Are we there yet? Advances in anytime-valid methods for hypothesis testing and prediction

Pérez-Ortiz, M.F.

Appendices

A. Appendix to Chapter 2

A.1. Computations

Proposition A.1.1. *Let $X \sim N(\gamma, I)$, and let $mS \sim W(m, I)$ be independent random variables. Let $LL' = S$ be the Cholesky decomposition of S, and let $M = \frac{1}{\sqrt{m}} L^{-1} X$. If $\mathbf{P}_{0,n}$ is the probability distribution under which $X \sim N(0, I)$, then, the likelihood $p^M_{\gamma,m}/p^M_{0,m}$ ratio is given by*

$$\frac{p^M_{\gamma,m}(M)}{p^M_{0,m}(M)} = \mathrm{e}^{-\frac{1}{2}\|\gamma\|^2} \int \mathrm{e}^{\langle \gamma, TA^{-1}M \rangle} \mathrm{d}\mathbf{P}_{m+1,I}(T)$$

where $A \in \mathcal{L}^+$ is the Cholesky factor $AA' = I + MM'$, and $\mathbf{P}^T_{m+1,I}$ is the probability distribution on $\mathcal{L}^+$ such that $TT' \sim W(m+1, I)$.

Proof. Let $\Sigma = \Lambda\Lambda'$ be the Cholesky decomposition of Σ. The density $p^X_{\gamma,\Lambda}$ of X with respect to the Lebesgue measure on $\mathbb{R}^d$ is

$$p^X_{\gamma,\Lambda}(X) = \frac{1}{(2\pi)^{d/2}\det(\Lambda)} \mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}X - \gamma)(\Lambda^{-1}X - \gamma)'\right),$$

where, for a square matrix A, we define $\mathrm{etr}(A)$ to be the exponential of the trace of A. Let $W = mS$. Then, the density $p^W_{\gamma,\Lambda}$ of W with respect to the Lebesgue measure on $\mathbb{R}^{d(d-1)/2}$ is

$$p^W_{\gamma,\Lambda}(W) = \frac{1}{2^{md/2}\Gamma_d(n/2)\det(\Lambda)^m} \det(S)^{(m-d-1)/2} \mathrm{etr}\left(-\frac{1}{2}(\Lambda\Lambda')^{-1}W\right).$$

Now, let $W = TT'$ be the Cholesky decomposition of W. We seek to compute the distribution of the random lower lower triangular matrix T. To this end, the change of variables $W \mapsto T$ is one-to-one, and has Jacobian determinant equal to $2^d \prod_{i=1}^d t_{ii}^{d-i+1}$. Consequently, the density $p^T_{\gamma,\Lambda}(T)$ of T with respect to the Lebesgue measure is

$$p^T_{\gamma,\Lambda}(T) = \frac{2^d}{2^{md/2}\Gamma_d(m/2)} \det(\Lambda^{-1}T)^m \mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)'\right) \prod_{i=1}^d t_{ii}^{-i}.$$

We recognize $\mathrm{d}\nu(T) = \prod_{i=1}^d t_{ii}^{-i} \mathrm{d}T$ to be a left Haar measure on $\mathcal{L}_+$, and consequently

$$\tilde{p}^T_{\gamma,\Lambda}(T) = \frac{2^d}{2^{md/2}\Gamma_d(m/2)} \det(\Lambda^{-1}T)^m \mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)'\right)$$

is the density of T with respect to $\mathrm{d}\nu(T)$. After this, the density $\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T)$ of the pair (X,T) with respect to $\mathrm{d}X \times \mathrm{d}\nu(T)$ is given by

$$\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T) = \frac{2^d}{K}\frac{\det(\Lambda^{-1}T)^m}{\det(\Lambda)}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}X-\gamma)(\Lambda^{-1}X-\gamma)'\right)$$

with $K = (2\pi)^{d/2}2^{md/2}\Gamma_d(n/2)$. The change of variables $(X,T) \mapsto (T^{-1}X,T)$ has Jacobian determinant equal to $\det(T)$. If $M = T^{-1}X$, then, the density $\tilde{p}^{M,T}_{\gamma,\Lambda}$ of (M,T) with respect to $\mathrm{d}M \times \mathrm{d}\nu(T)$ is given by

$$\tilde{p}^{M,T}_{\gamma,\Lambda}(M,T) = \frac{\det(\Lambda^{-1}T)^{m+1}}{K''}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}TM-\gamma)(\Lambda^{-1}TM-\gamma)'\right).$$

We now marginalize T to obtain the distribution of the maximal invariant M. Since the integral is with respect to the left Haar measure $\mathrm{d}\nu(T)$, we have that

$$\int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,\Lambda}(M,T)\mathrm{d}\nu(T) = \int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,I}(M,\Lambda^{-1}T)\mathrm{d}\nu(T) = \int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,I}(M,T)\mathrm{d}\nu(T),$$

and consequently,

$$\begin{aligned} p^{M}_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}TT' - \frac{1}{2}(TM-\gamma)(TM-\gamma)'\right)\mathrm{d}\nu(T) \\ &= \frac{2^d}{K}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}T(I+MM')T' + \gamma(TM)'\right)\mathrm{d}\nu(T). \end{aligned}$$

The matrix $I+MM'$ is positive definite and symmetric. It is then possible to perform its Cholesky decomposition $(I+MM') = AA'$. With this at hand, the previous display can be written as

$$p^{M}_{\gamma,\Lambda}(M) = \frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}}{K}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}(TA)(TA)' + \gamma(TM)'\right)\mathrm{d}\nu(T).$$

We now perform the change of variable $T \mapsto TA^{-1}$. To this end, notice that $\mathrm{d}\nu(A^{-1}) = \mathrm{d}\nu(T)\prod_{i=1}^d a_{ii}^{-(d-2i+1)}$, and consequently

$$\begin{aligned} p^{M}_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}TT' + \gamma(TA^{-1}M)'\right)\mathrm{d}\nu(T) \\ &= \frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)}\frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\mathbf{P}^T_{m+1}\left[\mathrm{e}^{\langle\gamma,TA^{-1}M\rangle}\right], \end{aligned}$$

so that that at $\gamma = 0$ the density $p^M_{0,\Lambda}(M)$ takes the form

$$p^M_{0,\Lambda}(M) = \frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)} \frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}},$$

and consequently the likelihood ratio is

$$\frac{p^M_{\gamma,\Lambda}(M)}{p^M_{0,\Lambda}(M)} = \mathrm{e}^{-\frac{1}{2}\|\gamma\|^2} \int \mathrm{e}^{\langle \gamma, TA^{-1}M\rangle} \mathrm{d}\mathbf{P}_{m+1}(T).$$

□

Remark A.1.2 (Numerical computation). Computing the optimal E-value is feasible numerically. We are interested in computing

$$\int \mathrm{e}^{\langle x, Ty\rangle} \mathrm{d}\mathbf{P}_{m+1}(T),$$

where T is a $\mathcal{L}^+$-valued random lower triangular matrix such that $TT' \sim W(m+1, I)$, and $x, y \in \mathbb{R}^d$. Define, for $i \geq j$, the numbers $a_{ij} = x_i y_j$. Then $\langle x, Ty\rangle = \sum_{i\geq j} a_{ij} T_{ij}$. By Bartlett's decomposition, the entries of the matrix T are independent and $T_{ii}^2 \sim \chi^2((m+1)-i+1)$, and $T_{ij} \sim N(0,1)$ for $i > j$. Hence, our target quantity satisfies

$$\int [\mathrm{e}^{\langle x, Ty\rangle}]\mathbf{P}_{m+1}(T) = \int \mathrm{e}^{\sum_{i\geq j} a_{ij}T_{ij}} \mathrm{d}\mathbf{P}_{m+1}(T) = \int \prod_{i\geq j} \mathrm{e}^{a_{ij}T_{ij}} \mathrm{d}\mathbf{P}_{m+1}(T).$$

On the one hand, for the off-diagonal elements satisfy, using the expression for the moment generating function of a standard normal random variable,

$$\mathbf{E}^{\mathbf{P}}_{m+1}[\mathrm{e}^{a_{ij}T_{ij}}] = \exp\left(\frac{1}{2}a_{ij}^2\right).$$

For the diagonal elements the situation is not as simple, but a numerical solution is possible. Indeed, for $a_{ii} \geq 0$, and $k_i = (m+1) - i + 1$

$$\begin{aligned}\mathbf{E}^{\mathbf{P}}_{m}[\mathrm{e}^{a_{ii}T_{ii}}] &= \frac{1}{2^{\frac{k_i}{2}}\Gamma\left(\frac{k_i}{2}\right)} \int_0^\infty x^{\frac{k_i}{2}-1} \exp\left(-\frac{1}{2}x + a_{ii}\sqrt{x}\right) \mathrm{d}x \\ &= {}_1F_1\left(\frac{k_i}{2}, \frac{1}{2}, \frac{a_{ii}^2}{2}\right) + \frac{\sqrt{2}a_{ii}\Gamma\left(\frac{k_i+1}{2}\right)}{\Gamma\left(\frac{k_i}{2}\right)} {}_1F_1\left(\frac{k_i+1}{2}, \frac{3}{2}, \frac{a_{ii}^2}{2}\right),\end{aligned}$$

where ${}_1F_1(a,b,z)$ is the Kummer confluent hypergeometric function. For $a_{ii} < 0$,

$$\frac{1}{2^{k_i/2}\Gamma\left(\frac{k_i}{2}\right)} \int_0^\infty x^{k_i/2-1} \exp\left(-\frac{1}{2}x + a_{ii}\sqrt{x}\right) \mathrm{d}x = \frac{\Gamma(k_i)}{2^{k_i-1}\Gamma\left(\frac{k_i}{2}\right)} U\left(\frac{k_i}{2}, \frac{1}{2}, \frac{a_{ii}^2}{2}\right),$$

and U is Kummer's U function.

