One-step ahead sequential Super Learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters

Geoffrey Ecoto[1,2], Aurélien F. Bibaut[3], Antoine Chambaz[2]

[1] Caisse Centrale de Réassurance

[2] MAP5 (UMR CNRS 8145), Université de Paris

[3] Netflix

July 28, 2021

Abstract

Suppose that we observe a short time series where each time-t-specific data-structure consists of many slightly dependent data indexed by a and that we want to estimate a feature of the law of the experiment that depends neither on t nor on a. We develop and study an algorithm to learn sequentially which base algorithm in a user-supplied collection best carries out the estimation task in terms of excess risk and oracular inequalities. The analysis, which uses dependency graph to model the amount of conditional independence within each t-specific data-structure and a concentration inequality by Janson [2004], leverages a large ratio of the number of distinct a-s to the degree of the dependency graph in the face of a small number of t-specific data-structures. The so-called one-step ahead Super Learner is applied to the motivating example where the challenge is to anticipate the cost of natural disasters in France.

1 Introduction

Caisse Centrale de Réassurance and the cost of natural disasters in France. In France, Law n°82-600 of July 13th 1982 imposes a compulsory extension of the guarantee for all property

insurance contracts for the coverage of natural catastrophes. This law defines the legal framework of the natural disasters compensation scheme, of which Caisse Centrale de Réassurance (CCR) is a major actor in France. With the French State guarantee, CCR provides its cedents[1] operating in France (*i.e.*, the insurance companies operating in France that CCR reinsures) with unlimited coverage against natural catastrophes. In order to better anticipate the risks, CCR has developed an expertise in natural disasters modeling. The so-called "cat models" [Mitchell-Wallace et al., 2017] exploit portfolios and claims data collected from CCR's cedents to enable a better appreciation of the exposures[2] of CCR, of its cedents and of the French State. Our study proposes a new method to better predict the aforementioned exposures. Termed "one-step ahead sequential Super Learning", rooted in statistical theory, the method allows to learn from short time series of many slightly dependent data.

Statistical challenges. Developing such a method presents several technical challenges. From a theoretical point of view, we have to deal with a time series $(\bar{O}_t)_{t\geq 1}$ whose time-t-specific component $\bar{O}_t$ consists of a large collection $(O_{\alpha,t})_{\alpha\in\mathcal{A}}$ of data that are dependent but such that there is a large amount of independence among them. The time series is observed only at a limited number of time steps, a drawback that could be mitigated by the large cardinality of $\mathcal{A}$. Furthermore, for reasons that we will present later on, we favor the development of a learning algorithm that works in an online fashion. The learning algorithm should build upon a library of competing algorithms, either to select the one that performs best or to combine the algorithms into a single meta-algorithm that performs almost as well as all possible combinations thereof (this is known as stacking, or aggregating, or Super Learning in the literature). Of course, assessing the said performances is not easy, notably because it requires some form of online cross-validation procedure. From the applied point of view, assembling the learning data set is difficult because the data come from many sources and take on various shapes. Moreover, some of the data are only partially available. Details will be given later on.

[1] A cedent is a party in an insurance contract that passes the financial obligation for certain potential losses to the insurer. In return for bearing a particular risk of loss, the cedent pays an insurance premium.

[2] The state of being subject to loss because of some hazard or contingency.

Organization of the article. Section 2 presents the theoretical development and analysis of the one-step ahead sequential Super Learner. Readers who are more interested in the application than in the theory could jump to Section 2.2 for a summary. Section 3 presents the complete application. The main objective is exposed in finer detail; the actual implementation of the algorithm is described; the obtained results are reported and commented upon. Section 4 closes the article on a discussion. Further details are given in the appendix.

2 A new result for the one-step ahead sequential Super Learner

Let $(\bar{O}_t)_{t\geq 1}$ be a time-t-ordered sequence of observations where each $\bar{O}_t$ is in fact a finite collection $(O_{\alpha,t})_{\alpha\in\mathcal{A}}$ of (α,t)-specific elements of a measured space $\mathcal{O}$. We are especially interested in situations where the variables $(O_{\alpha,t})_{\alpha\in\mathcal{A}}$ are conditionally dependent given the σ-field $F_{t-1} := \sigma(O_{\alpha,\tau} : \alpha\in\mathcal{A}, 1\leq\tau<t)$ generated by past observations (by convention, $F_0 := \emptyset$), but there is a large amount of conditional independence between them.

We rely on conditional dependency graphs to model the amount of conditional independence.[3]

Assumption 1. *There exists a graph $\mathcal{G}$ with vertex set $\mathcal{A}$ such that if $\alpha\in\mathcal{A}$ is not connected by any edge to any vertex in $\mathcal{A}'\subset\mathcal{A}$, then $O_{\alpha,t}$ is conditionally independent of $(O_{\alpha',t})_{\alpha'\in\mathcal{A}'}$ given F_{t-1} and (possibly) a known, fixed summary measure $\bar{Z}_t := \mathrm{Summ}(\bar{O}_t)$ of each observation $\bar{O}_t$.*[4]

For every $t\geq 1$ the summary measure $\bar{Z}_t$ writes as $\bar{Z}_t := (Z_{\alpha,t})_{\alpha\in\mathcal{A}} \in \mathcal{Z}^{\mathcal{A}}$. It is said *fixed* because it is derived from $\bar{O}_t$ by evaluating at $\bar{O}_t$ the fixed (in $t\geq 1$ and $\alpha\in\mathcal{A}$) function Summ. The adverb *possibly* hints at the case where Summ maps every $\bar{O}_t$ to an uninformative, empty summary.

We let $\deg(\mathcal{G})$ denote 1 plus the maximum degree of $\mathcal{G}$ (*i.e.*, 1 plus the largest number of edges that are incident to a vertex in $\mathcal{G}$). The smaller is $\deg(\mathcal{G})$, the more conditional independence we can rely on.

Our main objective is to estimate a feature $\theta^\star$ of the law $\mathbb{P}$ of $(\bar{O}_t)_{t\geq 1}$, an element of a parameter space Θ that is known to minimize over Θ the risk induced by a loss ℓ and $\mathbb{P}$. We consider the

[3] Janson [2004] exploits the finer notion of fractional chromatic numbers.

[4] This notion of conditional dependency graph is weaker than the one that requires that $(O_{\alpha,t})_{\alpha\in\mathcal{A}_1}$ and $(O_{\alpha,t})_{\alpha\in\mathcal{A}_2}$ be conditionally independent given F_{t-1} and $\bar{Z}_t$ whenever $\mathcal{A}_1,\mathcal{A}_2$ are disjoint subsets of $\mathcal{A}$ with no edge between them.

specific situation where the feature $\theta^\star$ can also be defined as the shared minimizer over Θ of all the risks induced by a loss ℓ and all the conditional marginal laws of $O_{\alpha,t}$ given $Z_{\alpha,t}$ ("all" refers to all $\alpha \in \mathcal{A}$ and $t \geq 1$).

For instance, we can address a situation where, firstly, each $O_{\alpha,t}$ decomposes as $O_{\alpha,t} := (X_{\alpha,t}, Y_{\alpha,t})$ with $X_{\alpha,t} \in \mathcal{X}$ a collection of (α,t)-specific covariates and $Y_{\alpha,t} \in [-1,1]$ a corresponding outcome of interest; secondly, under $\mathbb{P}$, the exists a (fixed) graph $\mathcal{G}$ with vertex set $\mathcal{A}$ such that, for all $t \geq 1$, if $\alpha \in \mathcal{A}$ is not connected by any edge to any vertex in $\mathcal{A}' \subset \mathcal{A}$, then $O_{\alpha,t}$ is conditionally independent of $(O_{\alpha',t})_{\alpha' \in \mathcal{A}'}$ given F_{t-1}; thirdly, there exists under $\mathbb{P}$ (a fixed) $\theta^\star : \mathcal{X} \to [-1,1]$ such that $\mathbb{E}(Y_{\alpha,t} | X_{\alpha,t} = x, F_{t-1}) = \theta^\star(x)$ for all $x \in \mathcal{X}$. In that situation, the loss ℓ can be the least-square loss function that maps any $\theta : \mathcal{X} \to [-1,1]$ to the function $(x,y) \mapsto (y - \theta(x))^2$. Note that here, every $Z_{\alpha,t}$ is empty.

Generally, we make the following assumption.

Assumption 2. *There exists a loss function $\ell : \Theta \to \mathbb{R}^{\mathcal{O} \times \mathcal{Z}}$ such that the feature of interest $\theta^\star$ minimizes all the risks $\theta \mapsto \mathbb{E}[\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) | Z_{\alpha,t}, F_{t-1}]$ over Θ, "all" referring to all $\alpha \in \mathcal{A}$ and $t \geq 1$. Moreover, for every $\theta \in \Theta$ and sequence $(\theta_t)_{t \geq 1}$ of elements of Θ adapted to $(F_t)_{t \geq 1}$ (i.e., such that each θ_t is F_t-measurable), for all $t \geq 2$ and non-negative integers $\varepsilon_1, \varepsilon_2$ such that $\varepsilon_1 + \varepsilon_2 = 2$,*

$$\mathbb{E}\left[\sum_{\alpha \in \mathcal{A}} (\ell(\theta_{t-1})(O_{\alpha,t}, Z_{\alpha,t}))^{\varepsilon_1} \times (\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}))^{\varepsilon_2} \middle| \bar{Z}_t, F_{t-1}\right] = \sum_{\alpha \in \mathcal{A}} \mathbb{E}\left[(\ell(\theta_{t-1})(O_{\alpha,t}, Z_{\alpha,t}))^{\varepsilon_1} \times (\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}))^{\varepsilon_2} | Z_{\alpha,t}, F_{t-1}\right].$$

Assumption **A2** guarantees some form of stationarity in $\mathbb{P}$ pertaining to its feature of interest $\theta^\star$. Thanks to it there is hope that we can learn $\theta^\star$ from $\bar{O}_1, \ldots, \bar{O}_t$ even with t small if the cardinality $|\mathcal{A}|$ of $\mathcal{A}$ is large (in fact, if the ratio $|\mathcal{A}| / \deg(\mathcal{G})$ is large).

Section 2.1 presents the one-step ahead sequential Super Learner, a collection of assumptions on the law $\mathbb{P}$ of the time series $(\bar{O}_t)_{t \geq 1}$ and on its feature $\theta^\star$, and our theoretical analysis of the one-step ahead sequential Super Learner's performance under these assumptions. Section 2.2 summarizes the content of Section 2.1 and Section 2.3 gathers comments on Section 2.1. The proofs are presented in Appendix A and B.

2.1 The one-step ahead sequential Super Learner and its oracular performances

The one-step ahead sequential Super Learner. Let $\widehat{\theta}_1, \ldots, \widehat{\theta}_J$ be J algorithms to learn $\theta^\star$ from $(\bar{O}_t)_{t\geq 1}$. In words, for each $j \in [\![J]\!] := \{1, \ldots, J\}$, $\widehat{\theta}_j$ is a procedure that, for every $t \geq 1$, maps $\bar{O}_1, \ldots, \bar{O}_t$ to an element of a j-specific subset Θ_j of Θ, namely $\theta_{j,t} \in \Theta_j$ (by convention, $\theta_{j,0}$ is a fixed, pre-specified element of Θ_j). The one-step ahead sequential Super Learner that we are about to introduce is a meta-algorithm that learns, as data accrue, which algorithm in the aforementioned collection performs best.

Strictly speaking, the one-step ahead sequential Super Learner really is an online algorithm if each of the J algorithms is online, that is, if for each $j \in [\![J]\!]$ and $t \geq 1$, the making of $\theta_{j,t}$ consists in an update of $\theta_{j,t-1}$ based on newly accrued data $\bar{O}_t$. If that is not the case, then the Super Learner is merely a sequential algorithm, updated at every time step t.

