Learning Theory for Conditional Risk Minimization: Supplementary Material

Alexander Zimin
IST Austria
azimin@ist.ac.at

Christoph H. Lampter
IST Austria
chl@ist.ac.at

1 Proofs

Proof of Theorem 1. After the application of (6) and (8) we can consider the two parts separately:

$$\mathbb{P}\left[R_n(h_n) - \inf_{h\in\mathcal{H}} R_n(h) > \alpha\right] \tag{31}$$

$$\leq \mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{t=1}^{n}(\ell(h,\mathbf{z}_t) - R_{t-1}(h))\right| > \alpha/4\right] \tag{32}$$

$$+\,\mathbb{P}\left[\frac{1}{n}\sum_{t=1}^{n} d_{t-1,n} > \alpha/4\right]. \tag{33}$$

The convergence of the probability in (32) is guaranteed by the result of [Rakhlin et al., 2014] for any stochastic process. The convergence of (33) follows from the definition of the convergent discrepancies and is a content of Lemma 2. □

Lemma 2. *If double array $d_{t,n}$ is convergent, then $\frac{1}{n}\sum_{t=1}^{n} d_{t-1,n}$ converges to 0 in probability.*

Proof. The proof is similar to that of the Toeplitz lemma, but adapted to our notion of convergence. Fix $\varepsilon > 0$ and $\delta > 0$. Then, by the definition of a convergent array, for $\varepsilon' = \delta' = \frac{\delta\varepsilon}{4}$

$$\exists n_0, \exists t_0 : 0 \leq t_0 < n_0, \forall n \geq n_0, \forall t_0 \leq t < n : \tag{34}$$

$$\mathbb{P}\left[d_{t,n} > \varepsilon'\right] \leq \delta'. \tag{35}$$

In particular, this means that for any $n \geq n_0$ and $\forall t_0 \leq t < n$ we have $\mathbb{E}\left[d_{t,n}\right] \leq \varepsilon' + \delta' = \frac{\delta\varepsilon}{2}$, because of the boundedness of $d_{t,n}$.

Now, choose any $n_1 \geq n_0$ that satisfies $\frac{n_0}{n_1} \leq \frac{\varepsilon}{2}$. Then for any $n \geq n_1$ we get

$$\mathbb{P}\left[\frac{1}{n}\sum_{t=1}^{n} d_{t-1,n} > \varepsilon\right] \leq \mathbb{P}\left[\frac{1}{n}\sum_{t=n_0+1}^{n} d_{t-1,n} > \frac{\varepsilon}{2}\right] \tag{36}$$

$$\leq 2\frac{\sum_{t=n_0+1}^{n} \mathbb{E}\left[d_{t-1,n}\right]}{n\varepsilon} \tag{37}$$

$$\leq \delta, \tag{38}$$

where the last line follows from the bound on the expectations. □

To characterize a complexity of some function class we use covering numbers and a sequential fat-shattering dimension. But before we could give those definitions, we need to introduce a notion of $\mathcal{Z}$-valued trees.

A $\mathcal{Z}$-valued tree of depth n is a sequence $z_{1:n}$ of mappings $z_i : \{\pm 1\}^{i-1} \to \mathcal{Z}$. A sequence $\varepsilon_{1:n} \in \{\pm 1\}^n$ defines a path in a tree. To shorten the notations, $\mathbf{z}_t(\varepsilon_{1:t-1})$ is denoted as $\mathbf{z}_t(\varepsilon)$. For a double sequence $z_{1:n}, z'_{1:n}$, we define $\chi_t(\varepsilon)$ as z_t if $\varepsilon = 1$ and z'_t if $\varepsilon = -1$. Also define distributions $p_t(\varepsilon_{1:t-1}, z_{1:t-1}, z'_{1:t-1})$ over $\mathcal{Z}$ as $\mathbb{P}\left[\cdot \,|\, \chi_1(\varepsilon_1), \dots, \chi_{t-1}(\varepsilon_{t-1})\right]$, where $\mathbb{P}$ is a distribution of a process under consideration. Then we can define a distribution ρ over two $\mathcal{Z}$-valued trees $\mathbf{z}$ and $\mathbf{z}'$ as follows: $\mathbf{z}_1$ and $\mathbf{z}'_1$ are sampled independently from the initial distribution of the process and for any path $\varepsilon_{1:n}$ for $2 \leq t \leq n$, $\mathbf{z}_t(\varepsilon)$ and $\mathbf{z}'_t(\varepsilon)$ are sampled independently from $p_t(\varepsilon_{1:t-1}, \mathbf{z}_{1:t-1}(\varepsilon), \mathbf{z}'_{1:t-1}(\varepsilon))$.