A.2. Importance of the filtration for randomly stopped E-Statistics

Consider the the t-test as in Example 2.1.1. Fix some $0 < a < b$, and define the stopping time $\tau^* := 1$ if $|X_1| \notin [a, b]$. $\tau^* = 2$ otherwise. Then clearly τ^* is not adapted to (hence not a stopping time relative to) $(M_n)_n$ as defined in that example, since $M_1 \in \{-1, 1\}$ coarsens out all information in X_1 except its sign. Now let $\delta_0 := 0$ (so that $\mathcal{H}_0$ represents the normal distributions with mean $\mu = 0$ and arbitrary variance). Let $T_n^{*,\delta_1}(X^n)$ be equal to the GROW E-statistic $T_n^*(X^n)$ as in (2.13); here we make explicit its dependence on δ_1. For $\mathcal{H}_1$, to simplify computations, we put a prior $\tilde{\mathbf{\Pi}}_1^\delta$ on $\Delta_1 := \mathbb{R}$. We take $\tilde{\mathbf{\Pi}}_1^\delta$ to be a normal distribution with mean 0 and variance κ. We can now apply Corollary 2.8.3 (with prior $\tilde{\mathbf{\Pi}}_0^\delta$ putting mass 1 on $\delta = \delta_0 = 0$), which gives that $\tilde{T}_n(X^n)$ is an E-statistic, where

$$\tilde{T}_n(x^n) = \int \frac{1}{\sqrt{2\pi\kappa^2}} \exp\left(-\frac{\delta_1^2}{2\kappa^2}\right) \cdot T_n^{*,\delta_1}(x^n) \mathrm{d}\delta_1$$

coincides with a standard type of Bayes factor used in Bayesian statistics. By exchanging the integrals in the numerator, this expression can be calculated analytically. The Bayes factor $\tilde{T}_1(x_1)$ for $x^1 = x_1$ is found to be equal to 1 for all $x_1 \neq 0$, and the Bayes factor for (x_1, x_2) is given by:

$$\tilde{T}_2(x_1, x_2) = \frac{\sqrt{2\kappa^2 + 1} \cdot (x_1^2 + x_2^2)}{\kappa^2 (x_1 - x_2)^2 + (x_1^2 + x_2^2)}.$$

Now we consider the function

$$f(x) := \mathbf{E}_{X_2 \sim N(0,1)}[\tilde{T}_2(x, X_2)].$$

$f(x)$ is continuous and even. We want to show that, with τ^* as above, $\tilde{T}_{\tau^*}(X^{\tau^*})$ is not an E-variable for some specific choices of a, b and κ. Since, for any $\sigma > 0$, the null contains the distribution under which the X_i are i.i.d. $N(0, \sigma)$, the data may, under the null, in particular be sampled from $N(0, 1)$. It thus suffices to show that

$$\mathbf{E}_{X_1, X_2 \sim N(0,1)}[\tilde{T}_{\tau^*}(X^{\tau^*})] = \mathbf{P}_{X_1 \sim N(0,1)}\{|X_1| \notin [a, b]\} + \mathbf{E}_{X_1 \sim N(0,1)}[\mathbf{1}\{|X_1| \in [a, b]\} f(X_1)] > 1.$$

But from numerical integration we find that $f(x) > 1$ on $[a, b]$ and $[-b, -a]$ if we take $\kappa = 200$, $a \approx 0.44$ and $b \approx 1.70$. Using again numerical integration, we find that the above expectation is then approximately equal to 1.19, which shows that, even though $\tilde{T}_n$ is an E-statistic at each n by Corollary 2.8.3 (it is even a GROW one), $\tilde{T}_{\tau^*}$ is not an E-statistic (its expectation is 0.19 too large), providing the desired counterexample.

B. Appendix to Chapter 3

B.1. Omitted Proofs and Details

In this section we provide proofs and remarks omitted from previous sections. In Appendix B.1.1 we relate growth-rate optimality to the minimum expected stopping time. In Appendix B.1.2, we show that the AV logrank statistic is a continuous-time martingale, and show that this is also true for general patterns of incomplete observation, such as left truncation and filtering as a consequence of the results of Andersen et al. [1993]. In Appendix B.1.3, we proof the claims made in Section 3.3.3 about the martingale structure of the AV logrank test under the presence of ties. Lastly, in Appendix B.1.4, we give further details on the simulations used to compute the planned maximum sample sizes for a given targeted power. Under the alternative and optional stopping, the observed sample size is in many cases lower.

B.1.1. Expected Stopping Time, GROW and Wald's Identity

Here we motivate the GROW criterion by showing that it minimizes, in a worst-case sense, the expected number of events needed before there is sufficient evidence to stop. Let $\mathbf{P}_0$ represent our null model, and let, as before, the alternative hypothesis be $\mathcal{H}_1 : \theta \leq \theta_1$ for some $\theta_1 < \theta_0$. Suppose we perform a level-α test based on a test martingale $S^q_{\theta_0,t}$ using the stopping rule τ that stops as soon as $S^q_{\theta_0,t}$ exceeds the threshold $1/\alpha$, that is, $\tau^q = \inf_t\{t : S^q_{\theta_0,t} \geq 1/\alpha\}$. In the main text we elaborated on how $S^{\theta_1}_{\theta_0,t}$ is optimal with respect to the GROW criterion. We now show that the problem of minimizing the worst-case, the expected number of events $\mathbf{E}_\theta[\bar{N}_{\tau^q}]$ over q is approximately equivalent to finding the GROW test martingale. To do so, we make simplifying assumptions that reduce the problem to an i.i.d. experiment. This allows us to employ a standard argument based on an identity of Wald [1947], originally due to Breiman [1961]. For this we assume that the initial risk sets (i.e., $\bar{y}^A_0$ and $\bar{y}^B_0$) are large enough so that, for all sample sizes we will ever encounter, $\bar{y}^A_t/\bar{y}^B_t \approx \bar{y}^A_0/\bar{y}^B_0$. This allows us to treat the likelihood of the participant(s) $I_{(k)}$ having witnessed the event at time $T^{(k)}$ to be independent of t, that is, as an i.i.d. experiment.

The argument of Breiman [1961] relates the expected number of events to the expected value of our stopped AV logrank statistic. Suppose first that we happen to know that the data come from a specific θ in the alternative hypothesis. Then $S^q_{\theta_0,\tau}$ is the product of $\bar{N}_\tau$ factors of ratios $R^q_{\theta_0,(i)} = q_{(i)}(I_{(i)})/p_{\theta_0,(i)}(I_{(i)})$ at the ith event.

Wald's identity applied to its logarithm implies

$$\mathbf{E}_\theta[\bar{N}_\tau] = \frac{\mathbf{E}_\theta[\ln S^q_{\theta_0,\tau^q}]}{\mathbf{E}_\theta[\ln R^q_{\theta_0,(1)}]}. \tag{B.1}$$

For simplicity we will further assume that the number of participants at risk is large enough so that the probability that we run out of data before we can reject is negligible. Because of the choice of the stopping rule τ^q, the right-hand side of the last display can then be further rewritten as

$$\frac{\mathbf{E}_\theta[\ln S^q_{\theta_0,\tau^q}]}{\mathbf{E}_\theta[\ln R^q_{\theta_0,(1)}]} = \frac{\ln(1/\alpha) + \text{VERY SMALL}}{\mathbf{E}_\theta\left[\ln\left(q_{(1)}(I_{(1)})/p_{(1),\theta_0}(I_{(1)})\right)\right]},$$

where VERY SMALL between 0 and $\log|\theta_1/\theta_0|$. The equality follows because we reject as soon as $S^q_{\theta_0,t} \geq 1/\alpha$, so $S^q_{\theta_0,\tau}$ cannot be smaller than $1/\alpha$, and it cannot be larger by more than a factor equal to the maximum likelihood ratio at a single outcome (if we would not ignore the probability of stopping because we run out of data, there would be an additional small term in the numerator).

With (B.1) at hand, we can relate our choice of q to the expected number of events witnessed before stopping. If, for a fixed θ, we try find the q that minimizes the expected number of events $\mathbf{E}_\theta[\bar{N}_{\tau^q}]$, and, as is customary in sequential analysis, we approximate the minimum by ignoring the VERY SMALL part, we see that the expression is minimized by maximizing the numerator $\mathbf{E}_\theta\left[\ln\left(Q_{(1)}/P_{\theta_0,(1)}\right)\right]$ over q. The maximum is achieved by $Q_{(1)} = P_{\theta,(1)}$; the expression in the denominator then becomes the Kulback-Leibler divergence between two Bernoulli distributions. It follows that, under θ, the expected number of outcomes until rejection is minimized by $Q_{(1)} = P_\theta$. Thus, in this case, we use the GROW $S^\theta_{\theta_0,t}$ as test statistic. However, we still need to consider the fact that the real $\mathcal{H}_1$ is composite: as statisticians, we do not know the actual θ; we only know $0 < \theta \leq \theta_1$. A worst-case approach uses the q achieving

$$\max_q \min_{\theta \leq \theta_1} \mathbf{E}_\theta\left[\ln\left(p_{(1)}(I_{(1)})/q_{(1),\theta_0}(I_{(1)})\right)\right]$$

since, repeating the reasoning leading to (B.1), this q should be close to achieving the min-max number of events until rejection, given by

$$\min_q \max_{\theta \leq \theta_1} \mathbf{E}_\theta[\bar{N}_{\tau^q}]$$

But this just tells us to use the GROW E-variable relative to $\mathcal{H}_1$, which is what we were arguing for.

