$, we have $\Pr(Y_1^T(i)) = \Pr(Y_2^T(i))$, and so

$$\mathbb{E}[L(h(Y_2^T), h'(Y_2^T))] - \mathbb{E}[L(h(Y_1^{T-1}), h'(Y_1^{T-1}))] = 0$$

and so taking the supremum over $h, h'$ yields the desired result. □

**Proposition 4.** *If $Y(i)$ is covariance stationary for all $1 \leq i \leq m$, $L$ is the squared loss, and $\mathcal{H}$ is a linear hypothesis space $\{x \to w \cdot x \mid \|w\| \in \mathbb{R}^p \leq \Lambda\}$, then $\Delta_e = 0$.*

*Proof.* Recall that a time series $Y$ is covariance stationary if $\mathbb{E}_Y[Y_t]$ does not depend on $t$ and $\mathbb{E}_Y[Y_tY_s] = f(t-s)$ for some function $f$.

Let now $(h, h') \in \mathcal{H} \equiv (w, w') \in \mathbb{R}^p$. We write $\Sigma = \Sigma_2^T(Y) = \Sigma_1^T(Y)$ the covariance matrix of $Y$ where the equality follows from covariance stationarity. Without loss of generality, we consider $p = T-1$. Then,

$$\begin{aligned}&\mathbb{E}[L(h(Y_2^T), h'(Y_2^T))] - \mathbb{E}[L(h(Y_1^{T-1}), h'(Y_1^{T-1}))] \\ &\quad = \mathbb{E}[((w-w')^\top \Sigma_2^T(Y)(w-w')] \\ &\quad\quad - \mathbb{E}[((w-w')^\top \Sigma_1^{T-1}(Y)(w-w')] \\ &\quad = 0.\end{aligned}$$

Taking the supremum over $h, h'$ yields the desired result. □

**Proposition 5.** *If the $Y(i)$ are periodic of period $p$ and the observed starting time of each $Y(i)$ is distributed uniformly at random in $[p]$, then $\Delta_e = 0$.*

*Proof.* This proof is similar to the stationary case: indeed, we can write $\Pr(Y_1^{T-1}(i)) = \frac{1}{p}\Pr(Y(i))$ due to the uniform distribution on starting times. Then, by the same reasoning, we have also

$$\Pr(Y_2^T(i)) = \frac{1}{p}\Pr(Y(i)) = \Pr(Y_1^{T-1}(i)),$$

from which the result follows. □

## C Generalization bounds

**Proposition 6.** Yu (1994, Corollary 2.7). *Let $f$ be a real-valued Borel measurable function such that $0 \leq f \leq 1$. Then, we have the following guarantee:*

$$\Big|\mathbb{E}[f(\widetilde{C})] - \mathbb{E}[f(C)]\Big| \leq (|C|-1)\beta,$$

*where $\beta$ is the total variation distance between joint distributions of $C$ and $\widetilde{C}$.*

**Theorem 4.1.** *Let $\mathcal{H}$ be a hypothesis space, and $h \in \mathcal{H}$. Let $C_1, \ldots, C_k$ form a partition of the training input $\mathbf{Y}_1^T$, and consider that the loss function $L$ is bounded by 1. Then, we have for $\delta > 0$, with probability $1-\delta$,*

$$\begin{aligned}\Phi_{s2s}(h) \leq& \Delta + \max_j \Big[\mathfrak{R}_{|C_j|}(\widetilde{C}_j \mid \mathbf{Y})\Big] \\ &+ \frac{1}{\sqrt{2\min_j |I_j|}}\sqrt{\log\left(\frac{k}{\delta - \sum_j(|I_j|-1)\beta_{s2s}(I_j)}\right)}.\end{aligned}$$

For ease of notation, we write

$$\begin{aligned}\phi(\mathbf{Y}) &= \sup_{h\in\mathcal{H}} \mathcal{L}(h \mid \mathbf{Y}') - \widehat{\mathcal{L}}(h, \mathbf{Y}) \\ &= \sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[f(h, Y_1^T(i)) \mid \mathbf{Y}'] \\ &\quad - \frac{1}{m}\sum_{i=1}^{m} f(h, Y_1^T(i)).\end{aligned}$$

We begin by proving the following lemma.