For any random variable $\mathbf{y}$ that is measurable with respect to σ_n (a σ-algebra generated by $\mathbf{z}_{1:n}$), we define its symmetrized counterpart $\tilde{\mathbf{y}}$ as follows. We know that there exists a measurable function ψ such that $\mathbf{y} = \psi(\mathbf{z}_{1:n})$. Then we define $\tilde{\mathbf{y}} = \psi(\chi_1(\epsilon_1), \dots, \chi_n(\varepsilon_n))$, where the samples used by χ_t's are understood from the context.

Now we can define covering numbers.

Definition 5. A set, V, of $\mathbb{R}$-valued trees of depth n is a (sequential) θ**-cover** (with respect to the ℓ_∞-norm) of $\mathcal{F} \subset \{f : \mathcal{Z} \to \mathbb{R}\}$ on a tree $\mathbf{z}$ of depth n if

$$\forall f \in \mathcal{F}, \forall \varepsilon \in \{\pm 1\}^n, \exists v \in V : \tag{39}$$

$$\max_{1\leq t\leq n} |f(\mathbf{z}_t(\varepsilon)) - v_t(\varepsilon)| \leq \theta. \tag{40}$$

The (sequential) θ**-covering number** of a function class $\mathcal{F}$ on a given tree $\mathbf{z}$ is

$$\mathcal{N}_\infty(\mathcal{F}, \theta, \mathbf{z}) = \min\{|V| : V \text{ is an } \theta\text{-cover} \tag{41}$$

$$\text{w.r.t. } \ell_\infty\text{-norm of } \mathcal{F} \text{ on } \mathbf{z}\}. \tag{42}$$

The **maximal θ-covering number** of a function class $\mathcal{F}$ over depth-n trees is

$$\mathcal{N}_\infty(\mathcal{F}, \theta, n) = \sup_{\mathbf{z}} \mathcal{N}_\infty(\mathcal{F}, \theta, \mathbf{z}). \tag{43}$$

To control the growth of covering numbers we use the following notion of complexity.

Definition 6. A $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n is θ-shattered by a function class $\mathcal{F} \subseteq \{f : \mathcal{Z} \to \mathbb{R}\}$ if there exists an $\mathbb{R}$-valued tree s of depth n such that

$$\forall \varepsilon \in \{\pm 1\}^n, \exists f \in \mathcal{F} \text{ s.t. } 1 \leq t \leq n, \tag{44}$$

$$\varepsilon_t(f(\mathbf{z}_t(\varepsilon)) - s_t(\varepsilon)) \geq \theta/2. \tag{45}$$

The *(sequential) fat-shattering dimension* $\text{fat}_\theta(\mathcal{F})$ at scale θ is the largest d such that $\mathcal{F}$ θ-shatters a $\mathcal{Z}$-valued tree of depth d.

An important result of [Rakhlin et al., 2014] is the following connection between the covering numbers and the fat-shattering dimension.

Lemma 3 (Corollary 1 of [Rakhlin et al., 2014])**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{Z} \to [-1, 1]\}$. For any $\theta > 0$ and any $n \geq 1$, we have that*

$$\mathcal{N}_\infty(\mathcal{F}, \theta, n) \leq \left(\frac{2en}{\theta}\right)^{fat_\theta(\mathcal{F})}. \tag{46}$$

In the proofs we denote $\mathcal{L}(\mathcal{H})$ as $\mathcal{F}$.