B.1.2. Continuous time and anytime validity

In this section, we show the anytime validity of the AV logrank test. This is done via Ville's inequality for which it suffices to show that $S^q_{\theta_0} = (S^q_{\theta_0,t})_{t\geq 0}$ is a nonnegative (super) martingale. To do so, we use the counting process formalism. A few definitions

are in order. Only in this section, we assume knowledge of counting process theory [see Andersen et al., 1993, Fleming and Harrington, 2011]. Denote, for $i = 1, \dots, m$, $\tilde{N}_t^i = \mathbf{1}\left\{t \le T^i\right\}$ the counting processes associated to each participant, and let y_t^i be the at-risk process. For each participant, the censored process N_t^i, which is observed, is given by $\mathrm{d}N_t^i = y_t^i \mathrm{d}\tilde{N}_t^i$—we use this convention to signify that $N_t^i = \int_0^t y_s^i \mathrm{d}\tilde{N}_s^i$. We define the sigma-algebra $\mathcal{F}_t := \sigma(N_s^j : 0 \le s \le t, j = 1, \dots, n)$, which, as usual, can be interpreted as the information in the study up to time t.

One of the results of the counting process theory is that the processes $\mathrm{d}N_t^i - y_t^i \mathrm{d}\lambda_t^i$ are martingales, where, recall, $y_t^i = \mathbf{1}\left\{X^i \ge t\right\}$ is the at-risk process, and λ_t^i is the hazard function associated to T^i. In that case, $y_t^i \mathrm{d}\lambda_t^i$ is called the compensator of N_t^i. The result that the AV logrank test is a martingale hinges specifically on this structure. Thus, any pattern that preserves this martingale structure also preserves the martingale property for the AV logrank test, and consequently its type-I error guarantees. Andersen et al. [1993, III.4] show exactly this under general patterns of incomplete observation provided that the mechanisms are independent of the observations. With this in mind, in the following, we only assume that the counting processes N_t^i have compensators A_t^i given by $\mathrm{d}A_t^i = y_t^i \mathrm{d}\lambda_t^i$.

The filtration $\mathcal{F} = (\mathcal{F}_s)_{s \ge 0}$ is right-continuous and we can safely identify predictable processes with left-continuous process. For some θ_0, denote by $\mathbf{P}_0$ the distribution under which, for each $i = 1, \dots, m$, the hazard function for T^i is $\lambda_t^i = \theta_0^{z^i} \lambda_t^A$, where $g^i = 0$ if $i \in A$ and $g^i = 1$ if $i \in B$. Recall from Section 3.2, if participant i belongs to Group B, $\lambda_t^i = \theta_0 \lambda_t^B = \theta_0 \lambda_t^A$; otherwise, $\lambda_t^i = \lambda_t^A$. Let $q_t^1, \dots, q_t^m$ be predictable processes such that $\sum_{i \le m} q_t^i y_t^i = 1$ a.s. for all t, that is, $\{q_t^i\}_{i \in \mathcal{R}_t}$ at each t is a probability distribution over the participants at risk at time t. Define r_t^i to be each of the ratios $r_t^i = q_t^i / p_{\theta_0,t}^i$. Define the predictable process $S_{\theta_0,t^-}^q = \lim_{s \uparrow t} S_{\theta_0,t^-}^q$. As such, at each t, the change $\mathrm{d}S_{\theta_0,t}^q = S_{\theta_0,t}^q - S_{\theta_0,t^-}^q$ of the AV logrank statistic $S_{\theta_1}^q$ at time t, given in (3.10), can be computed as

$$\mathrm{d}S_{\theta_0,t}^q = \sum_{i \le m} S_{\theta_0,t^-}^q (r_t^i - 1) \mathrm{d}N_t^i,$$

because no two events happen simultaneously with positive probability. Since S_{θ_0,t^-}^q is predictable, it is enough to prove that the process M_t defined by $\mathrm{d}M_t = \sum_{i \le m} (1 - r_t^i) \mathrm{d}N_t^i$ is a martingale [see Fleming and Harrington, 2011, Theorem 1.5.1]. Recall that $\bar{y}_t^A = \sum_{i \in A} y_t^i$ and $\bar{y}_t^B = \sum_{i \in B} y_t^i$. Then both $\bar{y}^A$ and $\bar{y}^B$ are left-continuous processes.

Lemma B.1.1. Let $\{q_t^i\}_{i \le m}$ be a collection of nonnegative left-continuous processes $q^i = (q_t^i)_{t \ge 0}$ such that $\sum_{i \le m} y_t^i q_t^i = 1$ for all t. Let $\{p_{\theta_0,t}^i\}_{i \le m}$ be the collection of processes given by

$$p_{\theta_0,t}^i = \frac{\theta_0^{g^i} y_t^i}{\bar{y}_t^A + \theta_0 \bar{y}_t^B}.$$

The process $M = (M_t)_{t \ge 0}$ given by $\mathrm{d}M_t = \sum_{i \le m} (1 - r_t^i) \mathrm{d}N_t^i$ is a martingale under $\mathbf{P}_0$ with respect to the filtration $\mathcal{F} = (\mathcal{F}_t)_{t \ge 0}$.

Proof. It suffices to show that the compensator A_t of M_t, given by $\mathrm{d}A_t = \sum_{i \le m} \sum_{i \le m} (r_t^i - 1) y_t^i \lambda_t^i \mathrm{d}t$ is zero. Define $\bar{q}_t^A = \sum_{i \in A} y_t^i q_t^i$ and $\bar{q}_t^B = \sum_{i \in B} y_t^i q_t^i$. Notice that by assumption

$\bar{q}_t^A + \bar{q}_t^B = 1.$, and recall that, under the null $\lambda_t^B = \theta_0 \lambda_t^A$. We can compute

$$\begin{aligned}
\sum_{i \le m} (r_t^i - 1) y_t^i \lambda_t^i &= \sum_{i \in A} y_t^i \lambda_t^A (r_t^i - 1) + \sum_{i \in B} y_t^i \lambda_t^B (r_t^i - 1) \\
&= \lambda_t^A [(\bar{y}_t^A + \theta_0 \bar{y}_t^B) \bar{q}_t^A - \bar{y}_t^A + \\
&\qquad (\bar{y}_t^A + \theta_0 \bar{y}_t^B) \bar{q}_t^B - \theta_0 \bar{y}_t^B] \\
&= \lambda_t^A [(\bar{y}_t^A + \theta_0 \bar{y}_t^B) \overbrace{(\bar{q}_t^A + \bar{q}_t^B)}^{=1} - \\
&\qquad (\bar{y}_t^A + \theta \bar{y}_t^B)] \\
&= 0,
\end{aligned}$$

where we used the assumption that $\sum_{i \le m} y_t^i q_t^i = \bar{y}_t^A q_t^A + \bar{y}_t^B q_t^B = 1$. As the compensator A_t of M_t is zero at each t, we conclude that M_t is a martingale, as was to be shown. □

Our previous discussion and the preceding lemma have the following corollary as a consequence.

Corollary B.1.2. *$S_{\theta_0}^q = (S_{\theta_0,t}^q)_{t \ge 0}$ is a nonnegative martingale with expected value equal to one.*

Hence, Ville's inequality holds for $S_{\theta_0}^q$, which implies that

$$\mathbf{P}_0 \{ S_{\theta_0,t}^q \ge 1/\alpha \quad \text{for some} \quad t \ge 0 \} \le \alpha.$$

This implies the anytime validity of the test $\xi_{\theta_0}^q = (\xi_{\theta_0,t}^q)_{t \ge 0}$ given by the AV logrank test $\xi_{\theta_0,t}^q = \mathbf{1}\left\{ S_{\theta_0,t}^q \ge 1/\alpha \right\}$.

B.1.3. Ties

The purpose of this section is twofold. Firstly, we prove Lemma 3.3.2. Secondly, we show that the conditional likelihood given in Section 3.3.3 indeed approximates the true conditional partial likelihood ratio under any distribution such that the hazard ratio is θ_1.

Our general strategy in this case is similar to the one undertaken in the continuous-monitoring case: we build a test martingale with respect to a filtration $\mathcal{G}^\star$, and use Ville's inequality to derive anytime-valid type-I error guarantees. Define, for each $k = 1, 2, \ldots$, the sigma-algebra $\mathcal{G}_k$ generated by all observations made in times $t_1, \ldots, t_k$, that is, $\mathcal{G}_k = \sigma(N_{t_l}^i, \tilde{N}_{t_l}^i \ : \ i = 1, \ldots, m; \ l = 1, \ldots, k)$, and the corresponding filtration $\mathcal{G} = (\mathcal{G}_k)_{k=1,2,\ldots}$. Under Cox's proportional hazard model, conditionally on $\mathcal{G}_{k-1}$, our observations $\Delta \bar{N}_k^A$ and $\Delta \bar{N}_k^B$ are binomially distributed with parameters depending on the hazard function (see Lemma B.1.3 below). By conditioning both on $\mathcal{G}_{k-1}$ and on the total number of events $\Delta \bar{N}_k = \Delta \bar{N}_k^A + \Delta \bar{N}_k^B$, we use the likelihood of having observed $\Delta \bar{N}_k^B$, which follows Fisher's noncentral hypergeometric distribution, as detailed in Corollary B.1.4. We gather these observations in the following two lemmas.

Lemma B.1.3. Conditionally on $\mathcal{G}_{k-1}$, the following hold:

1. The number of events $\Delta\bar{N}_k^A$ has a binomial distribution with parameters $\bar{y}_k^A$ and p_k^A where $p_k^A = 1 - \exp\left(-\int_{t_{k-1}}^{t_k} \lambda_s^A \mathrm{d}s\right)$.
2. The number of events $\Delta\bar{N}_k^B$ has a binomial distribution with parameters $\bar{y}_k^B$ and p_k^B where $p_k^B = 1 - \exp\left(-\theta\int_{t_{k-1}}^{t_k} \lambda_s^A \mathrm{d}s\right)$ and θ is the hazard ratio.

Proof. The result is standard, and it follows from explicitly solving for λ in (3.1) and computing the conditional probability in (3.2) for each group. □

Next, we use a standard result: given two binomially distributed random variables X and Y, the distribution of X conditionally on $X+Y$ is Fisher's noncentral hypergeometric distribution. We apply this to $\Delta\bar{N}_k^A$ and $\Delta\bar{N}_k^B$ from the previous lemma and spell out the corresponding parameters in the following corollary.