The measure of performance takes the form of an average cumulative risk conditioned on the observed sequence $(\bar{Z}_t)_{t\geq 1}$. For every $j \in [\![J]\!]$, the risk (for short) of $\widehat{\theta}_j$ till time $t \geq 1$ is defined as

$$\widetilde{R}_{j,t} := \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[\bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau)\middle|\bar{Z}_\tau, F_{\tau-1}\right] \quad \text{where} \tag{1}$$

$$\bar{\ell}(\theta)(\bar{O}_\tau, \bar{Z}_\tau) := \frac{1}{|\mathcal{A}|}\sum_{\alpha\in\mathcal{A}} \ell(\theta)(O_{\alpha,\tau}, Z_{\alpha,\tau}) \quad \text{for all } \theta \in \Theta, \tau \geq 1. \tag{2}$$

The empirical counterpart of (1) is

$$\widehat{R}_{j,t} := \frac{1}{t}\sum_{\tau=1}^{t} \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) = \frac{1}{t|\mathcal{A}|}\sum_{\tau=1}^{t}\sum_{\alpha\in\mathcal{A}} \ell(\theta_{j,\tau-1})(O_{\alpha,\tau}, Z_{\alpha,\tau}). \tag{3}$$

At each time $t \geq 1$, the collection of (j,t)-specific empirical risks is minimized at index $\widehat{j}_t$:

$$\widehat{j}_t \in \operatorname*{arg\,min}_{j\in[\![J]\!]} \widehat{R}_{j,t} \tag{4}$$

(the unlikely ties are broken arbitrarily). The one-step ahead sequential Super Learner is the meta-algorithm that learns $\theta^\star$ by mapping $\bar{O}_1, \ldots, \bar{O}_t$ to $\theta_{\widehat{j}_t,t}$ for every $t \geq 1$.

To assess how well the one-step ahead sequential Super Learner performs, we compare its risk to that of the oracular algorithm that learns $\theta^\star$ by mapping $\bar{O}_1, \ldots, \bar{O}_t$ to $\theta_{\widetilde{j}_t,t}$ at each time $t \geq 1$,

where

$$\widetilde{j}_t \in \arg\min_{j \in [\![J]\!]} \widetilde{R}_{j,t} \tag{5}$$

(again, the unlikely ties are broken arbitrarily). This is discussed next.

Comparing the one-step ahead sequential Super Learner to its oracular counterpart. So far we have defined the risks of $\widehat{\theta}_1, \ldots, \widehat{\theta}_J$, see (1). By analogy, for every $\theta \in \Theta$ and $t \geq 1$, let the risk of θ at time t be

$$\widetilde{R}_t(\theta) := \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{E}\left[\bar{\ell}(\theta)(\bar{O}_\tau, \bar{Z}_\tau) \middle| \bar{Z}_\tau, F_{\tau-1}\right].$$

The risk $\widetilde{R}_t(\theta)$ can be interpreted as the risk till time $t \geq 1$ of a dummy algorithm that constantly maps $\bar{O}_1, \ldots, \bar{O}_t$ to θ (the algorithm is said dummy because it does not learn). Let $\theta^\circ \in \Theta$ be such that

$$\widetilde{R}_t(\theta^\circ) \leq \min_{j \in [\![J]\!]} \min_{\theta \in \Theta_j} \widetilde{R}_t(\theta).$$

Under **A2**, θ° could be set to $\theta^\star$, but other choices might be made on a case by case basis. Our main results compare the excess risks of the one-step ahead sequential Super Learner and of the oracle, that is, they compare

$$\widetilde{R}_{\widehat{j}_t,t} - \widetilde{R}_t(\theta^\circ) \quad \text{to} \quad \widetilde{R}_{\widetilde{j}_t,t} - \widetilde{R}_t(\theta^\circ).$$

They rely on the following assumptions.

For every $\theta \in \Theta$, let $\Delta^\circ \ell(\theta) := \ell(\theta) - \ell(\theta^\circ)$.

Assumption 3. *There exists $b_1 > 0$ such that $\sup_{\theta \in \Theta} \|\Delta^\circ \ell(\theta)\|_\infty \leq b_1$. Moreover there exists $b_2 \in]0, 2b_1]$ such that, almost surely, for all $\alpha \in \mathcal{A}$, $t \geq 1$ and $\theta \in \Theta$,*

$$|\Delta^\circ \ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) - \mathbb{E}\left[\Delta^\circ \ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) | Z_{\alpha,t}, F_{t-1}\right]| \leq b_2.$$

Assumption 4. *There exist $\beta \in]0, 1]$ and $\gamma > 0$ such that, almost surely, for all $\alpha \in \mathcal{A}$, $t \geq 1$ and*

$\theta \in \Theta$,

$$\mathbb{E}\left[\left(\Delta^{\circ}\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t})\right)^2 \middle| Z_{\alpha,t}, F_{t-1}\right] \leq \gamma\Big(\mathbb{E}\left[\Delta^{\circ}\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t})|Z_{\alpha,t}, F_{t-1}\right]\Big)^{\beta}.$$

Assumption 5. *There exists $v_1 > 0$ such that, almost surely, for all $\alpha \in \mathcal{A}$, $t \geq 1$ and $\theta \in \Theta$,*

$$\mathbb{V}\mathrm{ar}\left[\Delta^{\circ}\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t})|Z_{\alpha,t}, F_{t-1}\right] \leq v_1.$$

Assumption **A4** is a so-called "variance bound", a well-known concept in statistical learning theory [Bartlett et al., 2005, Koltchinskii, 2006, Bartlett et al., 2006]. Under **A3**, the radius of the loss class is bounded. Note that if **A3** is met, then so is **A5** necessarily. We can now state our main results.

Theorem 1 (High probability oracular inequality)**.** *Suppose that **A1**, **A2**, **A3**, **A4** and **A5** are met. Define*

$$v_2 := \frac{3\pi}{2}\left[\left(\frac{15b_2}{|\mathcal{A}|/\deg(\mathcal{G})}\right)^2 + \frac{64v_1}{|\mathcal{A}|/\deg(\mathcal{G})}\right]. \tag{6}$$

Fix arbitrarily two integers $N, N' \geq 2$ and a real number $a > 0$, then set $\underline{x}(a, N) := a[2^{-N}v_2/\gamma]^{1/\beta}$, $\underline{x}'(a, N') := ab_1 2^{-N'}$. For all $t \geq 1$ and $x \geq \underline{x}(a, N)$, it holds that

$$\mathbb{P}\left[\widetilde{R}_{\widehat{j}_t,t} - \widetilde{R}_t(\theta^{\circ}) \geq (1+2a)\left(\widetilde{R}_{\widetilde{j}_t,t} - \widetilde{R}_t(\theta^{\circ})\right) + x\right] \leq 2JN\left[\exp\left(-\frac{tx^{2-\beta}}{C_1(a)}\right) + \exp\left(-\frac{tx}{C_2(a)}\right)\right], \tag{7}$$

where $C_1(a) := 2^{5-\beta}(1+a)^2\gamma/a^{\beta}$, $C_2(a) := 8(1+a)b_2/3$. Moreover, for all $t \geq 1$ and $x \geq \underline{x}'(a, N')$, it also holds that

$$\mathbb{P}\left[\widetilde{R}_{\widehat{j}_t,t} - \widetilde{R}_t(\theta^{\circ}) \geq (1+2a)\left(\widetilde{R}_{\widetilde{j}_t,t} - \widetilde{R}_t(\theta^{\circ})\right) + x\right] \leq 2e^2JN'\left[\exp\left(-\frac{[|\mathcal{A}|/(t^{\beta}\deg(\mathcal{G}))]x^{2-\beta}}{C_1'(a)}\right) + \exp\left(-\frac{[|\mathcal{A}|/\deg(\mathcal{G})]x}{C_2'(a)}\right)\right], \tag{8}$$

where $C_1'(a) := 2^{6+2\beta}e^2(1+a)^2\gamma/a^{\beta}$, and $C_2'(a) := 60e(1+a)b_2$.

We derive the following oracular inequality in expectation from Theorem 1.

Corollary 2 (Oracular inequality for the expected risk)**.** *Suppose that* ***A1****,* ***A2****,* ***A3****,* ***A4*** *and* ***A5*** *are met. For any* $a \in]0, 1]$*, it holds that*

$$\mathbb{E}\left[\widetilde{R}_{\widehat{j}_t,t} - \widetilde{R}_t(\theta^\circ) - (1+2a)\left(\widetilde{R}_{\widetilde{j}_t,t} - \widetilde{R}_t(\theta^\circ)\right)\right] \leq 3\left(\frac{C_1(a)}{t}\log(2JN)\right)^{1/(2-\beta)} + \frac{2C_2(a)}{t}\log(2JN) \quad (9)$$

provided that $N \geq 2$ *is chosen so that*

$$N \geq \frac{\beta}{2-\beta}\frac{\log(t) + \log(C_3)}{\log(2)} \quad (10)$$

where $C_3 := (v_2/\gamma)^{(2-\beta)/\beta}/(2^{5-\beta}\gamma)$ *with* v_2 *given by* (6)*. Moreover, it also holds that*

$$\mathbb{E}\left[\widetilde{R}_{\widehat{j}_t,t} - \widetilde{R}_t(\theta^\circ) - (1+2a)\left(\widetilde{R}_{\widetilde{j}_t,t} - \widetilde{R}_t(\theta^\circ)\right)\right] \leq 3\left(\frac{C_1'(a)}{|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))}\log(2JN')\right)^{1/(2-\beta)} + \frac{2C_2'(a)}{|\mathcal{A}|/\deg(\mathcal{G})}\log(2JN') \quad (11)$$

provided that $N' \geq 2$ *is chosen so that*

$$N' \geq \frac{\beta}{2-\beta}\frac{\log(|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))) + \log(C_3')}{\log(2)} \quad (12)$$

where $C_3' := b_1/(2^{6+2\beta}e^2\gamma)$*.*

2.2 Summary of Section 2.1

Given J algorithms $\widehat{\theta}_1, \ldots, \widehat{\theta}_J$ to learn $\theta^\star$ from $(\bar{O}_t)_{t\geq 1}$, the one-step ahead sequential Super Learner is a meta-algorithm that learns, as data accrue, which one performs best. In words, for each $j \in [\![J]\!] := \{1, \ldots, J\}$, $\widehat{\theta}_j$ is a procedure that, for every $t \geq 1$, maps $\bar{O}_t$ to an element $\theta_{j,t}$ of Θ.

Strictly speaking, the one-step ahead sequential Super Learner really is an online algorithm if each of the J algorithms is online, that is, if for each $j \in [\![J]\!]$ and $t \geq 1$, the making of $\theta_{j,t}$ consists in an update of $\theta_{j,t-1}$ based on newly accrued data $\bar{O}_t$. If that is not the case, then the Super Learner is merely a sequential algorithm, updated at every time step t.

The (unknown) t-specific measure of performance of each $\widehat{\theta}_j$, $\widetilde{R}_{j,t}$ (1), takes the form of an

average cumulative risk conditioned on the observed sequence $(\bar{Z}_t)_{t\geq 1}$ introduced in Assumption **A1**. The (unknown) t-specific oracular meta-algorithm is indexed by the oracular $\widetilde{j}_t \in [\![J]\!]$ (5).

We use the (known) t-specific empirical counterpart $\widehat{R}_{j,t}$ (3) of $\widetilde{R}_{j,t}$ to estimate $\widetilde{j}_t$ with the (known) t-specific $\widehat{j}_t$ (4). Algorithm $\widehat{\theta}_j$ with $j = \widehat{j}_t$ is the one-step ahead sequential Super Learner at time t.

The oracular inequalities in Corollary 2 have a familiar flavor for whoever is interested in Super Learning or, more generally, the aggregation or stacking of algorithms. In essence, as more data accrue, the expected risk of the one-step ahead sequential Super Learner is smaller than $(1 + a)$ (a chosen small) times the expected risk of the oracular meta-algorithm up to an error term of the form constant times $(\log(J \log(\mathcal{I}^2))/\mathcal{I}^2)^{1/(2-\beta)}$ where $\mathcal{I}$ grows like the amount of information available (the constant $\beta \in]0, 1]$ appears in one of the assumptions). In (9), $\mathcal{I}^2$ equals t. In (11), $\mathcal{I}^2$ equals $|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))$. In the next section we show that if the ratio $|\mathcal{A}|/\deg(\mathcal{G})$ is sufficiently large (both in absolute terms and relative to t) (see (14)), then the oracular inequality (11) can be sharper than the oracular inequality (9) in Corollary 2, revealing that we managed to leverage a large ratio $|\mathcal{A}|/\deg(\mathcal{G})$ in the face of a small t.

2.3 Comments

Leveraging a large ratio $|\mathcal{A}|/\deg(\mathcal{G})$ in the face of a small t. Our results generalize those of Benkeser et al. [2018] in two aspects. First, they do not require assumptions akin to their assumptions $A3$ and $A4$, which are meant to deal with the randomness at play in $\widetilde{R}_{j,t}$ and $\mathbb{V}\mathrm{ar}[\Delta^\circ \ell(O_{\alpha,t}, Z_t)|Z_t, F_{t-1}]$. Instead we exploit a so-called stratification argument inspired by Cesa-Bianchi and Gentile [2008]. Second, our results leverage the fact that, as explained at the beginning of Section 2, each t-specific observation is a collection $(O_{\alpha,t})_{\alpha\in\mathcal{A}}$ of (α, t)-specific data points with a large amount of conditional independence between them, as modelled by the conditional dependency graph $\mathcal{G}$. Recall that $\deg(\mathcal{G})$ equals 1 plus the maximum degree of $\mathcal{G}$. The smaller is $\deg(\mathcal{G})$ the more conditional independence we can rely on.