Proof of Theorem 2. After equations (6), (8) and (10), we are left to study the large deviations of the following quantity

$$\Theta(J_n) = \sup_{f \in \mathcal{F}} \left| \sum_{t=1}^{n} w_t(J_n) \left(f(\mathbf{z}_t) - \mathbb{E}_{t-1}[f]\right) \right| \tag{47}$$

with the weights defined as in (11). Let us define events $A_r = \{J_n = r\}$ and $B_r(j) = \{r \leq \sum_{t=1}^n g(M_{t-1,j}) \leq r + 1\}$, such that $E_{k,m} = \{\cup_{r \leq k} A_r\} \cap \{\cup_{r \geq m} B_r(J_n)\}$. Then we have

$$\mathbb{P}[\Theta(J_n) \geq \alpha] \leq \mathbb{P}[\Theta(J_n) \geq \alpha \wedge E_{k,m}] + \mathbb{P}\left[E_{k,m}^c\right]. \tag{48}$$

Now we can take a union bound for the first summand over A_r's and get

$$\mathbb{P}[\Theta(J_n) \geq \alpha \wedge E_{k,m}] \tag{49}$$

$$\leq \sum_{j=1}^{k} \mathbb{P}[\Theta(j) \geq \alpha \wedge \{\cup_{r \geq m} B_r(j)\}]. \tag{50}$$

Taking another union bound for each j, we end up with

$$\mathbb{P}[\Theta(j) \geq \alpha \wedge \{\cup_{r \geq m} B_r(j)\}] \tag{51}$$

$$\leq \sum_{r \geq m} \mathbb{P}[\Theta(j) \geq \alpha \wedge B_r(j)]. \tag{52}$$

Now we study the last probability for a fixed r and j. On $B_r(j)$ we can lower bound the denominator of the weights $\sum_{t=1}^n g(M_{t-1,j}) \geq r$ leading to $\Theta(j) \leq \Theta_r(j) = \frac{1}{r} \sup_{f \in \mathcal{F}} |\sum_{t=1}^n g(M_{t-1,j}) (f(\mathbf{z}_t) - \mathbb{E}_{t-1}[f])|$. Let $\lambda > 0$ and denote $V = \frac{1}{r^2} \sum_{t=1}^n g^2(M_{t-1,j})$, $E = \frac{1}{r} \sum_{t=1}^n g(M_{t-1,j})$. Then, since $\frac{1}{r} g(M_{t-1,j}) \sim \sigma_{t-1}$ by the definition of an M-bound, Lemma 4 gives us

$$\mathbb{E}\left[e^{\lambda \Theta_r(j) - \lambda^2 V - 2\lambda\beta E - \ln 2\mathcal{N}_\infty(\mathcal{F}, \beta, n)}\right] \leq 1. \tag{53}$$

Let $C = \{\Theta_r(j) \geq \alpha \wedge B_r(j)\}$ and note that $E \leq \frac{r+1}{r} \leq 2$ and $V \leq \frac{r+1}{r^2} \leq \frac{2}{r}$ on $B_r(j)$ by the boundedness of g. Then we have the following chain of inequalities

$$1 \geq \mathbb{E}\left[e^{\lambda \Theta_r(j) - \lambda^2 V - 2\lambda\beta E - \ln 2\mathcal{N}_\infty(\mathcal{F}, \beta, n)}\right] \tag{54}$$

$$\geq \mathbb{E}\left[e^{\lambda \Theta_r(j) - \lambda^2 V - 2\lambda\beta E - \ln 2\mathcal{N}_\infty(\mathcal{F}, \beta, n)} \mathbb{I}[C]\right] \tag{55}$$

$$\geq e^{\lambda\alpha - \lambda^2 \frac{2}{r} - 4\lambda\beta - \ln 2\mathcal{N}_\infty(\mathcal{F}, \beta, n)} \mathbb{P}[C]. \tag{56}$$