Corollary B.1.4. *Let $\mathcal{G}_{k-1}^\star = \mathcal{G}_{k-1} \vee \sigma(\Delta\bar{N}_k)$, and let p_k^A and p_k^B be as in Lemma B.1.3. Define the odd ratios $\omega_k^A = p_k^A/(1-p_k^A)$, $\omega_k^B = p_k^B/(1-p_k^B)$ and the ratio $\omega_k = \omega_k^B/\omega_k^A$. Then, conditionally on $\mathcal{G}_{k-1}^\star$, the likelihood of having observed $\Delta\bar{N}_k^B$ events in group B is given by Fisher's noncentral hypergeometric distribution with probability mass function $p_{\mathrm{FNCH}}(\Delta\bar{N}_k^B \,;\, \bar{y}_{k-1}^B, \bar{y}_{k-1}^A, \Delta\bar{N}_k, \omega_k)$ given by*

$$p_{\mathrm{FNCH}}(n^B;\ \bar{y}^B, \bar{y}^A, n, \omega) = \frac{\binom{\bar{y}^B}{n^B}\binom{\bar{y}^A}{n-n^B}\omega^{n^B}}{\sum_{\max\{0, n^B-\bar{y}^B\}\le u\le \min\{\bar{y}^B, n^B\}}\binom{\bar{y}^B}{u}\binom{\bar{y}^A}{n^B-u}\omega^u}.$$

Naively, one could use a partial likelihood ratio just as in the absence of ties to derive a sequential test. This, however, is not satisfactory, because, in general, the parameter ω_k depends heavily on the unknown baseline hazard function λ^A. Contrary to the general case, when the hazard ratio θ is one, the parameter $\omega_k = 1$, and Fisher's noncentral hypergeometric distribution reduces to the conventional hypergeometric distribution. With this observation at hand, if $\{q_k\}_{k=1,2,\dots}$ is a sequence of conditional distributions $q_k(\,\cdot\,)$ on the possible values of $\Delta\bar{N}_k^B$, we can build a sequential tests for (3.3), with its corresponding type-I error guarantee. We give the details in the following corollary, and subsequently point at a useful choice for q that approximates the real likelihood.

The choice of q for our statistic presented in Section 3.3.3 follows from an approximation of the parameter ω for small $\Delta t_k = t_k - t_{k-1}$. As noted by Mehrotra and Roth [2001], if $\int_{t_{k-1}}^{t_k} \lambda_1(s)\mathrm{d}s$ is small, then $p_k^A \approx \lambda_{t_{k-1}}\Delta t_k$ and $p_k^A \approx \theta p_k^A$. With these two approximations, $\omega_k \approx \theta$. This means that the choice $q_k(\Delta\bar{N}_k^B) = p_{\theta_1,k}(\Delta\bar{N}_k^B) := p_{\mathrm{FNCH}}(\Delta\bar{N}_k^B;\ \bar{y}_k^B,\ \bar{y}_k^A,\ \Delta\bar{N}_k,\ \omega = \theta_1)$ approximates the real conditional likelihood under any alternative for which the true hazard ratio is θ_1. Hence, the sequentially computed statistic

$$S_k^{\theta_1} = \prod_{l\le k} \frac{p_{\theta_1,k}(\Delta\bar{N}_k^B)}{p_{1,k}(\Delta\bar{N}_k^B)}$$

approximates the true partial likelihood ratio of the data observed up to time t_k in the presence of ties, and we recommend its use.

B.1.4. Details of sample size comparison simulations

In this section we lay out the procedure that we used to estimate the expected and maximum number of events required to achieve a predefined power as shown in Figure 3.4 and Figure 3.1 in Section 3.7. First we describe how we sampled the survival processes under a specific hazard ratio. We then describe how we estimated the maximum and expected sample size required to achieve a predefined power (80% in our case) for any of the test martingales that we considered (that of the exact AV logrank, its Gaussian approximation, and the prequential plugin variant). Finally, we explain how the same quantitiees for the classical logrank test and the O'Brien-Fleming procedure were obtained.

In order to simulate the order in which the events in a survival processes happens, we used the sequential-multinomial risk-set process from Section 3.3. As before, we consider the general testing problem with $\theta_0 = 1$ and a minimal clinically relevant effect size $\theta_1 < 1$, and we denote the true data generating parameter by θ, typically, $\theta \leq \theta_1$. Under θ, the odds of the next event at the i^{th} event time happening in Group B are $\theta\bar{y}^B_{(i)} : \bar{y}^A_{(i)}$—the odds change at each time step. Thus, simulating in which group the next event happens only takes a biased coin flip. For the problem of testing (3.11) with θ_0 we fix the tolerate a type-I error to $\alpha = 0.05$ and the type-II error to $\beta = 0.2$. For each test martingale $S^q_{\theta_0}$ of interest we first consider the stopping rule $\tau^q = \inf\{k : S^q_{\theta_0,(k)} \geq 1/\alpha\}$, that is, we stop as soon as $S^q_{\theta_0,(i)}$ crosses the threshold $1/\alpha$. Recall that in the worst case, $\theta = \theta_1$ the expected stopping time τ^q is lowest when we use $S^{\theta_1}_{\theta_0,(k)}$, see Appendix B.1.1.

To estimate the maximum number of events needed to achieve a predefined power with a given test martingale, we turned our attention to a modified stopping rule $\tilde{\tau}^q$. Under $\tilde{\tau}^q$ we stop at the first of two moments: either when our test martingale $S^q_{\theta_0,(k)}$ crosses the threshold $1/\alpha$ (i.e., at τ) or once we have witnessed a predefined maximum number of events $n_{\max}$. More compactly, this means using the stopping rule $\tilde{\tau}^q$ given by $\tilde{\tau}^q = \min(\tau^q, n_{\max})$. In those cases in which the test based on the stopping rule τ^q achieves a power higher than $1-\beta$, a maximum number of events $n_{\max}$ smaller than the initial size of the combined risk groups can be selected to achieve approximate power $1-\beta$ using the rule $\tilde{\tau}^q$.

A quick computation shows that $n_{\max}$ has the following property: it is the smallest number of events n such that stopping after n events has probability smaller than $1-\beta$ under the alternative hypothesis, that is,

$$\mathbf{P}_\theta\{\tau^q \geq n\} \leq 1-\beta.$$

More succinctly, $n_{\max}$ is the (approximate) $(1-\beta)$-quantile of the stopping time τ^q, which can be estimated experimentally in a straightforward manner.

To estimate $n_{\max}$ for an initial risk set sizes m_1, m_0, we sampled 10^4 realizations of the survival process (under θ) using the method described at the beginning of this

section. This allowed us to obtain the same number of realizations of the stopping time τ^q. We then computed the $(1-\beta)$-quantile of the simulated first passage time distribution of τ^q, and reported it as an estimate of the number of events $n_{\max}$ in the 'maximum' column in Figure 3.4.

We assessed the uncertainty in the estimation $n_{\max}$ using the bootstrap. We performed 1000 bootstrap rounds on the sampled empirical distribution of τ^q, and found that the number of realizations that we sampled (10^4) was high enough so that plotting the uncertainty estimates was not meaningful relative to the scale of our plots. For this reason we omitted the error bars in Figure 3.4 and Figure 3.1.

In the "mean" column of Figure 3.4 and Figure 3.1 we plot an estimate of the expected number of events $\tilde{\tau}^q = \min(\tau^q, n_{\max})$. For this, we used the empirical mean of the stopping times that were smaller than $n_{\max}$ on the sample that we obtained by simulation, with 20% of the stopping times being $n_{\max}$ itself. In the "conditional mean" column, we plot an estimate of $\tilde{\tau}^q \mid \tilde{\tau}^q < n_{\max}$, i.e., the stopping time given that we stop early (and hence reject the null).

For comparison, we also show the number of events that one would need under the Gaussian non-sequential approximation of Schoenfeld [1981], and under the continuous monitoring version of the O'Brien-Fleming procedure. In order to judge Schoenfeld's approximation, we report the number of events required to achieve 80% power. This is equivalent to treating the logrank statistic as if it were normally distributed, and rejecting the null hypothesis using a z-test for a fixed number of events. The power analysis of this procedure is classic, and the number of events required is $n^S_{\max} = 4(z_\alpha + z_\beta)^2 / \log^2 \theta_1$, where z_α, and z_β are the α, and β-quantiles of the standard normal distribution. In the case of the continuous monitoring version of O'Brien-Fleming's procedure, we estimated the number of events $n^{OF}_{\max}$ needed to achieve 80% as follows. For each experimental setting (m^A, m^B, θ), we generated 10^4 realizations of the survival process under θ and computed the corresponding trajectories of the logrank statistic. For each possible value n of $n^{OF}_{\max}$, we computed the fraction of trajectories for which the O'Brien-Fleming procedure correctly stopped when used with the maximum number of events set to n. We report as an estimate of the true $n^{OF}_{\max}$ the first value of n for which this fraction is higher than 80%, our predefined power.