**Lemma 3.** *Let $\bar{\mathbf{Y}}$ be equal to $\mathbf{Y}$ on all time series except for the last, where we have $\bar{Y}(m) = Y(m)$ at all times except for time $t = T$. Then*

$$\left|\phi(\mathbf{Y}) - \phi(\bar{\mathbf{Y}})\right| \leq \frac{1}{m}$$

*Proof.* Fix $h^* \in \mathcal{H}$. Then,

$$\begin{aligned}&\mathcal{L}(h^* \mid \mathbf{Y}') - \widehat{\mathcal{L}}(h^*, \mathbf{Y}) - \sup_{h\in\mathcal{H}}\Big[\mathcal{L}(h \mid \bar{\mathbf{Y}}') - \widehat{\mathcal{L}}(h, \bar{\mathbf{Y}})\Big] \\ &\leq \mathcal{L}(h^* \mid \mathbf{Y}') - \widehat{L}(h^*, \mathbf{Y}) \\ &- \Big[\mathcal{L}(h^*, \mid \bar{\mathbf{Y}}') - \widehat{\mathcal{L}}(h^*, \bar{\mathbf{Y}})\Big] \\ &\overset{(a)}{\leq} \widehat{\mathcal{L}}(h^*, \bar{\mathbf{Y}}) - \widehat{\mathcal{L}}(h^*, \mathbf{Y}) \\ &\leq \frac{1}{m}\Big[f(h^*, \bar{Y}_1^T(m)) - f(h^*, Y_1^T(m))\Big] \leq \frac{1}{m}.\end{aligned}$$

where (a) follows from the fact that $\mathbf{Y}' = \bar{\mathbf{Y}}'$ and the last inequality follows from the fact that $f$ is bounded by 1.

By taking the supremum over $h^*$, the previous calculations show that $\phi(\mathbf{Y}) - \phi(\bar{\mathbf{Y}}) \leq 1/m$; by symmetry, we obtain $\phi(\bar{\mathbf{Y}}) - \phi(\mathbf{Y}) \leq 1/m$ which proves the lemma. ☐

We now prove the main theorem.

*Proof.* Observe that the following bounds holds

$$\begin{aligned}\Phi_{\mathrm{s2s}}(\mathbf{Y}) &= \mathcal{L}(h \mid \mathbf{Y}) - \widehat{\mathcal{L}}(h, \mathbf{Y}) \\ &\leq \sup_{h\in\mathcal{H}}\Big[\mathcal{L}(h \mid \mathbf{Y}) - \mathcal{L}(h \mid \mathbf{Y}')\Big] \\ &\quad + \sup_{h\in\mathcal{H}}\Big[\mathcal{L}(h \mid \mathbf{Y}') - \widehat{\mathcal{L}}(h, \mathbf{Y})\Big].\end{aligned}$$

and so

$$\Phi_{\mathrm{s2s}}(\mathbf{Y}) - \Delta \leq \underbrace{\sup_{h\in\mathcal{H}} \mathcal{L}(h, \mid \mathbf{Y}') - \widehat{\mathcal{L}}(h, \mathbf{Y})}_{\phi(\mathbf{Y})}.$$

Define $M = \max_j \mathbb{E}[\phi(\widetilde{C}_j) \mid \widetilde{\mathbf{Y}'}]$. Then,

$$\begin{aligned}&\Pr\Big(\Phi_{\mathrm{s2s}}(\mathbf{Y}) - \Delta - M > \epsilon \mid \mathbf{Y}'\Big) \\ &\leq \Pr(\phi(\mathbf{Y}) - M > \epsilon \mid \mathbf{Y}').\end{aligned} \tag{C.1}$$