If one chooses $N = N'$ in (9) and (11), then it is easy to check that the two terms in the right-hand side expression of (11) are smaller than their counterparts in (9) if and only if

$$t^{1+\beta} \leq \frac{|\mathcal{A}|/\deg(\mathcal{G})}{2e^2 8^\beta} \quad \text{and} \quad t \leq \frac{|\mathcal{A}|/\deg(\mathcal{G})}{45e/2}. \tag{13}$$

Furthermore, a simple sufficient condition for (13) to be met is

$$t^{1+\beta} \leq \frac{|\mathcal{A}|/\deg(\mathcal{G})}{2e^2 8^\beta} \quad \text{and} \quad \frac{|\mathcal{A}|}{\deg(\mathcal{G})} \geq 24e \times (3/(2e))^{1/\beta}\,. \tag{14}$$

Thus, if (14) is met (note that $24e \times 3/(2e) = 36 \geq 24e \times (3/(2e))^{1/\beta}$ whatever is $\beta \in]0,1]$) and if we make the following (valid) choices in Corollary 2,

$$N = N' \geq \frac{\beta}{2-\beta} \frac{\log(|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))) + \log(C_3') + \big(\log[C_3/(2e^2 8^\beta C_3')]\big)_+}{\log(2)},$$

then the oracular inequalities (9) and (11) for the expected risk hold true, the latter being sharper than the former. In words, our analysis does take advantage of the fact that $|\mathcal{A}|/\deg(\mathcal{G})$ is large in the face of t being comparatively small.

A few details on the proofs. Theorem 1 notably hinges on the Fan-Grama-Liu concentration inequality for martingales [Theorem 3.10 in Bercu et al., 2015] and on the following result, tailored to our needs and derived from a concentration inequality for sums of partly dependent random variables shown by Janson [2004]. For each $j \in [\![J]\!]$ and $t \geq 1$, introduce the two (j,t)-specific averages of conditional variances

$$\mathrm{var}_{j,t} := \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \mathbb{V}\mathrm{ar}\left[\Delta^\circ \ell(\theta_{j,\tau-1})(O_{\alpha,t} Z_{\alpha,t}) | Z_{\alpha,t}, F_{t-1}\right], \tag{15}$$

$$\widetilde{\mathrm{var}}_{j,t} := \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{V}\mathrm{ar}\left[\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) \big| \bar{Z}_\tau, F_{\tau-1}\right]. \tag{16}$$

Theorem 3. *Suppose that* ***A3*** *and* ***A4*** *are met. For each* $j \in [\![J]\!]$, $\widetilde{\mathrm{var}}_{j,t} \leq v_2$ *almost surely (see* (6) *for the definition of* v_2*). Moreover, for any* $V > 0$ *and all* $x \geq 0$*, if*

$$\widetilde{\mathcal{F}}_V := \left[\max_{\tau \in [\![t]\!]} \{\mathrm{var}_{j,\tau}\} \leq V\right], \tag{17}$$

then

$$\mathbb{P}\left[|[\widehat{R}_{j,t} - \widehat{R}_t(\theta^\circ)] - [\widetilde{R}_{j,t} - \widetilde{R}_t(\theta^\circ)]| \geq x, \widetilde{\mathcal{F}}_V\right] \leq \exp\left(2 - \frac{[|\mathcal{A}|/\deg(\mathcal{G})]x^2}{32e^2 V + 15e b_2 x}\right). \tag{18}$$

Our proof of Theorem 3 consists in deriving a Rosenthal inequality from Janson's concentration

inequality [2004], following Petrov's line of proof [1995], in using a convexity argument, then in applying the same method as in [Dedecker, 2001, Corollary 3(b)] (inspired by the proof of Theorem 6 in [Doukhan et al., 1984]). Inequality (18) plays a key role in the derivation of (11). The fact that the first term in the right-hand side expression in (11) features $|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))$ and not $|\mathcal{A}|/\deg(\mathcal{G})$ may be deemed pessimistic but is inherent to our scheme of proof. Note that substituting a sharp Marcinkiewicz-Zygmund-like inequality [Rio, 2009, Theorem 2.9] for the convexity argument that leads to (49) does not solve the issue.

Furthermore, it is noteworthy that our results extend seamlessly to the case that every expression $\ell(\theta_{\tau-1})(O_{\alpha,\tau}, Z_{\alpha,\tau})$ with $\theta_{\tau-1}$ $F_{\tau-1}$-measurable is replaced by an expression of the form $\ell(\theta_{\tau-1})(O_{\alpha,\tau}, Z_{\alpha,\tau}) \times \omega_\tau(O_{\alpha,\tau}, Z_{\alpha,\tau})$, where ω_τ is a $F_{\tau-1}$-measurable weighting function. This proves very useful in the context of reinforcement learning, allowing to rely on importance sampling weighting.

Comments on the assumptions. Assumptions **A2**, **A3**, **A4**, **A5** are quite typical. Like in the context of the application motivating our study, suppose for example that each $O_{\alpha,t}$ decomposes as $O_{\alpha,t} := (Z_{\alpha,t}, X_{\alpha,t}, Y_{\alpha,t}) \in \mathcal{Z} \times \mathcal{X} \times [-B, B] =: \mathcal{O}$ where $X_{\alpha,t} \in \mathcal{X}$ is a collection of covariates, $Z_{\alpha,t} \in \mathcal{Z}$ is a fixed summary measure thereof (*i.e.*, as explained earlier, $Z_{\alpha,t}$ is derived from $X_{\alpha,t}$ by evaluating at $X_{\alpha,t}$ a known, fixed (in $t \geq 1$ and $\alpha \in \mathcal{A}$) function), and $Y_{\alpha,t} \in [-B, B]$ is a bounded real-valued outcome of primary interest. Suppose moreover that, for all $\alpha \in \mathcal{A}$ and $t \geq 1$, the conditional law of $Y_{\alpha,t}$ given $(X_{\alpha,t}, Z_{\alpha,t}) = (x, z)$, $(Z_{\alpha',t})_{\alpha' \in \mathcal{A}}$ and F_{t-1} admits the conditional density $y \mapsto f^\star(y|x, z)$ with respect to some measure $\lambda(dy)$ on $[-B, B]$. In this context, the conditional expectation $y \mapsto \theta^\star(y|x, z)$ of $Y_{\alpha,t}$ given $(X_{\alpha,t}, Z_{\alpha,t}) = (x, z)$ (for all (x, z) in the support, under $\mathbb{P}$, of any $(X_{\alpha,t}, Z_{\alpha,t})$) is an eligible feature of interest.

Let Θ be the set of measurable functions on $\mathcal{X} \times \mathcal{Z}$ taking their values in $[-B, B]$. Given by $\ell(\theta) : ((z, x, y), z) \mapsto (y - \theta(x, z))^2$ (for all $\theta \in \Theta$), the least-square loss function $\ell : \Theta \to \mathbb{R}^{\mathcal{O} \times \mathcal{Z}}$ satisfies **A2**. In addition, we can choose $\theta^\circ := \theta^\star$ and **A3**, **A4** (with $\beta = 1$) and **A5** are met. The fact that **A4** is met follows from a classical argument of strong convexity recalled, for self-containedness, in Appendix A.

3 Anticipating the cost of natural disasters

In this section, we apply one-step ahead sequential Super Learning to anticipate the cost of natural disasters. Section 3.1 presents the context and objective in details, Section 3.2 describes the data that we exploit, and Section 3.3 models the problem in the terms of the theoretical Section 2. Then, Section 3.4 discusses the implementation of the one-step ahead sequential Super Learner and Section 3.5 presents and comments its results.

3.1 More on the context and the objective

To better anticipate the risks, CCR has developed an expertise in natural disasters modeling. Its cat models exploit portfolios and claims data collected from CCR's cedents to enable a better appreciation of the exposures of CCR, of its cedents and of the French State.

The natural disasters compensation scheme created by French Law n°82-600 is triggered when the following three necessary conditions are met:

1. a government decree declaring a natural disaster must be published in the French Journal Officiel;

2. the lost and/or damaged property must be covered by a property and casualty insurance policy;

3. a causal link must exist between the declared natural disaster and the sustained loss and/or damage.

The mayor of a city can request the government declaration of natural disaster by sending a form to their prefect. All over France, the prefects gather the forms and send them to the relevant Interministerial Commission. The commission examines all requests and delivers the declaration of natural disaster if additional criteria are met. These criteria characterize what is considered as a natural disaster. For instance, for drought events (the natural catastrophes that we focus on, also known as *subsidence*[5] *events* in the literature, for reasons that the next paragraph clarifies), the criteria evaluate the intensity of the drought. It is noteworthy that the criteria are regularly updated by the commission – we shall discuss further this point in Section 4. If the Interministerial

[5]The process by which land or buildings sink to a lower level.

Commission delivers a favorable opinion, confirmed by the publication in the Journal Officiel of a government decree declaring a disaster, then CCR indemnifies the insurance companies once they have indemnified the policyholders.

As revealed earlier, we focus on drought events. Such events are caused by the clay shrinking and swelling during a calendar year (and must be distinguished from *agricultural* drought events). Drought events entail cracks on buildings, which can be covered by an insurance policy. In order to manage the risks inherent in the natural disasters compensation scheme, CCR must anticipate the costs generated by drought events in particular. This is crucial for the pricing of non-proportional reinsurance treaties, and for reserving (that is to say, to anticipate forthcoming payments). The present study tackles the challenge of predicting the cost of drought events from a data set that we describe next.

3.2 Data

The data set that we exploit to predict the costs of drought events is composed of several data sets of different natures. The data are commonly grouped in two classes, depending on whether they concern the natural disaster itself or any of the remaining relevant characteristics that complete the description of the financial impact of the natural disaster on the insurance industry. We choose to group the data in two other classes, depending on whether they come from the cedents or from another source.

Data from cedents. CCR reinsures 90% of the French natural disasters insurance market. Contractually, CCR's cedents must share their portfolios (*i.e.*, the location and characteristics of the insured goods) and claims data. Thanks to this mechanism, CCR has gathered a large data set that runs from 1990 to this day.

Data from other sources. The data from cedents are enriched with other data collected from four French organizations. The National Institute for Statistical and Economic Studies (INSEE) and Geographic National Institute (IGN) provide information on the cities (population, area, proportions of buildings by years of construction for INSEE; tree coverage rate for IGN). The French Geological Survey (BRGM) provides a mapping of the clay shrinkage-swelling hazards in France.

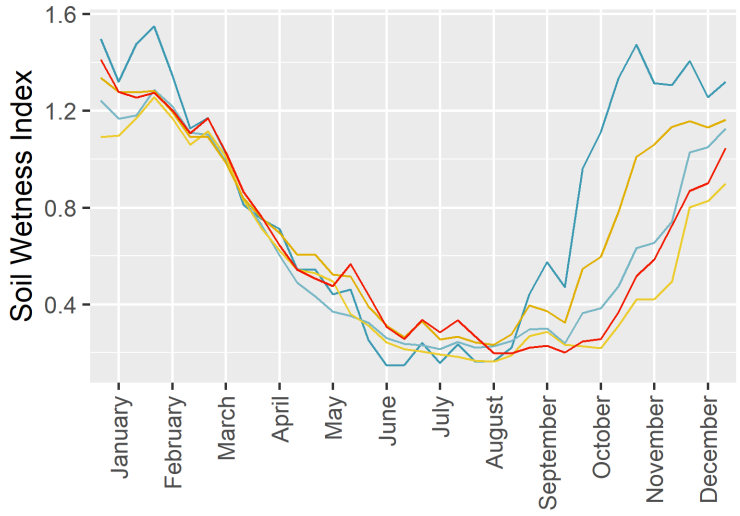


Figure 1: Chunks from five arbitrarily chosen time series of Soil Wetness Index (SWI) over the course of one year. It does not come as a surprise that the soil is drier during summer than during winter.

Finally, Météo France provides a soil wetness index (SWI). This last feature consists of time series of values (one every decade) ranging between -3.33 (very dry soil) and 2.33 (very wet soil). Figure 1 presents five one-year chunks of SWI time series.

Working at the city-level. It is noteworthy that the spatial resolution of SWI data is $8 \times 8\text{km}^2$, which is much larger than the 90%-quantile of the French cities area (30 km^2; only 1.3% of the French cities have an area larger than 65 km^2 – data from 2014). This issue will be discussed further in Section 4. It justifies why we choose to work at the city-level as opposed to the house level, by aggregating at every time point all data from each city into a single, time and city-specific observation.