B.2. Covariates: the full Cox Proportional Hazards E-Variable

We extend the AV logrank test to the situation when time-dependent covariates are present, as done in Section 3.3 with the same notation used there. Assume now the presence of d covariates and let, for each participant i, $\mathbf{z}^i_t = (z^i_{t,1}, \ldots, z^i_{t,d})$ be the covariate vector consisting left-continuous time-dependent covariates $z^i_{t,1}, \ldots, z^i_{t,d}$. Denote by $\mathbf{z}^i_{(k)}$ the value of the covariates of participant i at the time $T_{(k)}$ when the kth event is witnessed. We let random variable $I_{(k)}$ denote the index of the patient to which the kth event happens, and consider the extended process $I_{(1)}, I_{(2)}, \ldots$ where

the information that is available at time $T_{(k)}$ is, $I_{(1)}, I_{(2)}, \ldots, I_{(k)}$, and $z_{(1)}, \ldots, z_{(k)}$. The conditional partial likelihood underlying the process is now denoted $\mathbf{P}_{\beta,\theta}$ with $\theta > 0$, $\boldsymbol{\beta} \in \mathbb{R}^d$, and $\beta_\theta = \ln \theta \in \mathbb{R}$, defined as follows:

$$\begin{aligned}
&\mathbf{P}_{\boldsymbol{\beta},\theta}\{I_{(k)} = i \mid \mathbf{z}^j_{(l)}, y^j_{(l)};\ j = 1, \ldots, n;\ l = 1, \ldots, k\} := \\
&\qquad\qquad P_{\beta,\theta}\{I_{(k)} \mid \mathbf{z}^i_{(k)}, y^i_{(k)};\ i = 1, \ldots, n\}, \text{ and} \\
&\mathbf{P}_{\boldsymbol{\beta},\theta}\{I_{(k)} = i \mid \mathbf{z}^i_{(k)}, y^i_{(k)};\ i = 1, \ldots, n\} := \\
&\qquad p_{\boldsymbol{\beta},\theta,(k)}(\ i\ ) := \frac{y^i_{(k)} \exp\left(\langle \boldsymbol{\beta}, \mathbf{z}^i_{(k)} \rangle + g^i \beta_\theta\right)}{\sum_{j \in \mathcal{R}_{(k)}} \exp\left(\langle \boldsymbol{\beta}, \mathbf{z}^j_{(k)} \rangle + g^i \beta_\theta\right)},
\end{aligned}$$

This is consistent with Cox' (1972) proportional hazards regression model: the probability that the ith participant witnesses an event, assuming he/she is still at risk, is proportional to the exponentiated weighted covariates, with group membership being one of the covariates. In case $\boldsymbol{\beta} = 0$, this is easily seen to coincide with the definition of $\mathbf{P}_\theta$ via (3.5) with $\theta = \mathrm{e}^{\beta_\theta}$.

B.2.1. E-Variables and Martingales

Let $\mathbf{W}$ be a prior distribution on $\beta \in \mathbb{R}^d$ for some $d > 0$. ($\mathbf{W}$ may be degenerate, i.e., put mass one on a specific parameter vector β_1). For each such $\mathbf{W}$, we let $q_{\mathbf{W},\theta,(k)}$ be the probability distribution on $\mathcal{R}_{(k)}$ defined by

$$q_{\mathbf{W},\theta,(k)}(\ i\ ) := \int p_{\boldsymbol{\beta},\theta,(k)}(\ i\ ) \mathrm{d}\mathbf{W}(\boldsymbol{\beta}).$$

Consider a measure ρ on $\mathbb{R}^d$ (e.g., Lebesgue or some counting measure) and we let $\mathcal{W}$ be the set of all distributions on $\mathbb{R}^d$ which have a density relative to ρ, and $\mathcal{W}^\circ \subset \mathcal{W}$ be any convex subset of $\mathcal{W}$ (we may take $\mathcal{W}^\circ = \mathcal{W}$, for example). We define $\tilde{q}_{\leftarrow \mathbf{W},\theta_0}$ to be the *reverse information projection* [Li, 1999] (RIPr) of $q_{\mathbf{W},\theta,(k)}$ on $\{q_{\mathbf{W},\theta_0,(k)} : \mathbf{W} \in \mathcal{W}^\circ\}$, defined as the probability distribution on $\mathcal{R}_{(k)}$ such that

$$\mathrm{KL}(q_{\mathbf{W},\theta_1,(k)} \| \tilde{q}_{\leftarrow \mathbf{W},\theta_0,(k)}) = \inf_{\mathbf{W}^\circ \in \mathcal{W}^\circ} \mathrm{KL}(q_{\mathbf{W},\theta_1,(k)} \| q_{\mathbf{W}^\circ,\theta_0,(k)}).$$

We know from Li [1999] and Grünwald et al. [2020] that $\tilde{q}_{\leftarrow \mathbf{W},\theta_0,(k)}$ exists for each k. Grünwald et al. [2020] show, in the context of E-variables for 2×2 contingency tables, that the infimum in the previous display is in fact achieved by some distribution $\mathbf{W}^\star$ with finite support on $\mathbb{R}^d$ if the random variables $y^1_{(k)}, \ldots, y^m_{(k)}$ constituting our random process have a finite range. For given hazard ratios $\theta_0, \theta_1 > 0$, let

$$R^{\theta_1}_{\mathbf{W},\theta_0,(k)} = \frac{q_{\mathbf{W},\theta_1,(k)}(I_{(k)})}{q_{\leftarrow \mathbf{W},\theta_0,(k)}(I_{(k)})} \tag{B.2}$$

be our analogue of (3.8).

Theorem B.2.1 (Corollary of Theorem 1 from Grünwald et al. [2020])**.** *For every prior* $\mathbf{W}$ *on* $\mathbb{R}^d$*, for all* $\boldsymbol{\beta} \in \mathbb{R}^d$*,*

$$\mathbf{E}_{\boldsymbol{\beta},\theta_0}[R^{\theta_1}_{\mathbf{W},\theta_0,(k)} \mid \mathbf{z}^i_{(l)}, y^i_{(l)};\ i = 1,\ldots,m;\ l = 1,\ldots,k] = \sum_{i \in \mathcal{R}_{(k)}} q_{\boldsymbol{\beta},\theta_0,(k)}(i) \frac{q_{\mathbf{W},\theta_1,(k)}(i)}{q_{\leftarrow \mathbf{W},\theta_0,(k)}(i)} \leq 1$$

so that $R^{\theta_1}_{\mathbf{W},\theta_0,(k)}$ *is an E-variable conditionally on* $\mathbf{z}^i_{(l)}, y^i_{(l)}$ *with* $i = 1,\ldots,m;\ l = 1,\ldots,k$*.*

Note that the result does not require the prior $\mathbf{W}$ to be well specified in any way: under any $(\boldsymbol{\beta}, \theta_0)$ in the null distribution, even if $\boldsymbol{\beta}$ is completely disconnected to $\mathbf{W}$, $R^{\theta_1}_{\mathbf{W},\theta_0,(k)}$ is an E-variable conditional on past data.

In particular, since the result holds for arbitrary priors, it holds, at the kth event time, for the Bayesian posterior $\mathbf{W}_{k+1} = \mathbf{W}_1 \mid \mathbf{z}^i_{(l)}, y^i_{(l)};\ i = 1,\ldots,m;\ l = 1,\ldots,k$, based on arbitrary prior $\mathbf{W}_1$ with density w_1, i.e., the density of $\mathbf{W}_{k+1}$ is given by

$$w_{k+1}(\boldsymbol{\beta}) \propto \prod_{l \leq k} q_{\boldsymbol{\beta},\theta,(l)}(I_{(l)}) w_1(\boldsymbol{\beta}).$$

In parallel to the discussion in Section 3.3.1, we can therefore, for each prior $\mathbf{W}_1$, construct a test martingale $S_k := \prod_{l \leq k} R^{\theta_1}_{\mathbf{W}_l,\theta_0,(l)}$ that "learns" $\boldsymbol{\beta}$ from the data, analogously to (3.12), and computes a new RIPr at each event time k.

B.2.2. Finding the RIPr

While it is not clear how to calculate the RIPr $q_{\leftarrow \mathbf{W},\theta_0,(k)}$ in general, it can be well approximated with the efficient algorithm design by Li [1999] and Li and Barron [1999]. Their algorithm is computationally feasible as long as we restrict $\mathbf{W}^\circ_\delta$ to be the set of all priors $\mathbf{W}$ for which $\min_{i \in \mathcal{R}_{(k)}} q_{\mathbf{W},\theta_0,(k)}(i) \geq \delta$, for some $\delta > 0$. In that case, when run for M steps, the algorithm achieves an approximation error of $O(\ln(1/\delta)/M)$, where each step is linear in the dimension d. Since the approximation error is logarithmic in $1/\delta$, we can take a very small value of δ, which makes the requirement less restrictive. Exploring whether the Li-Barron algorithm really allows us to compute the RIPr for the Cox model, and hence $R^{\theta_1}_{\mathbf{W}_k,\theta_0,(k)}$ in practice, is a major goal for future work.

B.2.3. Ties

Without covariates, our E-variables allow for ties correspond to a likelihood ratio of Fisher's noncentral hypergeometric distributions (see Section 3.3.3), the situation is not so simple in the presence of covariates. Although deriving the appropriate extension of the noncentral hypergeometric partial likelihood is possible, one ends up with a hard-to-calculate formula [Peto, 1972]. Various approximations have been proposed in the literature [Cox, 1972, Efron, 1977]. In case these preserve the E-variable and martingale properties, they would retain type-I error probabilities under

optional stopping and we could use them without problems. We do not know whether this is the case however; for the time being, we recommend handling ties by putting the events in a worst-case order, leading to the smallest values of the E-variable of interest, as this is bound to preserve the type-I error guarantees.

B.3. Gaussian AV logrank test

In this section we derive the Gaussian AV logrank test of Section 3.4, and investigate the validity of the Gaussian approximation. In Appendix B.3.1, we show that this approximation is only valid when the allocation of participants to each group under investigation is balanced, that is, when $m^A = m^B$. In Appendix B.3.2 we investigate numerically the sample size needed to reject the null hypothesis under both the exact AV logrank test and its Gaussian approximation.