By sub-additivity of the supremum, we have

$$\phi(\mathbf{Y}) - M \leq \sum_j \frac{|C_j|}{m} \sup_{h\in\mathcal{H}}\Big[\mathcal{L}(h \mid \mathbf{Y}) - \widehat{\mathcal{L}}(h, C_j) - M\Big]$$

and so by union bound,

$$\Pr(\phi(\mathbf{Y}) - M \geq \epsilon \mid \mathbf{Y}') \leq \sum_j \Pr(\phi(C_j) - M \geq \epsilon \mid \mathbf{Y}').$$

By definition of $M$,

$$\begin{aligned}&\Pr\Big(\phi(C_j) - M \geq \epsilon \mid \mathbf{Y}'\Big) \\ &\leq \Pr(\phi(C_j) - \mathbb{E}[\phi(\widetilde{C}_j) \mid \widetilde{\mathbf{Y}'}] \geq \epsilon \mid \mathbf{Y}') \\ &\overset{(a)}{\leq} \Pr(\phi(\widetilde{C}_j) \\ &\quad - \mathbb{E}[\phi(\widetilde{C}_j) \mid \widetilde{\mathbf{Y}'}] \geq \epsilon \mid \mathbf{Y}') + (|I_j| - 1)\beta_{\mathrm{s2s}}(I_j \mid \mathbf{Y}') \\ &\overset{(b)}{\leq} e^{-2|C_j|\epsilon^2} + (|I_j| - 1)\beta_{\mathrm{s2s}}(I_j \mid \mathbf{Y}').\end{aligned}$$

where (a) follows by applying Prop. 6 to the indicator function of the event $\Pr(\phi(C_j) - \mathbb{E}[\phi(\widetilde{C}_j) \mid \widetilde{\mathbf{Y}'}] \geq \epsilon)$, and (b) is a direct application of McDiarmid's inequality, following Lemma 3. The notation $\beta_{\mathrm{s2s}}(I_j \mid \mathbf{Y}')$ indicates the total variation distance between the joint distributions of $C_j$ and $\widetilde{C}_j$ conditioned on $\mathbf{Y}'$. In particular, we have $\mathbb{E}_{\mathbf{Y}'}\beta_{\mathrm{s2s}}(C_j \mid \mathbf{Y}') = \beta_{\mathrm{s2s}}(C_j)$.

Finally, taking the expectation of the previous term over all possible $\mathbf{Y}'$ values and summing over $j$, we obtain

$$\begin{aligned}&\Pr(\mathcal{L}(h \mid \mathbf{Y}) - \widehat{\mathcal{L}}(h, \mathbf{Y}) - \mathbb{E}_{\widetilde{C}_{j'}}[\phi(\widetilde{C}_{j'}) \mid \widetilde{\mathbf{Y}}] \geq \epsilon) \\ &\leq \sum_j e^{-2|C_j|\epsilon^2} + \sum_j (|I_j| - 1)\beta_{\mathrm{s2s}}(I_j).\end{aligned}$$

Combining this bound with (C.1), we obtain

$$\begin{aligned}&\Pr\Big(\Phi_{\mathrm{s2s}}(\mathbf{Y}) - \Delta - M > \epsilon\Big) \\ &\leq \sum_j e^{-2|C_j|\epsilon^2} + \sum_j (|I_j| - 1)\beta_{\mathrm{s2s}}(I_j) \\ &\leq k e^{-2\min_j |C_j|\epsilon^2} + \sum_j (|I_j| - 1)\beta_{\mathrm{s2s}}(I_j)\end{aligned}$$

We invert the previous equation by choosing $\delta > \sum_j (|I_j| - 1)\beta_{\mathrm{s2s}}(I_j)$ and setting

$$\epsilon = \sqrt{\frac{\log \frac{k}{\delta - \sum_j (|I_j| - 1)\beta_{\mathrm{s2s}}(I_j)}}{2\min_j |I_j|}},$$

which yields that with probability $1 - \delta$, we have

$$\Phi_{\mathrm{s2s}}(Z) \leq M + \Delta + \sqrt{\frac{\log\left(\frac{k}{\delta - \sum_j (|I_j| - 1)\beta(I_j)}\right)}{2\min_j |I_j|}}.$$