- *City-level costs of drought events.* For every time point and each city, the city-level cost of drought event is the sum of all house claims over the city's area.

- *City-level SWI.* For every time point and each city, the city-level SWI is the convex average of the SWIs of the $8 \times 8\text{km}^2$ squares that overlap the city's area, the weights being proportional to the areas of the intersections.

- *City-level description.* For every time point and each city, a city-level description encapsulates the city's profile. The description is multi-faceted. It contains: an indicator of whether or

not a natural disaster was declared by the government; the overall insured value obtained by summing the insured values over the city's area; a summary of the city's clay hazard, defined as the proportions of houses falling in each of four categories of clay hazard; a summary of the city's dwelling age, *i.e.*, how old houses are, under the form of the proportions of houses falling in each of four categories; the climatic and seismic zones (a five-category and a four-category variables); a summary of the city's vegetation; the city's number of houses, population, area, average altitude, and density, defined as the ratio of the number of houses to the area. In addition, a variety of features are described by quantiles that summarize distributions (*e.g.*, the 30-quantiles of the distribution of the house-specific product of SWI and insured value, or the 30-quantile of the distribution of the house-specific product of the ground slope and insured value, to mention just a few). Overall, the city-level description consists of a little more than 430 variables.

3.3 Modeling

The sequence of observations. In the context of the anticipation of the cost of natural disasters, each $O_{\alpha,t}$ decomposes as $O_{\alpha,t} := (Z_{\alpha,t}, X_{\alpha,t}, Y_{\alpha,t})$ where

- $Y_{\alpha,t} \in [0, B]$ is the city-level cost of the drought event for city α at year t,
- $Z_{\alpha,t}$ is the city-level SWI for city α at year t,
- $X_{\alpha,t}$ is the city-level description that encapsulates city α's profile at year t, including an indicator $W_{\alpha,t} \in \{0, 1\}$ that equals 1 if and only if a natural disaster has been declared by the government for that city and that year.

By convention, $Y_{\alpha,t} = 0$ if $W_{\alpha,t} = 0$ (that is, in the absence of a declaration of natural disaster). Formally, $X_{\alpha,t}$ includes $Z_{\alpha,t}$. For notational simplicity, we rewrite each $X_{\alpha,t}$ as $(W_{\alpha,t}, X_{\alpha,t}, Z_{\alpha,t}) \in \{0, 1\} \times \mathcal{X} \times \mathcal{Z}$, the "new" $X_{\alpha,t}$ being the "old" $X_{\alpha,t}$ deprived of $Z_{\alpha,t}$ (but not of $W_{\alpha,t}$).

The feature of interest and related loss function. We assume that, for all $\alpha \in \mathcal{A}$ and $t \geq 1$, the conditional law of $Y_{\alpha,t}$ given $(W_{\alpha,t}, X_{\alpha,t}, Z_{\alpha,t}) = (1, x, z)$, $(Z_{\alpha',t})_{\alpha' \in \mathcal{A}}$ and F_{t-1} admits a conditional density $y \mapsto f^\star(y|x, z)$ with respect to some measure on $[0, B]$. In this context, the

conditional expectation $y \mapsto \theta^\star(y|x,z)$ of $Y_{\alpha,t}$ given $(W_{\alpha,t}, X_{\alpha,t}, Z_{\alpha,t}) = (1,x,z)$ (for all (x,z) in the support of any $(X_{\alpha,t}, Z_{\alpha,t})$ conditionally on $W_{\alpha,t} = 1$) is an eligible feature of interest.

Set $\mathcal{O} := \{0,1\} \times \mathcal{X} \times \mathcal{Z} \times [0,B]$ and let Θ be the set of measurable functions on $\mathcal{X} \times \mathcal{Z}$ taking their values in $[0,B]$ and such that $\theta(x,z) = 0$ if $w = 0$. Given by $\ell(\theta) : ((w,x,z,y),z) \mapsto (y - \theta(x,z))^2 \mathbf{1}\{w = 1\}$ (for all $\theta \in \Theta$), the least-square loss function $\ell : \Theta \to \mathbb{R}^{\mathcal{O} \times \mathcal{Z}}$ satisfies **A2**. In addition, we can choose $\theta^\circ := \theta^\star$ and **A3**, **A4** (with $\beta = 1$) and **A5** are met. The fact that **A4** is met follows from the classical argument of strong convexity recalled in Appendix A.

Of $\mathcal{A}$ and $\mathcal{G}$. Here, $\mathcal{A}$ represents the set of French cities. The dependency graph $\mathcal{G}$ used to model the amount of conditional independence operationalizes two different types of spatial dependence: one *geographical* and the other *administrative*. The former corresponds to the dependency caused by the proximity between two cities in geological and meteorological terms as well as in terms of vegetation. The latter corresponds to the dependency caused by the administrative proximity between two cities that belong to a same "communauté de communes" (*i.e.*, community of communes, a federation of municipalities). This second type of spatial dependence is less obvious than the first one. It arises from the fact that a declaration of natural disaster must be requested by the mayor of a city (see Section 3.1). If, in a small federation, a mayor makes such a request, then it is likely that the other mayors will as well.

The cardinality of $\mathcal{A}$ is of order $36,000$. In 2019, there were approximately $1,000$ federations of municipalities in France, regrouping approximately $26,000$ cities. The federation regrouped approximately 30 cities in average.

3.4 Implementation

We implement our statistical analysis in `R` [R Core Team, 2021]. All our Super Learners are implemented based on the `SuperLearner` library [Polley et al., 2019].

Base algorithms. The base algorithms $\widehat{\theta}_1, \ldots, \widehat{\theta}_J$ belong to one among four classes of algorithms: the class of algorithms based on small to moderate-dimensional working models (linear regression; lasso, ridge and elastic net regressions [Simon et al., 2011]; multivariate adaptive regression splines [Milborrow, 2020]; support vector regression [Karatzoglou et al., 2004]; gradient

boosting with linear boosters [Chen et al., 2021]); the class of algorithms based on trees (CART [Therneau and Atkinson, 2019], random forest [Wright and Ziegler, 2017], gradient boosting with tree boosters [Chen et al., 2021]); the class of k-nearest neighbors algorithms; the class of algorithms based on high-dimensional neural networks [Allaire and Chollet, 2021]. Most of the aforementioned algorithms contribute several base algorithms through the choice of different hyper-parameters. The k-nearest-neighbors algorithms are customized. Each of them focuses on one of the quantiles summarizing a feature of interest (see Section 3.2) and uses the Kolmogorov-Smirnov distance as a measure of similarity between every pair of quantiles (viewed as cumulative distribution functions).

Discrete and continuous one-step ahead sequential Super Learners. The one-step ahead sequential Super Learner that learns $\theta^\star$ by mapping $\bar{O}_1, \ldots, \bar{O}_t$ to $\theta_{\widehat{j}_t,t}$ for every $t \geq 1$ (4) is known as a *discrete* Super Learner. The *continuous* Super Learner is the *discrete* Super Learner when the collection of base algorithms consists of all convex combinations $\sum_{j\in[\![J]\!]} \sigma_j \widehat{\theta}_j$ of the base algorithms $\widehat{\theta}_1, \ldots, \widehat{\theta}_J$ where $(\sigma_1, \ldots, \sigma_J)$ ranges over the discretized simplex $\{(\sigma_1, \ldots, \sigma_J) \in \{(k-1)/K : k \in [\![K+1]\!]\}^J : \sum_{j\in[\![J]\!]} \sigma_j = 1\}$ with a large integer K. Note that the cardinality of this collection of base algorithms is of order $\Theta(K^J)$ and much larger than J. This is not overly problematic because J in (9) and (11) plays a role through $\log(J)/\mathcal{I}^2$ with $\mathcal{I}^2 = t$ or $\mathcal{I}^2 = |\mathcal{A}|/(t^\beta \deg(\mathcal{G}))$, one of them at least being supposed large.

For every $t \geq 1$, the σ-specific algorithm $\sum_{j\in[\![J]\!]} \sigma_j \widehat{\theta}_j$ maps $\bar{O}_1, \ldots, \bar{O}_t$ to $\sum_{j\in[\![J]\!]} \sigma_j \theta_{j,t}$. From a computational point of view, we do not use the larger collection of base algorithms obtained by convex combination. Instead, we directly solve

$$\arg\min_{\sigma\in\Sigma} \frac{1}{t} \sum_{\tau=1}^{t} \bar{\ell}\left(\sum_{j\in[\![J]\!]} \sigma_j \theta_{j,\tau-1}\right)(\bar{O}_\tau, \bar{Z}_\tau) \tag{19}$$

(where Σ is the whole simplex), which can be interpreted as a convexified version of (4).

More one-step ahead sequential Super Learners and the overarching one-step ahead sequential Super Learner. We propose and implement two more extensions. The first extension builds upon the interpretation of (19) as the so called meta-learning task consisting in predicting $Y_{\alpha,\tau}$ under the form $\sum_{j\in[\![J]\!]} \sigma_j \theta_{j,\tau-1}(X_{\alpha,\tau}, Z_{\alpha,\tau})$ for all $\alpha \in \mathcal{A}, 1 \leq \tau \leq t$. We consider other meta-learning methods m to predict $Y_{\alpha,\tau}$ based on $\theta_{1,\tau-1}(X_{\alpha,\tau}, Z_{\alpha,\tau})$, $\ldots$, $\theta_{J,\tau-1}(X_{\alpha,\tau}, Z_{\alpha,\tau})$, $X_{\alpha,\tau}, Z_{\alpha,\tau}$

for all $\alpha \in \mathcal{A}, 1 \leq \tau \leq t$. Each meta-learning method m yields its own m-specific (discrete or continuous) Super Learner.

The second extension builds upon the first one. The collection of m-specific Super Learners can be considered as a collection of base algorithms. This raises the question of learning which one performs best. To answer this question, we can rely on what we call the *overarching* (discrete or continuous) Super Learner.

Meta-learning methods and overarching meta-learning method. In view of the four classes of base algorithms described in the first paragraph of this section, the meta-learning methods belong to one among two classes of methods: the class of methods based on small to moderate-dimensional working models (linear regression with nonnegative coefficients [Mullen and van Stokkum, 2012]; lasso, ridge and elastic net regressions [Simon et al., 2011]; support vector regression [Karatzoglou et al., 2004]; gradient boosting with linear boosters [Chen et al., 2021]); the class of algorithms based on trees (extra trees, a variant of random forest [Wright and Ziegler, 2017]; gradient boosting with tree boosters [Chen et al., 2021]). The overarching Super Learner uses the meta-learning method based on linear regression with nonnegative coefficients. The discrete overarching Super Learner selects which among the Super Learners (viewed as base algorithms) performs best. The continuous overarching Super Learner learns which convex combination of the Super Learners (viewed as base algorithms) performs best.

Overall, we implement 27 base algorithms and 48 Super Learners.

3.5 Results

min.	1st qu.	median	mean	3rd qu.	max
23	162.5	607	1072.3	1921.5	4436

Table 1: Quantiles and mean of the yearly numbers of cities for which a declaration of natural disaster was delivered by the government as a result of a drought event. The time series runs from 1995 to 2017. Overall, we count 24,663 such declarations.

We observe the time series $(\bar{O}_t)_{t\geq 1}$ from 1995 to 2017. We also observe the years 2018 and 2019 but do not know yet the city-level or global costs of drought events for these two years. Overall we count 24,663 observations $\bar{O}_{\alpha,\tau}$ for which a declaration of natural disaster was delivered by the

government as a result of a drought event. The quantiles and mean of the yearly numbers of cities for which a declaration of natural disaster was delivered are reported in Table 1.

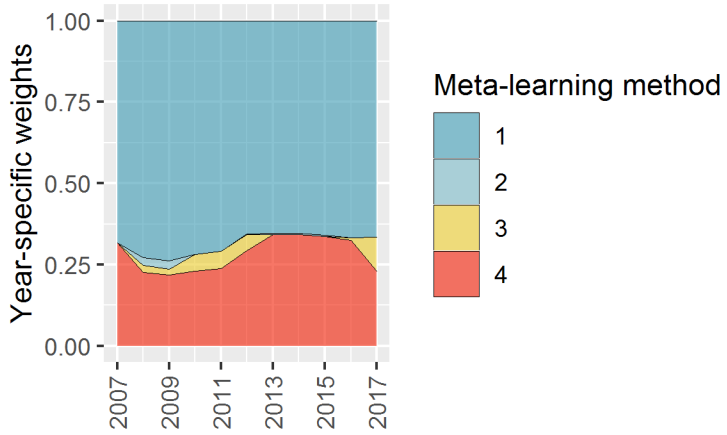


Figure 2: Evolution (from 2007 onward) of the weights attributed by the overarching Super Learner to 4 of its base algorithms, each one a Super Learner itself using its own meta-learning method. The other base algorithms get no weight at all (on this time window).