We start with the derivation of (3.15). For this we use (local) asymptotic normality of the Z-score (3.14). Under the null distribution, Z_k from (3.14) has an asymptotic standard Gaussian distribution. Under any alternative distribution under which the hazard ratio is θ, Schoenfeld [1981] showed that, in the absence of ties, the Z-statistic also follows a Gaussian distribution with unit variance, but this time with mean $\mu_1^\star$ given by

$$\mu_1^\star = \frac{\sum_{i \le k} E_i^B (1 - E_i^B)}{\sqrt{\sum_{i \le k} E_i^B (1 - E_i^B)}} \log(\theta).$$

Note that $\mu_1^\star$ depends on more than the summary statistic Z_k. In the case that the number of observed events is much smaller than the initial risk set sizes, the mean $\mu_1^\star$ under the alternative can be further approximated by

$$\mu_1^\star \approx \sqrt{\bar{N}_k}\mu_1 = \sqrt{\bar{N}_k}\sqrt{\frac{m^B m^A}{(m^B + m^A)^2}} \log(\theta), \tag{B.3}$$

where $\bar{N}_k$ is the total number of observations up until time t_k, and the resulting approximation only depends on summary statistics. It is exactly this value μ_1 that we use in the Gaussian AV logrank test. The asymptotic result of Schoenfeld relies on two conditions: (1) that the hazard ratio θ_1 under the alternative is close enough to one so that a first-order Taylor approximation around $\theta_0 = 1$ is adequate; (2) that the expected number of events E_k^B stays approximately constant over time, that is, close to the initial allocation proportion $E_1^B = m^B/(m^B + m^A)$. This indicates that the asymptotic approximation is reasonable for values of θ_1 close to 1 and the initial risk sets are both large in comparison to the number of events witnessed. Notice that in this regime of large risk sets the multiplicity correction in V_k is also negligible.

This raises the question whether a sequential Gaussian approximation is sensible for the logrank statistic— a priori it is not at all clear whether Schoenfeld's asymptotic fixed-sample result has a nonasymptotic counterpart. Define the the logrank statistic per observation time

$$Z_i = \frac{O_i^B - E_i^B}{\sqrt{V_i^B}}.$$

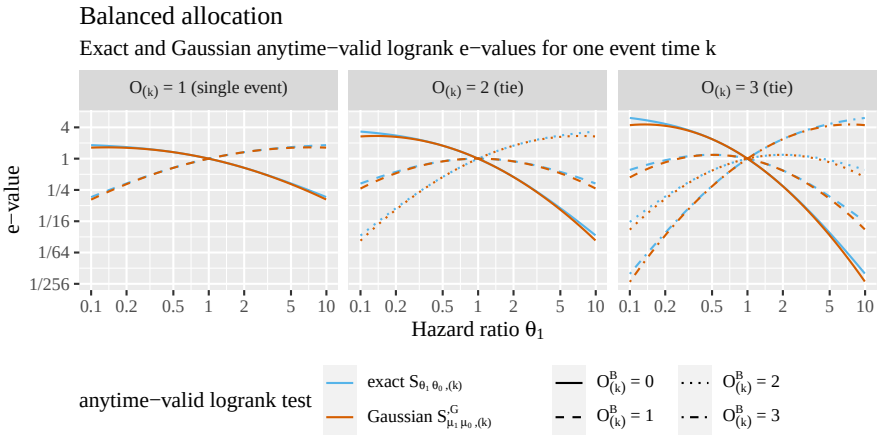

Figure B.1.: For balanced allocation ($m^A = m^B$) $S_1'^G$ is very similar to $S_{(1)}$ when $0.5 \leq \theta_1 \leq 2$. Here $\theta_0 = 1$, $\mu_0 = 0$, and $\mu_1 = \mu_1(\theta_1)$ as in in (B.3). Note that both axis are logarithmic.

We investigate whether the exact AV logrank statistic behaves similarly to the Gaussian likelihood ratio

$$S_k'^G = \prod_{i \leq k} \frac{\phi_{\mu_1\sqrt{O_i}}(Z_i)}{\phi_{\mu_0}(Z_i)} = \exp\left(-\frac{1}{2}\sum_{i\leq k}\left\{O_i\mu_1^2 - 2\mu_1\sqrt{O_i Z_i}\right\}\right)$$

for $\theta_0 = 1$ we have $\mu_0 = 0$, $\mu_1 = \log(\theta)\sqrt{m^B m^A/(m^A+m^B)^2}$, and ϕ_μ is the Gaussian density with unit variance and mean μ. Note that the statistic still depends on elements of the full data set; more approximations are needed. Write the Gaussian densities, and use that in the limit of large risk sets $p_i^B \approx m^B/(m^A+m^B)$ and that consequently $V_i \approx \sqrt{O_i \frac{m^A m^B}{(m^A+m^B)^2}}$. This approximations valid under Schoenfeld's second assumption. With these approximations at hand, the Z-statistic is approximated by

$$Z_k \approx \frac{\sum_{i\leq k}\left\{O_i^B - E_i^B\right\}}{\sqrt{O_i \frac{m^A m^B}{(m^A+m^B)^2}}}$$

and consequently

$$S_k'^G \approx S_k^G \approx S_k^G = \exp\left(-\frac{1}{2}\bar{N}_k\mu_1^2 + \sqrt{\bar{N}_k}\mu_1 Z_k\right),$$

where S_k^G is as in (3.15). In Figure B.1 we show, in case of balanced allocation, that the Gaussian approximation S_k^G a single event time from the Gaussian approximation are very similar to the exact $S_{\theta_0,(k)}^{\theta_1}$ for alternative hazard ratios θ_1 between 0.5 and 2.

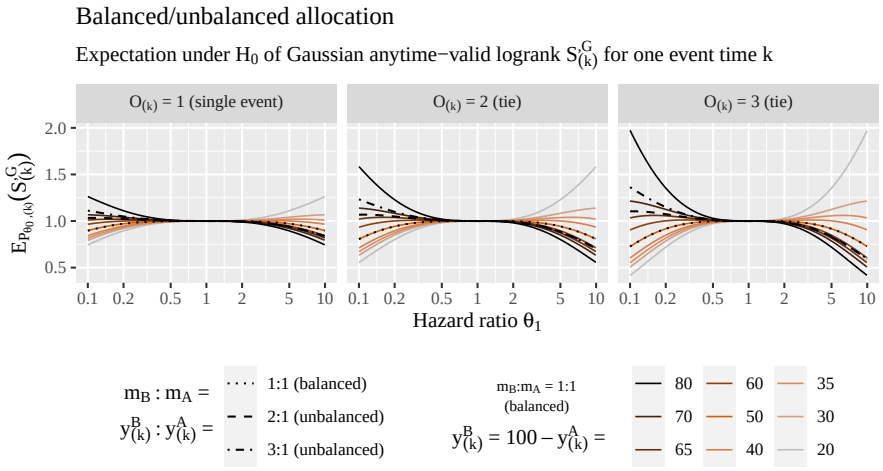

Figure B.2.: Expected value of the increments of the Gaussian AV logrank statistic as a function of the hazard ratio θ_1. For balanced allocation R_i^G is an E-variable, but it is not for unbalanced allocation. The risk set can also start out balanced but become unbalanced; this is unlikely under the null hypothesis (see Appendix B.3.1). Note that the x-axis is logarithmic.

B.3.1. Safety only for balanced allocation

In order to assess whether the Gaussian AV logrank test is indeed AV, that is, whether the type-I error guarantees holds, we inspect whether the expected value of each of its multiplicative increments is bellow 1. In relation to our discussion in Section 3.3.1, this would imply that all multiplicative increments are conditional E-variables and that the resulting test is, at least approximately, a test martingale. Figure B.2 shows the expectation of these increments as a function of the hazard ratio for several initial allocation ratios. In case of balanced 1:1 allocation S_k^G is an E-variable, since its expectation is 1 or smaller. However, in case of unbalanced 2:1 or 3:1 allocation and designs with hazard ratio $\theta_1 < 1$, S_k^G is not an E-variable. Of course, even if the initial allocation is balanced, it can become unbalanced. Figure B.2 shows that in case of designs outside the range $0.5 \leq \theta_1 \leq 2$ the deviations from expectation 1 can be problematic. Hence we do not recommend to use the Gaussian approximation on the logrank statistic for unbalanced designs and designs for $\theta_1 < 0.5$ or $\theta_1 > 2$. For balanced designs with $0.5 \leq \theta_1 \leq 2$, we found that in practice they are safe to use, the reason being that scenarios in which the allocation becomes highly unbalanced after some time (e.g. $y_i^B = 80, y_i^A = 20$) are extremely unlikely to occur under the null.

B.3.2. Sample size

In this section we compare the stopping time distribution $\tau^G := \inf\{k : \xi_k^G = 1\}$ of the Gaussian approximation to that of $\tau = \inf\{k : \xi_k = 1\}$. We use tests with tolerable type I error $\alpha = 0.05$, thus, the threshold $1/\alpha = 20$ for both tests. In the previous section we showed that the Gaussian approximation to the AV logrank statistic is valid when the initial allocation is 1:1 and for values $0.5 \leq \theta_1 \leq 2$, where θ_1 is the hazard ratio under the alternative. In these scenarios, we simulate a survival process from a distribution according to which the true data generating hazard ratio is $\theta = \theta_1$ and sampled realizations τ^G and τ for the same data set. The results of the simulation are shown in Figure B.3, where we plot the realizations of τ^G against those of τ. We see that in most cases both tests reject at the same time $\tau^G = \tau$, and that the approximation becomes better as θ_1 moves closer to $\theta_0 = 1$ (Schoenfeld's assumption 1). When both tests do not reject at the same time, the Gaussian approximation errs on the conservative side. The deviations from the constant large and balanced risk set do not seem to occur often for this range of hazard ratios. After all, the risk set needs to be large to observe the number of events to detect hazard ratios in the range $0.5 \leq \theta_1 \leq 2$.