Hence, by optimizing over λ, we get

$$\mathbb{P}[\Theta(j) \geq \alpha \wedge B_r(j)] \leq 2\mathcal{N}_\infty(\mathcal{F}, \beta, n) e^{-\frac{1}{2} r(\alpha - 4\beta)^2}. \tag{57}$$

Now, coming back to (51), we can evaluate it by computing the sum to obtain

$$\mathbb{P}[\Theta(J) \geq \alpha \wedge E_{k,m}] \leq \frac{2k\mathcal{N}_\infty(\mathcal{F}, \beta, n)}{(\alpha - 4\beta)^2} e^{-\frac{1}{2} m(\alpha - 4\beta)^2}. \tag{58}$$

□

Lemma 4. *Let $\mathbf{y}_{1:n}$ be a process such that each $\mathbf{y}_t \sim \sigma_{t-1}$ and denote $E = \sum_{t=1}^n |\mathbf{y}_t|$, $V = \sum_{t=1}^n \mathbf{y}_t^2$. Then for a fixed $\lambda, \beta > 0$ and $c = \ln 2\mathcal{N}_\infty(\mathcal{F}, \beta, n)$*

$$\mathbb{E}\left[e^{\lambda \sup_{f \in \mathcal{F}} \left|\sum_{t=1}^n \mathbf{y}_t (f(\mathbf{z}_t) - \mathbb{E}_{t-1}[f])\right| - \lambda^2 V - 2\lambda\beta E - c}\right] \leq 1 \tag{59}$$

Proof. Let $\mathbf{z}'_{1:n}$ be a decoupled tangent sequence to $\mathbf{z}_{1:n}$, i.e. a sequence that satisfies $\mathbb{E}_{t-1}[f(\mathbf{z}_t)] = \mathbb{E}_{t-1}[f(\mathbf{z}'_t)] = \mathbb{E}[f(\mathbf{z}'_t) | \mathbf{z}_{1:n}]$. Then

$$\mathbb{E}\left[e^{\lambda \sup_{f \in \mathcal{F}} \left|\sum_{t=1}^n \mathbf{y}_t (f(\mathbf{z}_t) - \mathbb{E}_{i-1}[f])\right| - \lambda^2 V - 2\lambda\beta E - c}\right] \tag{60}$$

$$\leq \mathbb{E}\left[e^{\lambda \sup_{f \in \mathcal{F}} \left|\sum_{t=1}^n \mathbf{y}_t \left(f(\mathbf{z}_t) - f(\mathbf{z}'_t)\right)\right| - \lambda^2 V - 2\lambda\beta E - c}\right]. \tag{61}$$

The Lemma 5 gives us that (61) equals to

$$\mathbb{E}_\rho \mathbb{E}_\varepsilon \left[e^{\lambda \sup_f \left|\sum_{t=1}^n \tilde{\mathbf{y}}_t \varepsilon_t \left(f(\mathbf{z}_t(\varepsilon)) - f(\mathbf{z}'_t(\varepsilon))\right)\right| - \lambda^2 \tilde{V} - 2\lambda\beta \tilde{E} - c}\right] \tag{62}$$

$$\leq \mathbb{E}_{\mathbf{z} \sim \rho} \mathbb{E}_\varepsilon \left[e^{2\lambda \sup_f \left|\sum_{t=1}^n \tilde{\mathbf{y}}_t \varepsilon_t f(\mathbf{z}_t(\varepsilon))\right| - \lambda^2 \tilde{V} - 2\lambda\beta \tilde{E} - c}\right], \tag{63}$$

where $\tilde{\mathbf{y}}$ is a symmetrized version of $\mathbf{y}$, $\tilde{E} = \sum_{t=1}^n |\tilde{\mathbf{y}}_t|$, $\tilde{V} = \sum_{t=1}^n \tilde{\mathbf{y}}_t^2$ and we used Jensen inequality to get the second line. Now we take a β-cover of $\mathcal{F}$ with respect to ℓ_∞-norm to get the following bound on (63)

$$\mathbb{E}_{\mathbf{z}\sim\rho}\mathcal{N}_\infty(\mathcal{F},\beta,n)\mathbb{E}_\varepsilon\left[e^{2\lambda\left|\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t(\varepsilon))\right|-\lambda^2\tilde{V}-c}\right] \tag{64}$$

$$= \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim\rho}\mathbb{E}_\varepsilon\left[e^{2\lambda\left|\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t(\varepsilon))\right|-\lambda^2\tilde{V}}\right] \tag{65}$$