Among the 48 Super Learners, the overarching continuous Super Learner attributes positive weights to the same four Super Learners consistently from 2007 to 2017, see Figure 2. Moreover, the discrete overarching Super Learner is consistently one of these four Super Learners.

Figure 3 represents the global costs of drought events as predicted by the discrete and continuous overarching Super Learners. We observe that the discrete and continuous overarching Super Learners make predictions that are consistent each year. The experts, who naturally focus on the years for which the real costs happen to be the largest because the financial stake is then the highest, deem them very satisfactory.

Overall, the averages (over the years) of the ratios of the predicted costs to the real costs equal 106% (discrete overarching Super Learner) and 112% (continuous overarching Super Learner). The ratios range from 67% (discrete overarching Super Learner) and 70% (continuous overarching Super Learner), in 2016, to 164% (discrete overarching Super Learner) and 180% (continuous overarching Super Learner), in 2012. The year 2016 is known by the experts to be atypical, and challenging, because the average cost (understood here as the ratio of the total cost of the year's drought event to the corresponding number of declarations of natural disaster delivered that year) is particularly large. Conversely, the average cost in the year 2012 is particularly small.

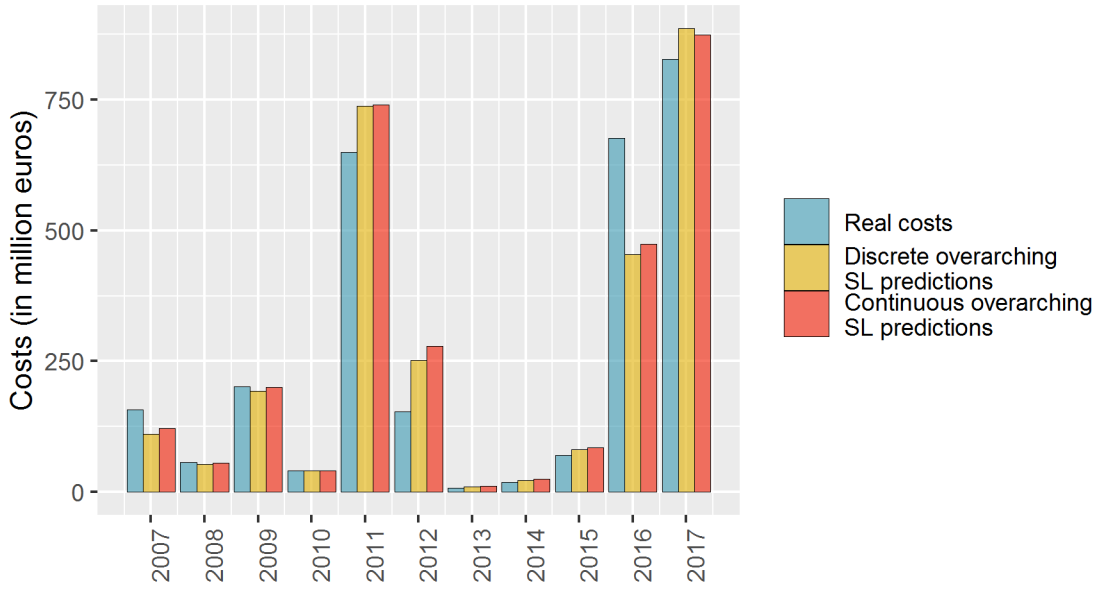


Figure 3: Presentation (from 2007 onward) of the real costs of drought events and their predictions made by the discrete and continuous overarching Super Learners.

4 Discussion

We develop and analyze a meta-algorithm that learns, as data accrue, which among J base algorithms better learns a feature $\theta^\star$ of the law $\mathbb{P}$ of a sequence $(\bar{O}_\tau)_{\tau\geq 1}$, where each $\bar{O}_\tau$ consists of a finite collection $(\bar{O}_{\alpha,\tau})_{\alpha\in\mathcal{A}}$ of many slightly dependent data. We show that the meta-algorithm, an example of Super Learner, leverages a large ratio $|\mathcal{A}|/\deg(\mathcal{G})$ (a measure of the amount of independence among the τ-specific $\bar{O}_{\alpha,\tau}$, $\alpha\in\mathcal{A}$) in the face of a small number t of time points where the time series is observed – see the summary presented in Section 2.2. The study is motivated by the challenge posed by the appreciation of the exposures to drought events of CCR, of its cedents and of the French State. We implement and use two Super Learners to learn to assess the (global) costs by predicting the (local) costs at the city-level – see Section 3.5 for a summary of our results.

Reliable prediction of the cost of a drought event *must* rely on some measure of the drought's intensity. We exploit a soil wetness index (SWI) provided by Météo France. Because the spatial resolution of SWI data is much larger than the 90%-quantile of the French cities area, we choose to work at the city-level rather than at the address-level, by aggregating all the address-specific information at the city-level. In future work, we will *learn* a better measure of the drought's intensity at a finer resolution by combining different sources of information pertaining to the soil

wetness (SWI, rainfall, nature of the soil, to name just a few). Since we know that costs can vary dramatically at the address-level, we also consider to later try and enhance our predictions by zooming in back to the address-level, thanks to the finer resolution measure of soil wetness.

In Section 3.1, we explained that the criteria characterizing what is considered as a natural disaster by the relevant Interministerial Commission are regularly updated. Moreover, even on the narrow time frame of our study, climate change may have affected the severity of droughts on the French territory. From a theoretical viewpoint, the marginal law of the (α, τ)-specific covariate $Z_{\alpha,\tau}$ that describes the severity of the drought depends on τ. We tried to give each $O_{\alpha,\tau}$ an (α, τ)-specific weight to target the (α, t)-specific marginal law of $Z_{\alpha,t}$ when addressing the prediction of the cost of year t. If any, the benefits were dwarfed by the increase in variability caused by the learned weighting scheme.

Acknowledgments. The authors thank Jérôme Dedecker (MAP5, Université de Paris) for his enlightened advice and Thierry Cohignac (Caisse Centrale de Réassurance) for his suggestions to improve the manuscript.

References

J. J. Allaire and F. Chollet. *keras: R Interface to 'Keras'*, 2021. URL https://CRAN.R-project.org/package=keras. R package version 2.4.0.

Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4): 467–493, 2000.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33 (4):1497–1537, 2005.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

D. Benkeser, C. Ju, S. Lendle, and M. J. van der Laan. Online cross-validation-based ensemble learning. *Stat. Med.*, 37(2):249–260, 2018.

B. Bercu, B. Delyon, and E. Rio. *Concentration inequalities for sums and martingales*. SpringerBriefs in Mathematics. Springer, Cham, 2015.

N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Trans. Inform. Theory*, 54(1):386–390, 2008.

T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y Geng, and Y. Li. *xgboost: Extreme Gradient Boosting*, 2021. URL https://CRAN.R-project.org/package=xgboost. R package version 1.4.1.1.

J. Dedecker. Exponential inequalities and functional central limit theorems for a random fields. *ESAIM Probab. Statist.*, 5:77–104, 2001.

P. Doukhan, J. León, and F. Portal. Vitesse de convergence dans le théorème central limite pour des variables aléatoires mélangeantes à valeurs dans un espace de Hilbert. *C. R. Acad. Sci. Paris Sér. I Math.*, 298(13):305–308, 1984.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.*, 2(2):131–154, 2005.

S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures Algorithms*, 24(3):234–248, 2004.

A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL http://www.jstatsoft.org/v11/i09/.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

S. Milborrow. *earth: Multivariate Adaptive Regression Splines*, 2020. URL https://CRAN.R-project.org/package=earth. R package version 5.3.0.

K. Mitchell-Wallace, M. Jones, J. , Hillie, and M. Foote, editors. *Natural Catastrophe Risk Management and Modelling: A Practitioner's Guide*. Wiley-Blackwell, 2017.

K. M. Mullen and I. H. M. van Stokkum. *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*, 2012. URL https://CRAN.R-project.org/package=nnls. R package version 1.4.

V. V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.

E. Polley, E. LeDell, C. Kennedy, and M. J. van der Laan. *SuperLearner: Super Learner Prediction*, 2019. URL https://CRAN.R-project.org/package=SuperLearner. R package version 2.0-26.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

E. Rio. Moment inequalities for sums of dependent random variables under projective conditions. *J. Theoret. Probab.*, 22(1):146–163, 2009.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. URL https://www.jstatsoft.org/v39/i05/.

T. Therneau and B. Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. URL https://CRAN.R-project.org/package=rpart. R package version 4.1-15.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6:Art. 25, 23, 2007.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

A A classical strong convexity argument

Suppose that Θ is convex, and that the loss function $\ell : \Theta \to \mathbb{R}^{\mathcal{O}}$ is a_1-Lipschitz,

$$|\ell(\theta_1) - \ell(\theta_2)| \leq a_1|\theta_1 - \theta_2| \tag{20}$$

and a_2-strongly convex: for all $s \in [0,1]$ and $\theta_1, \theta_2 \in \Theta$,

$$\ell(s\theta_1 + (1-s)\theta_2) - \tfrac{a_2}{2}(s\theta_1 + (1-s)\theta_2)^2 \leq s\left(\ell(\theta_1) - \left(\tfrac{a_2}{2}\theta_1\right)^2\right) + (1-s)\left(\ell(\theta_2) - \left(\tfrac{a_2}{2}\theta_2\right)^2\right)$$

(both inequalities above are understood pointwise). Then the modulus of continuity of ℓ is lower-bounded by $\rho \mapsto \frac{a_2}{8}\rho^2$ in the sense that, for all $\theta_1, \theta_2 \in \Theta$,

$$\tfrac{1}{2}\left(\ell(\theta_1) + \ell(\theta_2)\right) - \ell\left(\tfrac{1}{2}(\theta_1 + \theta_2)\right) \geq \tfrac{a_2}{8}(\theta_1 - \theta_2)^2 \tag{21}$$

(pointwise). Let P be a law on $\mathcal{O}$ such that $P\ell(\theta)$ is well defined for all $\theta \in \Theta$, where we note $Pf := \int f dP$. Choose $\theta^\circ \in \Theta$ such that $P\ell(\theta^\circ) \leq P\ell(\theta)$ for all $\theta \in \Theta$. Then, for all $\theta \in \Theta$,

$$\begin{aligned}
\tfrac{1}{2}P(\ell(\theta) + \ell(\theta^\circ)) &\geq P\ell(\tfrac{1}{2}(\theta + \theta^\circ)) + \tfrac{a_2}{8}P(\theta - \theta^\circ)^2 \\
&\geq P\ell(\theta^\circ) + \tfrac{a_2}{8}P(\theta - \theta^\circ)^2 \\
&\geq P\ell(\theta^\circ) + \tfrac{a_2}{8a_1^2}P(\ell(\theta) - \ell(\theta^\circ))^2,
\end{aligned}$$

where the first inequality follows from (21), the second holds by convexity of Θ and choice of θ°, and the third one follows from (20). Therefore,

$$P(\ell(\theta^\circ) - \ell(\theta))^2 \leq \tfrac{4a_1^2}{a_2}P(\ell(\theta) - \ell(\theta^\circ)),$$

which concludes the argument.

B Proofs

B.1 Proof of Theorem 1

The proof unfolds in three steps.