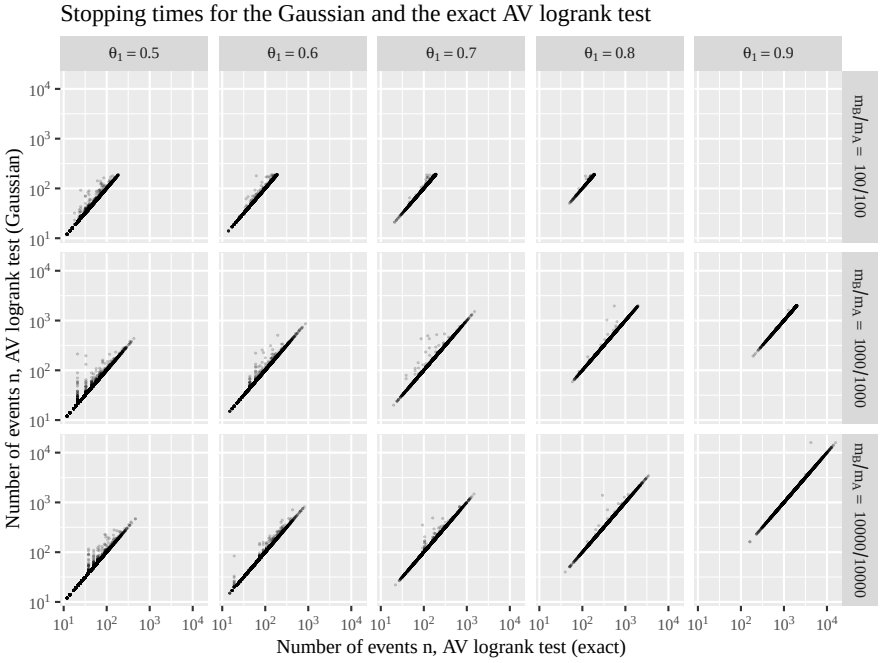

Figure B.3.: Stopping times for the Gaussian and exact AV logrank tests under continuous monitoring (no ties) with threshold $1/\alpha = 20$. The stopping times under the Gaussian approximation often coincide with the exact ones, and are often more conservative (see Appendix B.3.2). Note that both axes are logarithmic.

.

C. Appendix to Chapter 4

C.1. Saddle-Point Computation in Multiscale Games

One application of online learning is computing approximate mixed-strategy Nash equilibria in finite two-player zero-sum games (and more generally, to approximate saddle points of convex-concave functions). Here, we investigate a multiscale version of that problem. Our main focus is to find methods whose performance does not depend on the *maximum scale*, but on the *relevant scale* to the problem instance at hand. In this case, this means the scale of the payoffs in the subset of rows and columns in the support of the Nash equilibrium. In Section C.1.1 we lay out the setup of two-player zero-sum finite games. In Section C.1.2 we define the suboptimality gap, the main measure of performance in judging the solution to these games. In Section C.1.3 we define the payoff matrices used in the experiments that produced Figure 4.4. We conjecture that MUSCADA achieves fast scale-dependent convergence in Section C.1.5 and provide the additional details of the experiments that produced Figure 4.4(right) in Section C.1.6.

C.1.1. Two-player zero-sum finite games

Given a payoff matrix $A \in \mathbb{R}^{K\times M}$ (specifying losses for the row player and gains for the column player) we are looking for the mixed-strategy saddle point $(\boldsymbol{p}_*, \boldsymbol{q}_*) \in \mathcal{P}(K) \times \mathcal{P}(M)$ such that

$$\min_i \boldsymbol{e}_i^\top A \boldsymbol{q}_* \;\geq\; \max_j \boldsymbol{p}_*^\top A \boldsymbol{e}_j.$$

Our approach will be based on oracle access to the matrix-vector products $\boldsymbol{q} \mapsto A\boldsymbol{q}$ and $\boldsymbol{p} \mapsto A^\top \boldsymbol{p}$. We will use the scheme of running two online learners against each other, with loss vectors $\boldsymbol{\ell}_t^{\text{row}} = A\boldsymbol{q}_t$ and $\boldsymbol{\ell}_t^{\text{col}} = -A^\top \boldsymbol{p}_t$ and optimistic estimates given by the past loss vector $\boldsymbol{m}_t^{\text{row/col}} = \ell_{t-1}^{\text{row/col}}$. For the same-scale case, Rakhlin and Sridharan [2013] show that uncoupled adaptive schemes benefit from convergence of the gap of the pair of iterate averages at rate $O(\sigma_{\max}\frac{\ln K + \ln M}{T})$, while recently Hsieh et al. [2021] showed last iterate convergence as well. Here we investigate the advantage of using adaptive multiscale learners to improve the dependence in $\sigma_{\max}$.

C.1.2. The metric of success: suboptimality gap

We are looking for the equilibrium in mixed strategies, i.e. $\min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\top A \boldsymbol{q}$. The social exploitability of a candidate saddle point pair $\boldsymbol{p}, \boldsymbol{q}$ is defined as the gap

$$\text{gap}(\boldsymbol{p}, \boldsymbol{q}) \;=\; \max_j \boldsymbol{p}^\top A \boldsymbol{e}_j - \min_i \boldsymbol{e}_i^\top A \boldsymbol{q}.$$

We use the common technique of employing online learning with linear loss functions $\boldsymbol{p} \mapsto A\boldsymbol{q}_t$ and $\boldsymbol{q} \mapsto -A^\top \boldsymbol{p}_t$. A standard analysis [Freund and Schapire, 1997] bounds the gap of the iterate averages $\bar{\boldsymbol{p}}_t = \frac{1}{t}\sum_{s\le t}\boldsymbol{p}_s$ and $\bar{\boldsymbol{q}}_t = \frac{1}{t}\sum_{s\le t}\boldsymbol{q}_s$ from above by the social (sum-of) regret

$$\begin{aligned}\operatorname{gap}(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{q}}_t) &= \max_j \bar{\boldsymbol{p}}_t^\top A\boldsymbol{e}_j - \min_i \boldsymbol{e}_i^\top A\bar{\boldsymbol{q}}_t = \frac{1}{t}\left(\max_j \sum_{s\le t}\boldsymbol{p}_s^\top A\boldsymbol{e}_j - \min_i \sum_{s\le t}\boldsymbol{e}_i^\top A\boldsymbol{q}_s\right)\\ &= \frac{1}{t}\max_{i,j}\Big(\underbrace{\sum_{s\le t}\boldsymbol{p}_s^\top A\boldsymbol{e}_j - \sum_{s\le t}\boldsymbol{p}^\top A\boldsymbol{p}_s}_{R_t^{\boldsymbol{q}}(j)} + \underbrace{\sum_{s\le t}\boldsymbol{p}^\top A\boldsymbol{p}_s - \sum_{s\le t}\boldsymbol{e}_i^\top A\boldsymbol{q}_s}_{R_t^{\boldsymbol{p}}(i)}\Big).\end{aligned}$$

Having multiscale regret bounds at our disposal, it is natural to look at multiscale payoff matrices.

C.1.3. Multiscale structure

We will assume that our payoff matrix is multiscale in the sense that we are given row and column range vectors $\boldsymbol{\sigma}^{\text{row}}$ and $\boldsymbol{\sigma}^{\text{col}}$ such that $|A_{ij}| \le \min\{\sigma_i^{\text{row}}, \sigma_j^{\text{col}}\}$. The main point is to learn the saddle point faster if the maximum range is much larger than the range in the support of the saddle point, i.e. $\sigma_{\max}^{\text{row}} \gg \sigma_{\text{real}}^{\text{row}} := \max\{\sigma_i^{\text{row}} | \boldsymbol{e}_i^\top \boldsymbol{p}_* > 0\}$ and/or $\sigma_{\max}^{\text{col}} \gg \sigma_{\text{real}}^{\text{col}} := \max\{\sigma_j^{\text{col}} | \boldsymbol{e}_j^\top \boldsymbol{q}_* > 0\}$. We will denote that largest relevant scale by $\sigma_{\text{real}} = \max\{\sigma_{\text{real}}^{\text{row}}, \sigma_{\text{real}}^{\text{col}}\}$. Our aim is to get gap bounds that scale with σ_{real}, not $\sigma_{\max}$.

Example C.1.1 (Simple multiscale Game). For the purpose of our experiment, we will construct our multiscale payoff matrices following the template

$$A = \begin{bmatrix} B & -\mathbf{1}\mathbf{1}^\top \\ \mathbf{1}\mathbf{1}^\top & C \end{bmatrix}$$

where B_{ij} are i.i.d. Rademacher $\{\pm 1\}$ and C_{ij} are i.i.d. Rademacher $\{\pm\sigma_{\max}\}$ for some pre-specified $\sigma_{\max} \gg 1$. By construction, any saddle point for the submatrix B is (upon padding with zeros) also a saddle point for the full matrix A. Moreover, it is a strict saddle point for A if it is a strict saddle point for B with value $\min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\top B\boldsymbol{q} \in (\pm 1)$. We will assume throughout that we are in this latter strict case. Here $\sigma_{\text{real}} = 1$ regardless of $\sigma_{\max}$.

C.1.4. What can one hope to achieve?

Throughout the remainder we assume for simplicity that the saddle point $\boldsymbol{p}_*, \boldsymbol{q}_*$ of the payoff matrix A is unique (a common situation). We define the *optimality gap* of row i by $\delta^{\text{row}}(i) = (\boldsymbol{e}_i - \boldsymbol{p}_*)^\top A\boldsymbol{q}_* \ge 0$ and of column j by $\delta^{\text{col}}(j) = \boldsymbol{p}_*^\top A(\boldsymbol{q}_* - \boldsymbol{e}_j) \ge 0$. We are interested in scenarios where at least one player has strictly positive optimality gap on the action(s) of largest scale. We will show that multiscale regret bounds allow the learning to accelerate. Moreover, the learner does not need to know about this structure and will adapt automatically.