Introduce events $Y_+ = \{\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t) \geq 0\}$ and $Y_- = \{\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t) < 0\}$. Then the last line is equal to

$$\frac{1}{2}\mathbb{E}_{\mathbf{z}\sim\rho}\mathbb{E}_\varepsilon\left[e^{2\lambda\left|\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t(\varepsilon))\right|-\lambda^2\tilde{V}}\mathbb{I}\left[Y_+\right]\right] \tag{66}$$

$$+\frac{1}{2}\mathbb{E}_{\mathbf{z}\sim\rho}\mathbb{E}_\varepsilon\left[e^{2\lambda\left|\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t(\varepsilon))\right|-\lambda^2\tilde{V}}\mathbb{I}\left[Y_-\right]\right] \tag{67}$$

$$\leq \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim\rho}\mathbb{E}_\varepsilon\left[e^{2\lambda\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t(\varepsilon))-\lambda^2\tilde{V}}\right] \tag{68}$$

$$+\frac{1}{2}\mathbb{E}_{\mathbf{z}\sim\rho}\mathbb{E}_\varepsilon\left[e^{-2\lambda\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t(\varepsilon))-\lambda^2\tilde{V}}\right] \tag{69}$$

$$\leq 1, \tag{70}$$

where the last line follows by the standard martingale argument, since $\tilde{\mathbf{y}}_t\varepsilon_t f(\mathbf{z}_t(\varepsilon))$ is a martingale difference sequence (for a fixed tree $\mathbf{z}$). □

Lemma 5. *Let $\mathbf{z}_{1:n}$ be a sample from a process and $\mathbf{z}'_{1:n}$ its decoupled sequence. Let $\mathbf{y}_{1:n}$ be a process such that each $\mathbf{y}_t \sim \sigma_{t-1}$, then for any measurable functions $\varphi : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathcal{Z}^n \to \mathbb{R}$, we have*

$$\mathbb{E}\left[\varphi\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^n \mathbf{y}_t\left(f(\mathbf{z}_t) - f(\mathbf{z}'_t)\right)\right|\right)\psi(\mathbf{z}_{1:n})\right] \tag{71}$$

$$= \mathbb{E}_\rho\mathbb{E}_\varepsilon\left[\varphi\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^n \tilde{\mathbf{y}}_t\varepsilon_t\left(f(\mathbf{z}_t) - f(\mathbf{z}'_t)\right)\right|\right)\tilde{\psi}\right],$$

where $\tilde{\psi}$ is a symmetrized version of $\psi(\mathbf{z}_{1:n})$.

Proof. The proof is direct extension of Theorem 3 from Rakhlin et al. [2011] by using the fact $\mathbf{y}_t \sim \sigma_{t-1}$. □

Proof of Corollary 1. The proof follows from the Theorem 2 if we set $\beta = \frac{\alpha}{8}$ and use the Lemma 3. □

Proof of Lemma 1. The proof follows from the following bound

$$d_{t,n} = \sup_{f\in\mathcal{L}(\mathcal{H})}\left|\mathbb{E}_t f - \mathbb{E}_n\left[\mathbf{x}_f\right]\right| \tag{72}$$

$$\leq \mathbb{E}_n\left[\sup_{f\in\mathcal{L}(\mathcal{H})}\left|\mathbb{E}_t f - \mathbf{x}_f\right|\right]. \tag{73}$$

And then the convergence of the discrepancies follows from the definition of the uniformly convergent martingale. □