To conclude our proof, it remains to show that

$$M \leq \mathfrak{R}_{|C_j|}(\widetilde{C}_j \mid \widetilde{\mathbf{Y}'}).$$

$$\begin{aligned}\mathbb{E}[\phi(\widetilde{C}_j) \mid \widetilde{\mathbf{Y}'}] =& \mathbb{E}\Big[\sup_{h\in\mathcal{H}} \mathcal{L}(h \mid \widetilde{\mathbf{Y}'}) \\ &- \frac{1}{|C_j|}\sum_{i=1}^m f(h, \widetilde{Y}_1^T(i)) \mid \widetilde{\mathbf{Y}'}\Big] \\ =& \frac{1}{|C_j|}\mathbb{E}\Big[\sup_{h\in\mathcal{H}} \sum_{\widetilde{Y}_1^T\in\widetilde{C}_j} \mathbb{E}[f(h,\widetilde{Y}_1^T) \mid \widetilde{\mathbf{Y}'}] \\ &- f(h, \widetilde{Y}_1^T(i)) \mid \widetilde{\mathbf{Y}'}\Big] \\ \leq& \frac{1}{|C_j|}\mathbb{E}\Big[\sup_{h\in\mathcal{H}} \sum_{\widetilde{Y}_1^T\in\widetilde{C}_j} g(h, \widetilde{Y}_1^T(i)) \mid \widetilde{\mathbf{Y}'}\Big]\end{aligned}$$

where we've defined

$$g(h, \widetilde{Y}_1^T(i)) \triangleq \mathbb{E}[f(h, \widetilde{Y}_1^T(i)) \mid \widetilde{\mathbf{Y}'}] - f(h, \widetilde{Y}_1^T(i)).$$

Similar arguments to those used at the end of Appendix B yield the desired result, which concludes the proof of Theorem 4.1. ☐

## D Generalization bounds for local models

**Theorem 5.1.** *Let $h = (h_1, \ldots, h_m)$ where each $h_i$ is a hypothesis learned via a local method to predict the univariate time series $Z_i$. For $\delta > 0$ and any $\alpha > 0$, we have w.p. with $1-\delta$*

$$\begin{aligned}\Phi_{loc}(\mathbf{Z}) \leq& \frac{1}{m}\sum_i \Delta(Y(i)) + 2\alpha \\ &+ \sqrt{\frac{2}{T}\log\frac{m\max_i(\mathbb{E}_{v\sim T(Y(i))}[\mathcal{N}_1(\alpha,\mathcal{F},v)])}{\delta}}\end{aligned}$$

*Proof.* Write

$$\begin{aligned}\Phi(Y_1^T(i)) =& \sup_{h\in\mathcal{H}} \mathbb{E}[f(h, Y_1^{T+1}) \mid Y_1^T] \\ &- \frac{1}{T}\sum_{t=1}^T f(h, Y_t^{t+T}(i)).\end{aligned}$$

By (Kuznetsov and Mohri, 2015, Theorem 1), we have that for $\epsilon > 0$, and $1 \leq i \leq m$,

$$\begin{aligned}\Pr(\Phi(Y_1^T(i) - \Delta(Y(i)) > \epsilon) \leq& \mathbb{E}_{v\sim T(p)}[\mathcal{N}_1(\alpha,\mathcal{F},v)] \\ &\times \exp\Big(-\frac{T(\epsilon-2\alpha)^2}{2}\Big).\end{aligned}$$

By union bound,

$$\begin{aligned}\Pr(\frac{1}{m}\sum_i &\Phi(Y_1^T(i)) - \Delta(Y(i)) > \epsilon) \\ &\leq m\max_i(\mathbb{E}_{v\sim T(Y(i))}[\mathcal{N}_1(\alpha,\mathcal{F},v)]) \\ &\quad\times \exp\Big(-\frac{T(\epsilon-2\alpha)^2}{2}\Big)\end{aligned}$$