Step 1: an algebraic decomposition. For all $j \in [\![J]\!]$, $t \geq 1$ and $\theta \in \Theta$, let us define

$$\widetilde{H}_{j,t} := \widetilde{R}_{j,t} - \widetilde{R}_t(\theta^\circ), \quad \widehat{H}_{j,t} := \widehat{R}_{j,t} - \widehat{R}_t(\theta^\circ), \quad \text{and}$$

$$\Delta^\circ \bar{\ell}(\theta)(\bar{O}_t, \bar{Z}_t) := \bar{\ell}(\theta)(\bar{O}_t, \bar{Z}_t) - \bar{\ell}(\theta^\circ)(\bar{O}_t, \bar{Z}_t)$$

($\bar{\ell}(\theta)$ is defined in (2)). Fix arbitrarily $a > 0$. An algebraic decomposition at the heart of all studies of the Super Learner [see, *e.g*, Dudoit and van der Laan, 2005, van der Laan et al., 2007, Benkeser et al., 2018]) states that the excess risk of the Super Learner (that is, $\widetilde{H}_{\widehat{j}_t,t}$) can be bounded by $(1+2a)$ times the excess risk of the oracle (that is, $\widetilde{H}_{\widetilde{j}_t,t}$), plus some remainder terms:

$$\begin{aligned}\widetilde{H}_{\widehat{j}_t,t} &\leq (1+2a)\widetilde{H}_{\widetilde{j}_t,t} + A_{\widehat{j}_t,t}(a) + B_{\widetilde{j}_t,t}(a) \\ &\leq (1+2a)\widetilde{H}_{\widetilde{j}_t,t} + \max_{j\in[\![J]\!]}\{A_{j,t}(a)\} + \max_{j\in[\![J]\!]}\{B_{j,t}(a)\} \end{aligned} \tag{22}$$

where

$$A_{j,t}(a) := (1+a)\left(\widetilde{H}_{j,t} - \widehat{H}_{j,t}\right) - a\widetilde{H}_{j,t} \quad \text{and} \quad B_{j,t}(a) := (1+a)\left(\widehat{H}_{j,t} - \widetilde{H}_{j,t}\right) - a\widetilde{H}_{j,t}.$$

The first terms in the definitions of $A_{j,t}(a)$ and $B_{j,t}$ equal $\pm(1+a)$ times

$$\frac{1}{t}\sum_{\tau=1}^{t}\left(\Delta^\circ\bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) - \mathbb{E}\left[\Delta^\circ\bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau)\middle|\bar{Z}_\tau, F_{\tau-1}\right]\right),$$

that is as the average of the t first terms of a martingale difference sequence. As for the shared second term in the definitions of $A_{j,t}(a)$ and $B_{j,t}$, it satisfies $-a\widetilde{H}_{j,t} \leq 0$. The second step of the proof consists in exploiting two so-called Bernstein's inequalities to control the probabilities $\mathbb{P}[A_{j,t}(a) \geq x]$ and $\mathbb{P}[B_{j,t}(a) \geq x]$ for $x \geq 0$.

Step 2: Bounding positive deviations of $A_{j,t}(a)$ and $B_{j,t}(a)$. Set arbitrarily two integers $N, N' \geq 2$ and a real number $x \geq 0$. The analysis of $\mathbb{P}[B_{j,t}(a) \geq x]$ is exactly the same as that of $\mathbb{P}[A_{j,t}(a) \geq x]$, so we present only the latter. The key to the analysis is a so-called stratification argument inspired by Cesa-Bianchi and Gentile [2008].

For every $j \in [\![J]\!]$ and $t \geq 1$, recall the definitions (15) and (16) of $\mathrm{var}_{j,t}$ and $\widetilde{\mathrm{var}}_{j,t}$. On the one hand, by **A4** and because the functions of a real variable $u \mapsto u^2$ and $u \mapsto u^\beta$ are respectively

convex and concave, it holds that

$$\begin{aligned}\widetilde{\mathrm{var}}_{j,t} &\leq \frac{1}{t}\sum_{\tau=1}^{t}\mathbb{E}\left[\left(\Delta^{\circ}\bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau,\bar{Z}_\tau)\right)^2\Big|\bar{Z}_\tau,F_{\tau-1}\right]\\ &\leq \gamma\left(\frac{1}{t}\sum_{\tau=1}^{t}\mathbb{E}\left[\Delta^{\circ}\bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau,\bar{Z}_\tau)\big|\bar{Z}_\tau,F_{\tau-1}\right]\right)^{\beta}=\gamma\left(\widetilde{H}_{j,t}\right)^{\beta}\end{aligned} \tag{23}$$

almost surely. Moreover, it also holds that $\widetilde{\mathrm{var}}_{j,t}\leq v_2$ almost surely by Theorem 3. The previous upper bound and (23) play a key role in the first version of Step 2 (Step 2 (v1)) presented below. On the other hand, by **A2**, **A4**, and because the function $u\mapsto u^{\beta}$ is concave it holds almost surely that, for all $\tau\in[\![t]\!]$,

$$\begin{aligned}\mathrm{var}_{j,\tau} &\leq \frac{1}{|\mathcal{A}|}\sum_{\alpha\in\mathcal{A}}\mathbb{E}\left[\left(\Delta^{\circ}\ell(\theta_{j,\tau-1})(O_{\alpha,\tau},Z_{\alpha,\tau})\right)^2\Big|Z_{\alpha,\tau},F_{\tau-1}\right]\\ &\leq \gamma\left(\mathbb{E}\left[\Delta^{\circ}\bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau,\bar{Z}_\tau)\big|\bar{Z}_\tau,F_{\tau-1}\right]\right)^{\beta}.\end{aligned}$$

Consequently if $\widetilde{H}_{j,t}\leq B$ (an inequality that holds almost surely when $B=b_1$, by **A3**), then it also holds that

$$B\geq\widetilde{H}_{j,t}=\frac{1}{t}\sum_{\tau=1}^{t}\mathbb{E}\left[\Delta^{\circ}\bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau,\bar{Z}_\tau)\big|\bar{Z}_\tau,F_{\tau-1}\right]\geq\frac{1}{t}\sum_{\tau=1}^{t}(\mathrm{var}_{j,\tau}/\gamma)^{1/\beta}.$$

In summary we will use that, for any $B>0$,

$$\mathbf{1}\left\{\widetilde{H}_{j,t}\leq B\right\}\leq\mathbf{1}\left\{\max_{1\leq\tau\leq t}\{\mathrm{var}_{j,\tau}\}\leq\gamma(tB)^{\beta}\right\}=\mathbf{1}\{\widetilde{\mathcal{F}}_{\gamma(tB)^{\beta}}\} \tag{24}$$

($\widetilde{\mathcal{F}}_V$ is defined for any $V>0$ in (17)). The upper bound $\widetilde{H}_{j,t}\leq b_1$ and (24) play a key role in the second version of Step 2 (Step 2 (v2)) presented below.

Step 2 (v1). Set $v_2^{(-1)}:=0$ and, for all $i\in[\![N-1]\!]$, $v_2^{(i)}:=2^{i+1-N}\times v_2$. In view of (23) and since $\widetilde{\mathrm{var}}_{j,t}\in\cup_{i=0}^{N}[v_2^{(i-1)},v_2^{(i)}]$ almost surely, it holds that

$$\begin{aligned}\mathbb{P}\left[A_{j,t}(a)\geq x\right]&=\mathbb{P}\left[\widetilde{H}_{j,t}-\widehat{H}_{j,t}\geq\frac{1}{1+a}\left(x+a\widetilde{H}_{j,t}\right)\right]\\ &\leq\mathbb{P}\left[\widetilde{H}_{j,t}-\widehat{H}_{j,t}\geq\frac{1}{1+a}\left(x+a(\widetilde{\mathrm{var}}_{j,t}/\gamma)^{1/\beta}\right)\right]\end{aligned}$$

$$\leq \sum_{i=0}^{N-1} \mathbb{P}\left[\widetilde{H}_{j,t} - \widehat{H}_{j,t} \geq \frac{1}{1+a}\left(x + a(\widetilde{\mathrm{var}}_{j,t}/\gamma)^{1/\beta}\right), \widetilde{\mathrm{var}}_{j,t} \in [v_2^{(i-1)}, v_2^{(i)}]\right]$$
$$\leq \sum_{i=0}^{N-1} \mathbb{P}\left[\widetilde{H}_{j,t} - \widehat{H}_{j,t} \geq \frac{1}{1+a}\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right), \widetilde{\mathrm{var}}_{j,t} \leq v_2^{(i)}\right]. \tag{25}$$

Note that $(\widetilde{H}_{j,t} - \widehat{H}_{j,t})_{t\geq 1}$ is a martingale adapted to the filtration $(\sigma(F_t, \sigma(\bar{Z}_{t+1})))_{t\geq 1}$. By **A3** and the Fan-Grama-Liu concentration inequality for martingales [Theorem 3.10 in Bercu et al., 2015], (25) implies

$$\mathbb{P}\left[A_{j,t}(a) \geq x\right] \leq \sum_{i=0}^{N-1} \exp\left(-\frac{1}{2}\frac{tD_i(x)}{(1+a)^2}\right), \tag{26}$$

where, for all $i \in [\![N-1]\!]$,

$$D_i(x) := \frac{\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right)^2}{v_2^{(i)} + \frac{1}{3}\frac{b_2}{1+a}\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right)}.$$

Set arbitrarily $i \in [\![N-1]\!] \cup \{0\}$ and define $x_i := 3(1+a)v_2^{(i)}/b_2 - a(v_2^{(i-1)}/\gamma)^{1/\beta}$.

- If $x \leq x_i$, then $v_2^{(i)} \geq (x + a(v_2^{(i-1)}/\gamma)^{1/\beta}) \times b_2/(3(1+a))$ hence

$$D_i(x) \geq \frac{\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right)^2}{2v_2^{(i)}} = \frac{\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right)^{2-\beta}}{2v_2^{(i)}/\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right)^{\beta}}$$
$$\geq \frac{x^{2-\beta}}{2v_2^{(i)}/\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right)^{\beta}}. \tag{27}$$

  If $i \neq 0$, then (27) entails

$$D_i(x) \geq \frac{x^{2-\beta}}{2\gamma v_2^{(i)}/(a^{\beta} v_2^{(i-1)})} = \frac{x^{2-\beta}}{4\gamma/a^{\beta}}. \tag{28}$$

  If $i = 0$, then (28) is also met if and only if $x \geq \underline{x}(a, N)$, where $\underline{x}(a, N)$ is defined in the theorem.

- Moreover if $x \geq x_i$, then $v_2^{(i)} \leq (x + a(v_2^{(i-1)}/\gamma)^{1/\beta}) \times b_2/(3(1+a))$ hence

$$D_i(x) \geq \frac{\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right)^2}{\frac{2}{3}\frac{b_2}{1+a}\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta}\right)} = \frac{x + a(v_2^{(i-1)}/\gamma)^{1/\beta}}{\frac{2}{3}\frac{b_2}{1+a}} \geq \frac{x}{\frac{2}{3}\frac{b_2}{1+a}}. \tag{29}$$

Therefore, in light of (26), (28), (29) and the definitions of $C_1(a), C_2(a)$ given in the theorem, for all $x \geq \underline{x}(a, N)$, it holds that

$$\begin{aligned}\mathbb{P}[A_{j,t}(a) \geq x] &\leq \sum_{i=0} N - 1\left[\mathbf{1}\{x \leq x_i\}\exp\left(-\frac{t \times (2x)^{2-\beta}}{C_1(a)}\right) + \mathbf{1}\{x \geq x_i\}\exp\left(-\frac{t \times (2x)}{C_2(a)}\right)\right] \\ &\leq N\left[\exp\left(-\frac{t \times (2x)^{2-\beta}}{C_1(a)}\right) + \exp\left(-\frac{t \times (2x)}{C_2(a)}\right)\right]. \end{aligned} \tag{30}$$

Step 2 (v2). This step is very similar to Step 2 (v1). Set $b_1^{(-1)} := 0$ and, for all $i \in [\![N' - 1]\!]$, $b_1^{(i)} := 2^{i+1-N'} \times b_1$. In view of (24) and since $\widetilde{H}_{j,t} \in \cup_{i=0}^{N'-1}[b_1^{(i-1)}, b_1^{(i)}]$ almost surely, it holds that

$$\begin{aligned}\mathbb{P}\left[A_{j,t}(a) \geq x\right] &= \mathbb{P}\left[\widetilde{H}_{j,t} - \widehat{H}_{j,t} \geq \frac{1}{1+a}\left(x + a\widetilde{H}_{j,t}\right)\right] \\ &\leq \sum_{i=0}^{N'-1} \mathbb{P}\left[\widetilde{H}_{j,t} - \widehat{H}_{j,t} \geq \frac{1}{1+a}\left(x + a\widetilde{H}_{j,t}\right), \widetilde{H}_{j,t} \in [b_1^{(i-1)}, b_1^{(i)}]\right] \\ &\leq \sum_{i=0}^{N'-1} \mathbb{P}\left[\widetilde{H}_{j,t} - \widehat{H}_{j,t} \geq \frac{1}{1+a}\left(x + ab_1^{(i-1)}\right), \widetilde{H}_{j,t} \leq b_1^{(i)}\right] \\ &\leq \sum_{i=0}^{N'-1} \mathbb{P}\left[\widetilde{H}_{j,t} - \widehat{H}_{j,t} \geq \frac{1}{1+a}\left(x + ab_1^{(i-1)}\right), \widetilde{\mathcal{F}}_{\gamma(tb_1^{(i)})^\beta}\right]. \end{aligned} \tag{31}$$

By **A3** and **A4**, Theorem 3 applies and (31) yields

$$\mathbb{P}\left[A_{j,t}(a) \geq x\right] \leq \sum_{i=0}^{N'-1} \exp\left(2 - \frac{|\mathcal{A}|/\deg(\mathcal{G})}{(1+a)^2} D_i'(x)\right), \tag{32}$$

where, for all $i \in [\![N' - 1]\!]$,

$$D_i'(x) := \frac{\left(x + ab_1^{(i-1)}\right)^2}{32e^2\gamma(tb_1^{(i)})^\beta + \frac{15eb_2}{1+a}\left(x + ab_1^{(i-1)}\right)}.$$

Set arbitrarily $i \in [\![N'-1]\!] \cup \{0\}$ and define $x'_i := 32e(1+a)\gamma(tb_1^{(i)})^\beta/(15b_2) - ab_1^{(i-1)}$.