2 Exceptional set examples

Markov chains. First, we bound the probability of A_k:

$$\mathbb{P}\left[J_n > k\right] \leq \mathbb{P}\left[F_{\mathbf{z}_n} > k\right] \leq |S|\max_s \mathbb{P}\left[F_s > k\right]. \tag{74}$$

On the event $B_{k,m}$ we have the following chain of inequalities.

$$\sum_{t=J_n}^n \mathbb{I}\left[d_{t,J_n} \leq b_n\right] \geq \sum_{t=k}^n \mathbb{I}\left[d_{t,J_n} \leq b_n\right] \tag{75}$$

$$\geq \sum_{t=k}^n \mathbb{I}\left[d_{t,J_n} = 0\right] \tag{76}$$

$$\geq \sum_{t=k}^n \mathbb{I}\left[\mathbf{z}_t = \mathbf{z}_{J_n}\right], \tag{77}$$

which gives us

$$\mathbb{P}\left[J_n \leq k \wedge \sum_{t=J_n}^n \mathbb{I}\left[d_{t,J_n} \leq b_n\right] < m\right] \tag{78}$$

$$\leq \mathbb{P}\left[J_n \leq k \wedge \sum_{t=k}^n \mathbb{I}\left[\mathbf{z}_t = \mathbf{z}_{J_n}\right] < m\right] \tag{79}$$

$$\leq |S|\max_s \mathbb{P}\left[J_n \leq k \wedge \sum_{t=k}^n \mathbb{I}\left[\mathbf{z}_t = s\right] < m \wedge \mathbf{z}_{J_n} = s\right]. \tag{80}$$

Now, for a given state s, $\sum_{t=k}^n \mathbb{I}\left[\mathbf{z}_t = s\right]$ can be lower bounded by the number of times we hit the state s again. Let $T_s^i, i \geq 1$, be independent copies of the recurrence times. Then $\sum_{t=k}^n \mathbb{I}\left[\mathbf{z}_t = s\right] \geq m$ for any $m \geq 0$, such that $\sum_{i=1}^m T_s^i \leq n - k$. We also have the following sequence of inclusions.

$$\left\{1 \leq i \leq m : T_s^i \leq \lfloor\frac{n-k}{m}\rfloor \wedge J_n \leq k \wedge \mathbf{z}_{J_n} = s\right\} \tag{81}$$

$$\subseteq \left\{\sum_{i=1}^m T_s^i \leq n - k \wedge J_n \leq k \wedge \mathbf{z}_{J_n} = s\right\} \tag{82}$$

$$\subseteq \left\{\sum_{t=k}^n \mathbb{I}\left[\mathbf{z}_t = s\right] \geq m \wedge J_n \leq k \wedge \mathbf{z}_{J_n} = s\right\}. \tag{83}$$

And this gives us

$$\mathbb{P}\left[J_n \leq k \wedge \sum_{t=k}^n \mathbb{I}\left[\mathbf{z}_t = s\right] < m \wedge \mathbf{z}_{J_n} = s\right] \tag{84}$$

$$\leq \mathbb{P}\left[\exists\, 1 \leq i \leq m : T_s^i > \lfloor\frac{n-k}{m}\rfloor\right] \tag{85}$$

$$\leq m\mathbb{P}\left[T_s > \lfloor\frac{n-k}{m}\rfloor\right]. \tag{86}$$

Dynamical systems. The bound on $\mathbb{P}[A_k]$ follows from the fact that $J_n \leq F(C_n)$. For the $B_{k,m}$ we get

$$\mathbb{P}[B_{k,m}] \leq k \max_{1\leq j\leq k} \mathbb{P}\left[J_n = j \wedge \sum_{t=j}^{n} \mathbb{I}[d_{t,j} \leq b_n]\right]. \tag{87}$$

And similarly to the Markov chain example,

$$\mathbb{P}\left[J_n = j \wedge \sum_{t=j}^{n} \mathbb{I}[d_{t,j} \leq b_n]\right] \leq \mathbb{P}\left[T(C_j) > \lfloor\frac{n-j}{m}\rfloor\right]. \tag{88}$$

General stationary processes. The bound for this case is done analogously to the previous two examples, thus we omit the argument.