We invert the previous equation by letting

$$\epsilon = 2\alpha + \sqrt{\frac{2}{T}\log\frac{m\max_i(\mathbb{E}_{v\sim T(Y(i))}[\mathcal{N}_1(\alpha,\mathcal{F},v)])}{\delta}}.$$

which yields the desired result. ☐

## E Analysis of expected mixing coefficients

**Lemma 2.** *Two AR processes $Y(i), Y(j)$ generated by (4.1) such that $\sigma = Cov(Y(i), Y(j)) \leq \sigma_0 < 1$ verify $\beta_{s2s}(i,j) = \max\left(\frac{3}{2(1-\sigma_0^2)}, \frac{1}{1-2\sigma_0}\right)\sigma$.*

*Proof.* For simplicity, we write $U = Y(i)$ and $V = Y(j)$.

Write

$$\begin{aligned}\beta =& \|P(U_T|\mathbf{Y}')P(V_T|\mathbf{Y}') - P(U_T, V_T|\mathbf{Y}')\|_{TV} \\ =& \sup_{u,v} |P(U_T = u)P(V_T = v) - P(U_T = u, V_T = v)| \\ =& \sup_{u,v} \Big| P(U_T = u \mid U_0^{T-1})P(V_T = v \mid v_0^{T-1}) \\ &- P(U_T = u, V_T = v \mid v_0^{T-1}, u_0^{T-1})\Big| \\ =& \sup_{u,v} \Big|\Big[P(u, v \mid U_0^{T-1}, V_0^{T-1}) + f(\sigma,\delta,\epsilon)\Big] \\ &- P(u, v \mid U_0^{T-1}, V_0^{T-1})\Big|\end{aligned}$$

where we've written $\delta = u - \Theta_i(U_0^{T-1})$ (and $\epsilon$ similarly for $v$), and we've defined

$$\begin{aligned}f(\sigma,\delta,\epsilon) &= P(u|U_0^{T-1})P(v|V_0^{T-1}) - P(u,v|U_0^{T-1},V_0^{T-1}) \\ &= e^{-\frac{1}{2}(\delta^2+\epsilon^2)} - \frac{1}{1-\sigma^2}e^{-\frac{1}{2}\frac{1}{1-\sigma^2}(\delta^2+\epsilon^2-2\sigma\epsilon\delta)}.\end{aligned}$$

Assuming we can bound $f(\sigma,\delta,\epsilon)$ by a function $g(\sigma)$ independent of $\delta,\epsilon$, we can then derive a bound on $\beta$.

Let $x = \sqrt{\delta^2+\epsilon^2}$ be a measure of how far the AR process noises lie from their mean $\mu = 0$. Using the inequality

$$|\delta\epsilon| \leq \delta^2 + \epsilon^2,$$

we proceed to bound $|f(\sigma,\delta,\epsilon)|$ by bounding $f$ and $-f$.

$$
\begin{aligned}
f(\sigma,\delta,\epsilon) &\le e^{-\frac{1}{2}(\delta^2+\epsilon^2)} - e^{-\frac{1}{2}\frac{1}{1-\sigma^2}(\delta^2+\epsilon^2+2\sigma|\delta\epsilon|)} \\
&\le e^{-\frac{1}{2}x^2} - e^{-\frac{1}{2}\frac{1}{1-\sigma^2}(1+2\sigma)x^2} \\
&\le e^{-\frac{1}{2}x^2}\left(1 - e^{-\frac{1}{2}\frac{2\sigma+\sigma^2}{1-\sigma^2}x^2}\right)
\end{aligned}
$$

Using the inequality $1 - x \le e^{-x}$, it then follows that

$$
\begin{aligned}
f(\sigma,\delta,\epsilon) &\le e^{-\frac{1}{2}x^2}(1-(1-\tfrac{1}{2}\tfrac{2\sigma+\sigma^2}{1-\sigma^2}x^2)) \\
&\le \tfrac{1}{2}\tfrac{3}{1-\sigma^2}\sigma x^2 e^{-\frac{1}{2}x^2} \\
&\overset{(a)}{\le} \tfrac{3}{e(1-\sigma^2)}\sigma
\end{aligned} \tag{E.1}
$$

where inequality $(a)$ follows from the fact that $y \to ye^{-y}$ is bounded by $1/e$.