- If $x \le x'_i$, then $32e^2\gamma(tb_1^{(i)})^\beta \ge (x + ab_1^{(i-1)}) \times 15eb_2/(1+a)$ hence

$$D'_i(x) \ge \frac{\left(x + ab_1^{(i-1)}\right)^2}{64e^2\gamma(tb_1^{(i)})^\beta} = \frac{\left(x + ab_1^{(i-1)}\right)^{2-\beta}}{64e^2\gamma(tb_1^{(i)})^\beta / \left(x + ab_1^{(i-1)}\right)^\beta} \ge \frac{x^{2-\beta}}{64e^2\gamma(tb_1^{(i)})^\beta / \left(x + ab_1^{(i-1)}\right)^\beta}. \tag{33}$$

  If $i \neq 0$, then (33) entails

$$D'_i(x) \ge \frac{x^{2-\beta}}{64e^2\gamma(tb_1^{(i)})^\beta/(ab_1^{(i-1)})^\beta} = \frac{x^{2-\beta}}{64e^2\gamma(2t/a)^\beta}. \tag{34}$$

  If $i = 0$, then (34) is also met if and only if $x \ge \underline{x}'(a, N')$, where $\underline{x}'(a, N')$ is defined in the theorem.

- Moreover if $x \ge x'_i$, then $32e^2\gamma(tb_1^{(i)})^\beta \le (x + ab_1^{(i-1)}) \times 15eb_2/(1+a)$ hence

$$D'_i(x) \ge \frac{\left(x + ab_1^{(i-1)}\right)^2}{\frac{30eb_2}{1+a}\left(x + ab_1^{(i-1)}\right)} = \frac{x + ab_1^{(i-1)}}{\frac{30eb_2}{1+a}} \ge \frac{x}{\frac{30eb_2}{1+a}}. \tag{35}$$

Therefore, in light of (32), (34), (35) and the definitions of $C'_1(a), C'_2(a)$ given in the theorem, for all $x \ge \underline{x}'(a, N')$, it holds that

$$\begin{aligned}
\mathbb{P}[A_{j,t}(a) \ge x] &\le \sum_{i=0}^{N'-1} \left[1\{x \le x'_i\} \exp\left(2 - \frac{[|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))] \times (2x)^{2-\beta}}{C'_1(a)}\right) \right. \\
&\qquad \left. + \mathbf{1}\{x \ge x'_i\} \exp\left(2 - \frac{[|\mathcal{A}|/\deg(\mathcal{G})] \times (2x)}{C'_2(a)}\right) \right] \\
&\le N' \left[\exp\left(2 - \frac{[|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))] \times (2x)^{2-\beta}}{C'_1(a)}\right) \right. \\
&\qquad \left. + \exp\left(2 - \frac{[|\mathcal{A}|/\deg(\mathcal{G})] \times (2x)}{C'_2(a)}\right) \right].
\end{aligned} \tag{36}$$

Step 3: end of the proof. In view of (22), a union bound implies that

$$\mathbb{P}\left[\widetilde{H}_{\widehat{j}_t,t} - (1+2a)\widetilde{H}_{\widetilde{j}_t,t} \geq x\right] \leq \mathbb{P}\left[\max_{j\in[\![J]\!]}\{A_{j,t}(a)\} + \max_{j\in[\![J]\!]}\{B_{j,t}(a)\} \geq x\right]$$
$$\leq \sum_{j=1}^{J}\left(\mathbb{P}\left[A_{j,t}(a) \geq x/2\right] + P\left[B_{j,t}(a) \geq x/2\right]\right).$$

For all $x \geq \underline{x}(a,N)$, (7) follows from (30) and the above inequality; for all $x \geq \underline{x}'(a,N')$, (8) follows from (36) and the above inequality. This completes the proof of Theorem 1. □

B.2 Proof of Corollary 2

Corollary 2 follows from the straightforward application, twice, of the next technical lemma, based on (7) on the one hand and on (8) on the other hand.

Lemma 4. *Let $a,b,c>0$, $\beta \in]0,1]$ be some constants and $(\underline{x}(N))_{N\geq 2}$ be a sequence of positive numbers that decreases to 0. Let U be a real valued random variable such that $E[|U|]<\infty$ and, for all integer $N\geq 2$ and all $x \geq \underline{x}(N) > 0$,*

$$\mathbb{P}[U \geq x] \leq aN\left[\exp(-x^{2-\beta}/b) + \exp(-x/c)\right]. \tag{37}$$

If $N \geq \min\{n \geq 2 : \underline{x}(n) \leq b^{1/(2-\beta)}, \log(an) \geq 1\}$, then

$$\mathbb{E}[U] \leq 3(b\log(aN))^{1/(2-\beta)} + 2c\log(aN). \tag{38}$$

Proof of Lemma 4. It is well known that

$$\mathbb{E}[U] \leq \mathbb{E}[U\mathbf{1}\{U\geq 0\}] = \int_0^\infty \mathbb{P}[U\mathbf{1}\{U\geq 0\} \geq x]dx = \int_0^\infty \mathbb{P}[U\geq x]dx.$$

Therefore, for any $N \geq 2$,

$$\mathbb{E}[U] \leq \int_0^\infty \left(\mathbf{1}\{x < \underline{x}(N)\} + \mathbf{1}\{x \geq \underline{x}(N)\}aN\left[\exp(-x^{2-\beta}/b) + \exp(-x/c)\right]\right)dx$$
$$\leq \underline{x}(N) + \int_0^\infty \min\{1, aN\exp(-x^{2-\beta}/b)\}dx + \int_0^\infty \min\{1, aN\exp(-x/c)\}dx. \tag{39}$$

Let us denote by $\mathcal{L}$ and $\mathcal{R}$ the above left-hand side and right-hand side integrals. Choose $N \geq \min\{n \geq 2 : \underline{x}(n) \leq b^{1/(2-\beta)}, \log(an) \geq 1\}$.

Bounding $\mathcal{L}$. Let $x_{\mathcal{L}}$ be chosen so that $aN\exp(-x_{\mathcal{L}}^{2-\beta}/b) = 1$, *i.e.*, $x_{\mathcal{L}} := (b\log(aN))^{1/(2-\beta)}$. Now, thanks to the change of variable $u = x^{2-\beta}/b$ and because $u \mapsto u^{1/(2-\beta)-1}$ is nonincreasing,

$$\begin{aligned}
\mathcal{L} &= x_{\mathcal{L}} + aN\int_{x_{\mathcal{L}}}^{\infty} \exp(-x^{2-\beta}/b)dx \\
&= x_{\mathcal{L}} + aNb^{1/(2-\beta)}\int_{\log(aN)}^{\infty} \exp(-u)u^{1/(2-\beta)-1}du \\
&\leq x_{\mathcal{L}} + \frac{aN(b\log(aN))^{1/(2-\beta)}}{\log(aN)}\int_{0}^{\infty} \exp(-u)du \\
&= x_{\mathcal{L}}(1 + 1/\log(aN)) \leq 2(b\log(aN))^{1/(2-\beta)}. \qquad (40)
\end{aligned}$$

Bounding $\mathcal{R}$. Let $x_{\mathcal{R}}$ be chosen so that $aN\exp(-x_{\mathcal{R}}/c) = 1$, *i.e.*, $x_{\mathcal{R}} := c\log(aN)$. It is readily seen that

$$\mathcal{R} = x_{\mathcal{R}} + aN\int_{x_{\mathcal{R}}}^{\infty} \exp(-x/c)dx = x_{\mathcal{R}} + acN\exp(-x_{\mathcal{R}}/c) = c(1 + \log(aN)). \qquad (41)$$

In view of (39), (40), (41), and by choice of N, we obtain

$$\begin{aligned}
\mathbb{E}[U] &\leq b^{1/(2-\beta)} + 2(b\log(aN))^{1/(2-\beta)} + c(1 + \log(aN)) \\
&\leq 3(b\log(aN))^{1/(2-\beta)} + 2c\log(aN).
\end{aligned}$$

This completes the proof. □

Set $t \geq 1$ and $a \in]0, 1]$. In view of (7), Lemma 4 yields (9) under the sufficient condition that $N \geq 2$ also satisfy (10). Moreover, in view of (8), Lemma 4 also yields (11) under the sufficient condition that $N \geq 2$ also satisfy (12). This completes the proof of the corollary. □

B.3 Proof of Theorem 3

The proof of Theorem 3 hinges on a Bernstein-like concentration inequality for sums of partly dependent random variables shown by Janson [2004, Theorem 2.3]. Janson emphasizes that his

theorem uses the independence of suitable (large) subsets of $(\zeta_\alpha)_{\alpha\in\mathcal{A}}$, not any other information on the dependencies, so that the result must be suboptimal when the dependencies that exist are weak. We recall the theorem for completeness.

Theorem 5 (Janson [2004]). *Let $(\zeta_\alpha)_{\alpha\in\mathcal{A}}$ be a collection of random variables with dependency graph $\mathcal{G}$ such that $\zeta_\alpha - \mathbb{E}[\zeta_\alpha] \le B$ for some $B > 0$ and all $\alpha \in \mathcal{A}$. Define $\mathcal{Z} := |\mathcal{A}|^{-1}\sum_{\alpha\in\mathcal{A}}\zeta_\alpha$ and $V := |\mathcal{A}|^{-1}\sum_{\alpha\in\mathcal{A}}\mathbb{V}\mathrm{ar}[\zeta_\alpha]$. Then, for all $x \ge 0$,*

$$\mathbb{P}\left[\mathcal{Z} - \mathbb{E}[\mathcal{Z}] \ge x\right] \le \exp\left(-\frac{|\mathcal{A}|V}{B^2\deg(\mathcal{G})}h\left(\frac{4Bx}{5V}\right)\right), \tag{42}$$

where $h : u \mapsto (1+u)\log(1+u) - u$.

Note that (18) from Theorem 3 also writes as

$$\mathbb{P}\left[|\widehat{H}_{j,t} - \widetilde{H}_{j,t}| \ge x, \widetilde{\mathcal{F}}_V\right] \le \exp\left(2 - \frac{[|\mathcal{A}|/\deg(\mathcal{G})]x^2}{32e^2V + 15eb_2x}\right).$$

Following the line of proof of the Rosenthal inequality by Petrov [1995, page 59] (see also the proof of Theorem 5.2 in [Baraud, 2000]), we use (42) to control $\mathbb{E}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|^p]$ hence $\mathbb{E}[|\widehat{H}_{j,t} - \widetilde{H}_{j,t}|^p]$ (by convexity) for all $p \ge 2$. The bound (18) follows as in [Dedecker, 2001, proof of Corollary 3(b)], a method inspired by the proof of Theorem 6 in [Doukhan et al., 1984].

We first prove this corollary of Theorem 5. The constants are in no way optimal.