3 Counter-example for learnability

Theorem 3. *Let $\mathcal{Z} = \{0, 1\}$, $\mathcal{H} = [0, 1]$ and $\ell(h, z) = (h-z)^2$. Also, let $\mathcal{C}$ be a class of all stationary ergodic processes taking values in $\mathcal{Z}$. Then for any learning algorithm that produces a sequence of hypotheses h_n, there is a process $P \in \mathcal{C}$ such that*

$$\mathbb{P}\left[\limsup_{n\to\infty}\left(R_n(h_n) - \inf_{h\in\mathcal{H}} R_n(h)\right) > \frac{1}{16}\right] \geq \frac{1}{8}. \tag{89}$$

Proof. Using the fact that the minimizer of $\mathbb{E}_n\left[(h - \mathbf{z}_{n+1})^2\right]$ is $\mathbb{E}_n\mathbf{z}_{n+1}$, we can rewrite for any $h_n \sim \sigma_n$

$$R_n(h_n) - \inf_{h\in\mathcal{H}} R_n(h) \tag{90}$$
$$= \mathbb{E}_n\left[(h_n - \mathbf{z}_{n+1})^2\right] - \inf_{h\in\mathcal{H}} \mathbb{E}_n\left[(h - \mathbf{z}_{n+1})^2\right] \tag{91}$$
$$= \mathbb{E}_n\left[(h_n - \mathbf{z}_{n+1})^2\right] - \mathbb{E}_n\left[(\mathbb{E}_n\mathbf{z}_{n+1} - \mathbf{z}_{n+1})^2\right] \tag{92}$$
$$= (h_n - \mathbb{E}_n\mathbf{z}_{n+1})^2. \tag{93}$$

A minor modification of the proof of Theorem 1 of [Györfi et al., 1998] gives that for every algorithm that produces a sequence h_n of hypotheses, there is a stationary and ergodic process such that

$$\mathbb{P}\left[\limsup_{n\to\infty} (h_n - \mathbb{E}_n\mathbf{z}_{n+1})^2 > \frac{1}{16}\right] \geq \frac{1}{8}, \tag{94}$$

which shows that no algorithm can be a limit learner for the class of all stationary and ergodic binary processes. □

4 Connection to time series prediction

The goal of this section is to show the connection of our framework to existing theoretical approaches to time series prediction. In particular, we consider two frameworks, which are close enough to conditional risk minimization. In both cases, we show that the conditional risk minimization solves harder problem in a sense that its solutions can be used to solve these particular problems, but it requires more assumptions to be valid.

We start with a framework of time series prediction by statistical learning, considered for example in [Alquier et al., 2013, McDonald et al., 2012]. Fixing some point n in time, we consider a hypotheses class $\tilde{\mathcal{H}} \subseteq \{h : \mathcal{Z}^n \to \mathcal{Z}\}$, where each hypotheses h gives us a prediction of the next step by evaluating the whole history. For any loss function $\ell : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$, we consider the following risk minimization problem:

$$\min_{h\in\tilde{\mathcal{H}}} \mathbb{E}\left[\ell(h(\mathbf{z}_{1:n}), \mathbf{z}_{n+1})\right]. \tag{95}$$

To set up the conditional risk minimization, we define a class of constant functions $\mathcal{H} = \{h_{z'}(z) = z', \forall z' \in \mathcal{Z}\}$. Then if the process belongs to a class learnable with $\mathcal{H}$ and ℓ, we can guarantee that there is an algorithm to choose a point $\mathbf{z}'_n$, such that with probability $1-\delta$

$$\mathbb{E}\left[\ell(\mathbf{z}'_n, \mathbf{z}_{n+1}) | \, \mathbf{z}_{1:n}\right] \leq \inf_{z'} \mathbb{E}\left[\ell(z', \mathbf{z}_{n+1}) | \, \mathbf{z}_{1:n}\right] + \varepsilon_n(\delta), \tag{96}$$