Similarly, we now bound $-f$:

$$
\begin{aligned}
-f(\sigma,\delta,\epsilon) &\le \frac{1}{1-\sigma^2}e^{-\frac{1}{2}\frac{1}{1-\sigma^2}(\delta^2+\epsilon^2-2\sigma|\epsilon\delta|)} - e^{-\frac{1}{2}(\delta^2+\epsilon^2)} \\
&\le \frac{1}{1-\sigma^2}e^{-\frac{1}{2}\frac{1-2\sigma}{1-\sigma^2}x^2} - e^{-\frac{1}{2}x^2} \\
&\le \frac{1}{1-\sigma^2}e^{-\frac{1}{2}(1-2\sigma)x^2} - e^{-\frac{1}{2}x^2}.
\end{aligned}
$$

One shows easily that this last function reaches its maximum for $x_0^2 = \frac{1}{\sigma}\log(\frac{1-\sigma^2}{1-2\sigma})$, at which point it verifies

$$
-f(\sigma, x_0) = \frac{2\sigma}{1-2\sigma}e^{-\frac{1}{2\sigma}\log(\frac{1-\sigma^2}{1-2\sigma})} \le \frac{2\sigma}{1-2\sigma} \tag{E.2}
$$

Putting (E.1) and (E.2) together, we obtain

$$
\begin{aligned}
|f(\sigma,\delta,\epsilon)| &\le \sigma \max\left(\frac{3}{e(1-\sigma^2)}, \frac{1}{1-2\sigma}\right) \\
&\le \max\left(\frac{3}{2(1-\sigma_0^2)}, \frac{1}{1-2\sigma_0}\right)\sigma
\end{aligned}
$$

Taking the expectation over all possible realizations of $\mathbf{Y}'$ yields the desired result. □

*Proof.* Recall that $\mathbf{Y}$ contains $m' = mT$ examples, which we denote $Y_{t-p}^t(i)$ for $1 \le i \le m$ and $1 \le t \le T$ (when $t - p < 0$, we truncate the time series appropriately). We define

$$
\mathcal{L}_{\text{hyb}}(h \mid \mathbf{Y}) = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[L(h(Y_{T-p+1}^T(i)), Y_{T+1}(i)) \mid \mathbf{Y}]
$$

$$
\mathcal{L}_{\text{hyb}}(h \mid \mathbf{Y}') = \frac{1}{m}\sum_{i=1}^m \frac{1}{T}\sum_{t=1}^T \mathbb{E}[L(h(Y_{t-p}^{t-1}(i)), Y_t(i)) \mid \mathbf{Y}']
$$

$$
\widehat{\mathcal{L}_{\text{hyb}}}(h) = \frac{1}{m}\sum_{i=1}^m \frac{1}{T}\sum_{t=1}^T L(h(Y_{t-p}^{t-1}(i)), Y_t(i))
$$

where we note that here $\mathbf{Y}'$ indicates each of the $mT$ training samples excluding their last time point.

Observe that the following chain of inequalities holds:

$$
\begin{aligned}
\Phi_{\text{hyb}}(\mathbf{Y}) &= \sup_{h\in\mathcal{H}} \mathcal{L}_{\text{hyb}}(h \mid \mathbf{Y}) - \widehat{\mathcal{L}_{\text{hyb}}}(h) \\
&\le \sup_{h\in\mathcal{H}}\left[\mathcal{L}_{\text{hyb}}(h \mid \mathbf{Y}) - \mathcal{L}_{\text{hyb}}(h \mid \mathbf{Y}')\right] \\
&\quad + \sup_{h\in\mathcal{H}}\left[\mathcal{L}_{\text{hyb}}(h \mid \mathbf{Y}') - \widehat{\mathcal{L}_{\text{hyb}}}(h, \mathbf{Y})\right] \\
&\le \frac{1}{T}\sum_{t=1}^T \sup_{h\in\mathcal{H}}\Big[\mathcal{L}_{\text{hyb}}(h \mid \mathbf{Y}) \\
&\quad - \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\mathcal{D}'}[L(h(Y_{t-p}^{t-1}(i)), Y_t(i)) \mid \mathbf{Y}']\Big] \\
&\quad + \sup_{h\in\mathcal{H}}\left[\mathcal{L}_{\text{hyb}}(h \mid \mathbf{Y}') - \widehat{\mathcal{L}_{\text{hyb}}}(h, \mathbf{Y})\right].
\end{aligned}
$$

and so

$$
\Phi_{\text{hyb}}(\mathbf{Y}) - \frac{1}{T}\sum_t \Delta_t \le \underbrace{\sup_{h\in\mathcal{H}} \mathcal{L}_{\text{hyb}}(h, \mid \mathbf{Y}') - \widehat{\mathcal{L}_{\text{hyb}}}(h, \mathbf{Y})}_{\phi(\mathbf{Y})}.
$$

Then, following the exact same reasoning as above for $\Phi_{\text{s2s}}$ shows that for $\delta > 0$, we have with probability $1 - \delta/2$

$$
\begin{aligned}
&\Phi_{\text{hyb}}(\mathbf{Y}) \\
&\le \underbrace{\max_j \widehat{\mathfrak{R}}_{\widetilde{C}_j}(\mathcal{F}) + \frac{1}{T}\sum_t \Delta_t + \sqrt{\frac{\log\left(\frac{2k}{\delta - \sum_j(|I_j|-1)\beta(I_j)}\right)}{2\min_j |I_j|}}}_{B_1}
\end{aligned}
$$

However, upper bounding $\Phi_{\text{hyb}}$ can also be approached using the same techniques as Kuznetsov and Mohri (2015), which we now describe. Let $\alpha > 0$. For a given $h$, computing $\mathcal{L}_{\text{hyb}}(h, \mathbf{Y})$ is similar in expectation to running $h$ on each of the $m$ time series, yielding for each time series $Y_{T-p+1}^T(i)$ the bound

$$
\begin{aligned}
&\mathbb{E}[L(h(Y_{T-p+1}^T(i)), Y_{T+1}(i)) \mid \mathbf{Y}] \\
&\le \frac{1}{T}\sum_{t=1}^T L(h(Y_{t-p}^{t-1}(i)), Y_t(i)) + \Delta(\mathbf{Y}_i) \\
&\quad + 2\alpha + \sqrt{\frac{2}{T}\log\frac{\max_i \mathbb{E}_{v\sim T(\mathbf{Y}_i)}[\mathcal{N}_i(\alpha, \mathcal{F}, v)]}{\delta}}
\end{aligned}
$$

and so by union bound, as above, we obtain with probability $1 - \delta/2$

$$
\begin{aligned}
\Phi_{\text{hyb}}(\mathbf{Y}) &\le \frac{1}{m}\sum \Delta(\mathbf{Y}_i) + 2\alpha \\
&\quad + \sqrt{\frac{2}{T}\log\frac{2m\max_i \mathbb{E}_{v\sim T(\mathbf{Y}_i)}[\mathcal{N}_i(\alpha, \mathcal{F}, v)]}{\delta}} \\
&\le B_2
\end{aligned}
$$

We conclude by a final union bound on the event $\{\Phi_{\mathrm{hyb}}(\mathbf{Y}) \geq B_1 \cup \Phi_{\mathrm{hyb}}(\mathbf{Y}) \geq B_2\}$, we obtain with probability $1 - \delta$,

$$\Phi_{\mathrm{hyb}}(\mathbf{Y}) \leq \min(B_1, B_2)$$

□