Corollary 6. *In the context of Theorem 5, for all $p \ge 2$,*

$$\mathbb{E}\left[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|^p\right] \le \frac{3\pi}{2}\left[\left(\frac{15B\deg(\mathcal{G})}{2|\mathcal{A}|}\right)^p p^p + \left(\frac{32V\deg(\mathcal{G})}{|\mathcal{A}|}\right)^{p/2} p^{p/2}\right]. \tag{43}$$

Proof of Corollary 6. Fix arbitrarily $p \ge 2$. It is well known that $\mathbb{E}[U^p] = \int_0^\infty ps^{p-1}\mathbb{P}[U \ge s]ds$ for any nonnegative random variable U. Let $r > 0$ be a constant that we will carefully choose later on. Set arbitrarily $s \ge 0$, define $m := s/r$, and introduce

$$\tilde{\mathcal{Z}}_m := |\mathcal{A}|^{-1}\sum_{\alpha\in\mathcal{A}}(\zeta_\alpha - \mathbb{E}[\zeta_\alpha])\mathbf{1}\{|\zeta_\alpha - \mathbb{E}[\zeta_\alpha]| < m\}.$$

It holds that

$$\begin{aligned}\mathbb{P}(|\mathcal{Z}-\mathbb{E}[\mathcal{Z}]|\geq s) &\leq \mathbb{P}[\mathcal{Z}-\mathbb{E}[\mathcal{Z}]\neq\tilde{\mathcal{Z}}_m]+\mathbb{P}[|\mathcal{Z}-\mathbb{E}[\mathcal{Z}]|\geq s,\mathcal{Z}-\mathbb{E}[\mathcal{Z}]=\tilde{\mathcal{Z}}_m]\\ &\leq \mathbb{P}[r\max_{\alpha\in\mathcal{A}}|\zeta_\alpha-\mathbb{E}[\zeta_\alpha]|\geq s]+\mathbb{P}[|\mathcal{Z}-\mathbb{E}[\mathcal{Z}]|\geq s,\mathcal{Z}-\mathbb{E}[\mathcal{Z}]=\tilde{\mathcal{Z}}_m]\\ &\leq \mathbb{P}[r\max_{\alpha\in\mathcal{A}}|\zeta_\alpha-\mathbb{E}[\zeta_\alpha]|\geq s]+\mathbb{P}[|\tilde{\mathcal{Z}}_m-\mathbb{E}[\tilde{\mathcal{Z}}_m]|\geq s-\mathbb{E}[\tilde{\mathcal{Z}}_m]]\end{aligned}$$

hence

$$\mathbb{E}[|\mathcal{Z}-\mathbb{E}[\mathcal{Z}]|^p]\leq r^p\mathbb{E}[\max_{\alpha\in\mathcal{A}}|\zeta_\alpha-\mathbb{E}[\zeta_\alpha]|^p]+\int_0^\infty ps^{p-1}\mathbb{P}[|\tilde{\mathcal{Z}}_m-\mathbb{E}(\tilde{\mathcal{Z}}_m)]\geq s-\mathbb{E}[\tilde{\mathcal{Z}}_m]]ds. \tag{44}$$

We now note that

$$\begin{aligned}|\mathbb{E}[\tilde{\mathcal{Z}}_m]| &= |\mathbb{E}[\tilde{\mathcal{Z}}_m-(\mathcal{Z}-\mathbb{E}[\mathcal{Z}])]|\\ &= |\mathcal{A}|^{-1}\left|\mathbb{E}\left[\sum_{\alpha\in\mathcal{A}}(\zeta_\alpha-\mathbb{E}[\zeta_\alpha])\mathbf{1}\{|\zeta_\alpha-\mathbb{E}[\zeta_\alpha]|\geq m\}\right]\right|\\ &\leq (m|\mathcal{A}|)^{-1}\sum_{\alpha\in\mathcal{A}}\mathbb{V}\mathrm{ar}[\zeta_\alpha]=V/m.\end{aligned}$$

Therefore if $s\geq s_0:=\sqrt{2rV}$, then $s/2\geq V/(s/r)=V/m$ hence $s-|\mathbb{E}[\tilde{\mathcal{Z}}_m]|\geq s/2$. In light of (42) and (44), the rightmost term in (44), say I_p, satisfies

$$\begin{aligned}I_p &\leq \int_0^{s_0}ps^{p-1}ds+\int_{s_0}^\infty ps^{p-1}\mathbb{P}[|\tilde{\mathcal{Z}}_m-\mathbb{E}[\tilde{\mathcal{Z}}_m]|\geq s/2]ds\\ &\leq s_0^p+2\int_{s_0}^\infty ps^{p-1}\exp\left(-\frac{|\mathcal{A}|\tilde{V}}{4m^2\deg(\mathcal{G})}h\left(\frac{8ms/2}{5\tilde{V}}\right)\right)ds,\end{aligned} \tag{45}$$

where

$$\begin{aligned}\tilde{V} &:= |\mathcal{A}|^{-1}\sum_{\alpha\in\mathcal{A}}\mathbb{V}\mathrm{ar}[(\zeta_\alpha-\mathbb{E}[\zeta_\alpha])\mathbf{1}\{|\zeta_\alpha-\mathbb{E}[\zeta_\alpha]|\leq m\}]\\ &\leq |\mathcal{A}|^{-1}\sum_{\alpha\in\mathcal{A}}\mathbb{E}[(\zeta_\alpha-\mathbb{E}[\zeta_\alpha])^2\mathbf{1}\{|\zeta_\alpha-\mathbb{E}[\zeta_\alpha]|\leq m\}]\\ &\leq |\mathcal{A}|^{-1}\sum_{\alpha\in\mathcal{A}}\mathbb{E}[(\zeta_\alpha-\mathbb{E}[\zeta_\alpha])^2]=V.\end{aligned}$$

Because $h(u) \geq \frac{u}{2}\log(1+u)$ for all $u \geq 0$, (45) yields

$$\begin{aligned} I_p &\leq s_0^p + 2\int_{s_0}^{\infty} ps^{p-1} \exp\left(-\frac{|\mathcal{A}|s}{10m\deg(\mathcal{G})}\log\left(1+\frac{4ms}{5\tilde{V}}\right)\right) ds \\ &= s_0^p + 2\int_{s_0}^{\infty} ps^{p-1} \exp\left(-\frac{|\mathcal{A}|r}{10\deg(\mathcal{G})}\log\left(1+\frac{4s^2}{5r\tilde{V}}\right)\right) ds. \end{aligned} \tag{46}$$

If $u := s/(5r\tilde{V}/4)^{1/2}$, then $s^{p-1} \leq (5r\tilde{V}/4)^{(p-1)/2}(1+u^2)^{(p-1)/2}$. A change of variable and the bound $\tilde{V} \leq V$ thus imply that the rightmost term in (46) is smaller than

$$2p\left(\frac{5rV}{4}\right)^{p/2} \int_0^{\infty} (1+u^2)^{(p-1)/2 - r|\mathcal{A}|/(10\deg(\mathcal{G}))} du.$$

We now choose $r := 5(p+1)\deg(\mathcal{G})/|\mathcal{A}|$ to guarantee the convergence of the above integral, to $\pi/2$, and conclude that (44) and (46) imply

$$\begin{aligned} \mathbb{E}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|^p] &\leq r^p \mathbb{E}[\max_{\alpha\in\mathcal{A}} |\zeta_\alpha - \mathbb{E}[\zeta_\alpha]|^p] + \pi(rV)^{p/2}(2^{p/2} + p(5/4)^{p/2}) \\ &\leq (rB)^p + \pi(p+1)(2rV)^{p/2}. \end{aligned} \tag{47}$$

Finally, since $(p+1)/p \leq 3/2$ and $p^{2/p} \leq e^{2/e} \approx 2.61$, we can simplify (47) to (43), thus completing the proof of Corollary 6. □

Fix arbitrarily $j \in [\![J]\!]$, $t \geq 1$, $V > 0$, and $p \geq 2$. To save space introduce, for each $\tau \in [\![t]\!]$,

$$\mathcal{Z}_{j,\tau} := \Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) - \mathbb{E}\left[\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) \middle| \bar{Z}_\tau, F_{\tau-1}\right].$$

In view of **A1**, **A2**, **A3** and **A5**, Corollary 6 applies and guarantees that almost surely, for all $\tau \in [\![t]\!]$,

$$\begin{aligned} &\mathbb{E}\left[|\mathcal{Z}_{j,\tau}|^p \middle| \bar{Z}_\tau, F_{\tau-1}\right] \mathbf{1}\{\mathrm{var}_{j,\tau} \leq V\} \\ &\quad \leq \frac{3\pi}{2}\left[\left(\frac{15b_2\deg(\mathcal{G})}{2|\mathcal{A}|}\right)^p p^p + \left(\frac{32V\deg(\mathcal{G})}{|\mathcal{A}|}\right)^{p/2} p^{p/2}\right] \mathbf{1}\{\mathrm{var}_{j,\tau} \leq V\} \\ &\quad \leq \frac{3\pi}{2}\left[\left(\frac{15b_2\deg(\mathcal{G})}{2|\mathcal{A}|}\right)^p p^p + \left(\frac{32V\deg(\mathcal{G})}{|\mathcal{A}|}\right)^{p/2} p^{p/2}\right]. \end{aligned} \tag{48}$$

It is now easy to show that $\widetilde{\mathrm{var}}_{t,j} \le v_2$ almost surely (see (6) for the definition of v_2). By **A5**, it holds that $\mathrm{var}_{j,\tau} \le v_1$ almost surely for each $\tau \in [\![t]\!]$, hence

$$\widetilde{\mathrm{var}}_{j,t} = \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[(\mathcal{Z}_{j,\tau})^2 \middle| \bar{Z}_\tau, F_{\tau-1}\right] = \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[(\mathcal{Z}_{j,\tau})^2 \middle| \bar{Z}_\tau, F_{\tau-1}\right] \mathbf{1}\{\mathrm{var}_{j,\tau} \le v_1\} \le v_2$$

because of (48) with $p = 2$.

We now turn to the proof of (18). In view of (48), by convexity of $u \mapsto |u|^p$, it holds that

$$\begin{aligned}
\mathbb{E}\left[\left|\widehat{H}_{j,t} - \widetilde{H}_{j,t}\right|^p \mathbf{1}\{\widetilde{\mathcal{F}}_V\}\right] &\le \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[|\mathcal{Z}_{j,\tau}|^p \mathbf{1}\{\widetilde{\mathcal{F}}_V\}\right] \\
&\le \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[|\mathcal{Z}_{j,\tau}|^p \mathbf{1}\{\mathrm{var}_{j,\tau} \le V\}\right] \\
&= \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[\mathbb{E}\left[|\mathcal{Z}_{j,\tau}|^p \middle| \bar{Z}_\tau, F_{\tau-1}\right] \mathbf{1}\{\mathrm{var}_{j,\tau} \le V\}\right] \\
&\le \frac{3\pi}{2}\left[\left(\frac{15 b_2 \deg(\mathcal{G})}{2|\mathcal{A}|}\right)^p p^p + \left(\frac{32 V \deg(\mathcal{G})}{|\mathcal{A}|}\right)^{p/2} p^{p/2}\right]. \qquad (49)
\end{aligned}$$

Therefore Markov's inequality implies that, for all $x > 0$,

$$\begin{aligned}
\mathbb{P}\left[\left|\widehat{H}_{j,t} - \widetilde{H}_{j,t}\right| \ge x, \widetilde{\mathcal{F}}_V\right] &\le \mathbb{E}\left[x^{-p}\left|\widehat{H}_{j,t} - \widetilde{H}_{j,t}\right|^p \mathbf{1}\{\widetilde{\mathcal{F}}_V\}\right] \\
&\le \frac{3\pi}{2}\left(\frac{15 b_2 \deg(\mathcal{G}) p/2 + \sqrt{32|\mathcal{A}|V\deg(\mathcal{G})p}}{x|\mathcal{A}|}\right)^p. \qquad (50)
\end{aligned}$$

By the technical Lemma 7, there exists $p_x > 0$ such that

$$\begin{aligned}
x|\mathcal{A}| &= 15 e b_2 \deg(\mathcal{G}) p_x/2 + \sqrt{32 e^2 |\mathcal{A}| V \deg(\mathcal{G}) p_x}, \quad \text{and} \\
p_x \ge q_x &:= (x|\mathcal{A}|)^2 \left(32 e^2 |\mathcal{A}| V \deg(\mathcal{G}) + 15 e b_2 \deg(\mathcal{G}) x |\mathcal{A}|\right)^{-1} \\
&= x^2 |\mathcal{A}| \left(32 e^2 V \deg(\mathcal{G}) + 15 e b_2 \deg(\mathcal{G}) x\right)^{-1}.
\end{aligned}$$

If $q_x \ge 2$, then p_x is a valid choice for p in (50). This choice yields the inequality

$$\mathbb{P}\left[\left|\widehat{H}_{j,t} - \widetilde{H}_{j,t}\right| \ge x, \widetilde{\mathcal{F}}_V\right] \le \frac{3\pi}{2}\exp(-p_x) \le \frac{3\pi}{2}\exp(-q_x) \le \exp\left(2 - q_x\right).$$

Otherwise, $\mathbb{P}[|\widehat{H}_{j,t} - \widetilde{H}_{j,t}| \geq x, \widetilde{\mathcal{F}}_V] \leq \exp(2 - q_x)$ holds trivially. This completes the proof of Theorem 3. □

Lemma 7. *For any $a, b, c > 0$, there exists $p > 0$ such that $c = b\sqrt{p} + ap$. Moreover, $c^2 \leq (b^2 + 2ac)p$.*

Proof of Lemma 7. The quadratic equation $c = bX + aX^2$ has a positive solution, so there does exist $p > 0$ such that $c = b\sqrt{p} + ap$. Moreover, $c^2/p = b^2 + 2ab\sqrt{p} + a^2p$ on the one hand and $2ac = 2ab\sqrt{p} + 2a^2p \geq 2ab\sqrt{p} + a^2p$ on the other hand, implying that $c^2/p \leq b^2 + 2ac$. This completes the proof. □