where $\varepsilon_n(\delta)$ is a sequence of errors guaranteed by the algorithm for a given confidence δ and $\varepsilon_n(\delta) \to 0$. Converting this to the bound on the expectation, we get

$$\mathbb{E}\left[\ell(\mathbf{z}'_n, \mathbf{z}_{n+1})\right] \leq \mathbb{E}\left[\inf_{z'} \mathbb{E}\left[\ell(z', \mathbf{z}_{n+1}) | \, \mathbf{z}_{1:n}\right]\right] \tag{97}$$
$$+ \varepsilon_n(\delta) + \delta. \tag{98}$$

Notice that

$$\mathbb{E}\left[\inf_{z'} \mathbb{E}\left[\ell(z', \mathbf{z}_{n+1}) | \, \mathbf{z}_{1:n}\right]\right] \tag{99}$$
$$\leq \mathbb{E}\left[\inf_{h\in\tilde{\mathcal{H}}} \mathbb{E}\left[\ell(h(\mathbf{z}_{1:n}), \mathbf{z}_{n+1}) | \, \mathbf{z}_{1:n}\right]\right] \tag{100}$$
$$\leq \inf_{h\in\tilde{\mathcal{H}}} \mathbb{E}\left[\ell(h(\mathbf{z}_{1:n}), \mathbf{z}_{n+1})\right]. \tag{101}$$

Therefore, if the process is from a learnable class, there is an algorithm that always give good predictions according to this framework as well.

The second setting, which was considered by [Wintenberger, 2014], is very close to the online sequence prediction. In order to reduce the notations and simplify the presentation, we assume that the learner has an access to a (usually finite) hypothesis class $\mathcal{H}$ and at every step t he should choose a distribution π_t over $\mathcal{H}$ in a way that minimizes the regret:

$$\sum_{t=1}^{n} \mathbb{E}_{t-1}\left[\ell(\mathbb{E}_{\pi_t} h, \mathbf{z}_t)\right] - \min_{h\in\mathcal{H}} \sum_{t=1}^{n} \mathbb{E}_{t-1}\left[\ell(h, \mathbf{z}_t)\right]. \tag{102}$$

Again, if the process belongs to a learnable class with $\mathcal{H}$ and ℓ, then there is an algorithm, which produce the sequence h_t that satisfies with probability $1-\delta$

$$\mathbb{E}_{t-1}\left[\ell(h_t, \mathbf{z}_t)\right] \leq \min_{h\in\mathcal{H}} \mathbb{E}_{t-1}\left[\ell(h, \mathbf{z}_t)\right] + \varepsilon_t(\delta/n) \quad (103)$$

for all $1 \leq t \leq n$. Summing up over t, we get

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{E}_{t-1}\left[\ell(h_t, \mathbf{z}_t)\right] && (104)\\
&\leq \sum_{t=1}^{n} \min_{h\in\mathcal{H}} \mathbb{E}_{t-1}\left[\ell(h, \mathbf{z}_t)\right] + \sum_{t=1}^{n} \varepsilon_t(\delta/n) && (105)\\
&\leq \min_{h\in\mathcal{H}} \sum_{t=1}^{n} \mathbb{E}_{t-1}\left[\ell(h, \mathbf{z}_t)\right] + \sum_{t=1}^{n} \varepsilon_t(\delta/n). && (106)
\end{aligned}$$

Thus giving us $\sum_{t=1}^{n} \varepsilon_t(\delta/n)$ bound on the regret with high probability. For nice sequences (like i.i.d.) $\varepsilon_t(\delta/n)$ is of order $\mathcal{O}\left(\sqrt{\frac{\log n}{t}}\right)$, which gives a regret bound of order $\mathcal{O}\left(\sqrt{n \log n}\right)$. On the downside, we can get guarantees only for a class of learnable processes, while the results of [Wintenberger, 2014] hold for any stochastic process. The reason for this is that conditional risk minimization is inherently more difficult problem, since it requires to optimize at every step and not in the cumulative sense.