Let us assume without loss of generality that $\delta^{\text{row}}(k) > 0$ while $\sigma_k^{\text{row}} = \max_i \sigma_i^{\text{row}}$ where $\sigma_i^{\text{row}} = \max_j |A_{i,j}|$. The general idea now is to use that $\bar{\boldsymbol{p}}_T \to \boldsymbol{p}_*$. This means that from some point t on,

$$\max_j \bar{\boldsymbol{p}}_t^\top A \boldsymbol{e}_j = \max_{j:\boldsymbol{q}_*(j)>0} \bar{\boldsymbol{p}}_t^\top A \boldsymbol{e}_j = \\ \frac{1}{t} \max_{j:\boldsymbol{q}_*(j)>0} \sum_{s\le t} \boldsymbol{p}_s^\top A \boldsymbol{e}_j \le \frac{1}{t} \sum_{s\le t} \boldsymbol{p}_s^\top A \boldsymbol{q}_s + \max_{j:\boldsymbol{q}_*(j)>0} \frac{1}{t} R_t^{\text{col}}(j)$$

A similar argument for the row player then allows us to conclude

$$\begin{aligned} \text{gap}(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{q}}_t) &\le \frac{1}{t}\left(\sum_{s\le t} \boldsymbol{p}_s^\top A \boldsymbol{q}_s + \max_{j:\boldsymbol{q}_*(j)>0} R_t^{\text{col}}(j) - \sum_{s\le t} \boldsymbol{p}_s^\top A \boldsymbol{q}_s + \max_{i:\boldsymbol{p}_*(i)>0} R_t^{\text{row}}(i) \right) \\ &= \frac{1}{t}\left(\max_{j:\boldsymbol{q}_*(j)>0} R_t^{\text{col}}(j) + \max_{i:\boldsymbol{p}_*(i)>0} R_t^{\text{row}}(i) \right). \end{aligned}$$

The main point is that this bound scales with $\max_{i:\boldsymbol{p}_*(i)>0} \sigma_i^{\text{row}} + \max_{j:\boldsymbol{q}_*(j)>0} \sigma_j^{\text{col}}$ and not with the respective unconstrained maxima.

Proposition C.1.2. *Any pair of multiscale online learning algorithms with bounds of order $R_t^i \le O(\sigma_i \sqrt{T})$ ensures iterate average gap*

$$\text{gap}(\bar{p}_t, \bar{q}_t) \;=\; O(\sigma_{\text{real}}/\sqrt{t})$$

as $t \to \infty$. In particular, this holds for Muscada *with Tuning 3 (see Lemma C.2.1).*

Note that single-scale algorithms would only deliver $\text{gap}(\bar{p}_t, \bar{q}_t) = O(\sigma_{\max}/\sqrt{t})$; a weaker guarantee.

C.1.5. Why our approach may achieve the hope optimistically

Rakhlin and Sridharan [2013] show that using optimism in saddle point interactions can improve the rate to $O(\sigma_{\max}/t)$. We first show that this is true for Muscada as well, after which we will investigate achieving $O(\sigma_{\text{real}}/t)$. The mechanism for this proof is to show that the social regret is constant. Technically, one would explicitly keep track of the slack in (C.5) and (C.6), and use these harvested slacks to cancel the $\sqrt{t}$ term of the regret bound. Only the constant-order term measuring the entropy of the initial weights remains. For this to be a constant, we further need that the learning rate stops decreasing once the regret stabilizes. Following exactly the steps of Rakhlin and Sridharan [2013], we can prove the following proposition.

Proposition C.1.3. *For same-scale games, the optimistic version (see Figure 4.3) of* Muscada *with Tuning 3 and uniform prior (see Lemma C.2.1) achieves average iterate gap $\text{gap}(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{q}}_t) = O(\sigma_{\max}/t)$ as $t \to \infty$.*

The same-scale assumption makes all $\boldsymbol{\sigma}$ equal, while the uniform-prior assumption in addition makes all $\boldsymbol{\eta}$ equal. This makes the standard argument from the literature apply.

We further forward the natural conjecture that we state next.

Conjecture C.1.4. For the multiscale case, the optimistic version (see Figure 4.3) of MUSCADA with Tuning 3 (see Lemma C.2.1) and any nondegenerate prior achieves average iterate gap bounded by $gap(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{q}}_t) = O(\sigma_{\text{real}}/t)$.

The reason that our Tuning 3 has any chance here is that *no* terms (not even the additive constant) in the regret bound scale with $\sigma_{\max}$. This in contrast to the algorithms of Foster et al. [2017], Cutkosky and Orabona [2018], Bubeck et al. [2019], Chen et al. [2021], whose existing multiscale analyses all result in a lower-order term scaling with $\sigma_{\max}$.[1] We next provide empirical support for our conjecture.

C.1.6. Numerical results

We investigate three algorithms: Hedge with classic time-decreasing learning rate $\eta_t = \sqrt{\frac{\ln(K)}{\sigma_{\max}^2 t}}$, MUSCADA with all scales set to $\sigma_{\max}$ and MUSCADA with actual knowledge of the multiscale vectors. All algorithms are run in optimistic mode with guesses $\boldsymbol{m}_t = \ell_{t-1}$, the loss vector of the previous round (and $\boldsymbol{m}_{1,k} = 0$). We choose a matrix of structure given in Example C.1.1, with B and C of size 10×10, and pick $\sigma_{\max} = 100$. We give all algorithms the uniform prior $\pi_k = 1/20$. The results are displayed in Figure C.1, where we show the saddle point gap for the average iterate, the last iterate and the theoretical regret bounds that we obtain from the analysis. In the main text, Figure 4.4(right) shows only the saddle point gap for the average iterate of optimistic MUSCADA with the optimistic modification of Tuning 3 from Figure C.2. Generating this figure with the code from the supplementary material takes 30 minutes on an Intel i7-7700 processor. Memory usage is negligible.

We see in Figure C.1 that the gap of optimistic Hedge decays at the slow rate $O(\sigma_{\max}/\sqrt{t})$. This means that optimism alone is insufficient to obtain a faster $O(\sigma_{\max}/t)$ convergence rate; it is also necessary that the learning rates stop decreasing when the regret plateaus. It is also apparent that MUSCADA tuned to $\sigma_{\max}$ has the fast $O(1/t)$ rate, but at the $\sigma_{\max}$ scale. Finally, the numerical experiments show evidence that our multiscale algorithm does exploit the small scale of the actions in the support of the saddle point, exhibiting the desired $O(\sigma_{\text{real}}/t)$ regret conjectured above. The plot also includes the quality of the last iterate. Hsieh et al. [2021] prove convergence of the last iterate for the common scale, common prior case. In our experiment the iterate average can be seen to converge quickly in the multiscale case, but convergence is terribly slow in the same-scale case. This is not inconsistent; no rates are currently known for the last iterate.

C.2. Tuning 3

In this section we describe a third tuning, defined in Figure C.2. In contrast to Tunings 1 and 2, the learning rates in Tuning 3 start *higher*, namely at $1/(2\sigma_k)$ instead of $1/(2\sigma_{\max})$. The downside of this aggressive tuning is that the variance bound is not available (though the weaker, uncentered second-moment analog is). The upside

[1] Which is hard to spot in some of the literature because of a global $\sigma_{\max} = 1$ convention.

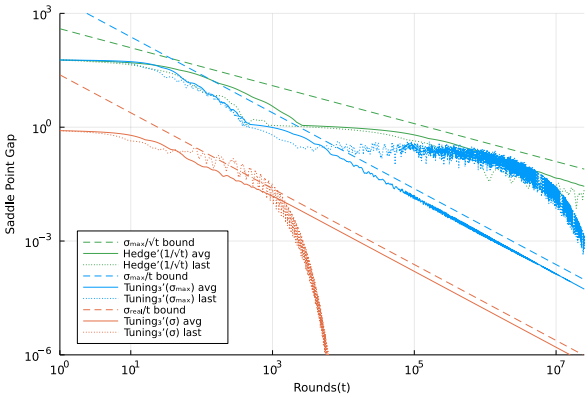

Figure C.1.: Quality of average iterate (solid) and last iterate (dotted) for three optimistic algorithms, compared to their relevant bounds (dashed). The multiscale-aware algorithm (red) outperforms the non-scale-aware competitors by the factor $\sigma_{\max}/\sigma_{\text{real}} = 100$. See Section C.1.6 for further discussion.

Tuning 3 $u = \pi \frac{\sigma_{\min}}{\sigma_k}$, $\eta_{0,k} = \frac{1}{2\sigma_k}$, $\gamma = 8\ln(1/u_k)$, and

$$H_{1,k}(v_t) = \frac{\mathrm{d}}{\mathrm{d}v_t}\left[\frac{v_t}{\sqrt{1+v_t/\gamma_k}}\right] = \frac{v_t/\gamma_k + 2}{2(1+v_t/\gamma_k)^{3/2}}.$$

Figure C.2.: Tuning 3 for MUSCADA

is that the resulting regret bound compared to expert k features only σ_k and has *no* occurrence of $\sigma_{\max}$ whatsoever, not even in the additive constants.

Lemma C.2.1. Let $\boldsymbol{\pi}$ be a probability distribution on K experts. MUSCADA run with Tuning 3 depicted in Figure C.2 guarantees that, for any $t = 1, 2, \ldots,$

$$R_{t,k} \le 2\sigma_k\sqrt{2v_t\ln(1/u_k)} + c_{\boldsymbol{\sigma},\boldsymbol{\pi}}\sigma_{\min}\sqrt{2v_t} + 8\sigma_k\ln(1/u_k) + 4\sigma_{\min} + \frac{\sigma_k}{2}\max_{s\le t}\Delta v_s, \quad \text{(C.1)}$$

where $c_{\boldsymbol{\sigma},\boldsymbol{\pi}} = \sum_{k\in K}\pi_k(1/\sqrt{\ln(1/u_k)})$ and $u_k = \pi_k\frac{\sigma_{\min}}{\sigma_k}$. Additionally, we have that $v_t \le 4\sum_{s\le t}\frac{\langle\tilde{\boldsymbol{w}}_s,\ell_s^2\rangle}{\langle\tilde{\boldsymbol{w}}_s,\boldsymbol{\sigma}^2\rangle} \le 4t$, where, for each $t = 1, 2, \ldots,$ the weights are $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$.

Proof. Follow the same steps as in the proof of the regret bound for Tuning 1 in

Lemma 4.2.3. Obtain that

$$\mu_{t,k} \le \sigma_k \sqrt{2v_t \ln(1/u_k)} + 4\sigma_k \ln(1/u_k) \tag{C.2}$$

$$\frac{\ln(1/u_k)}{\eta_{t,k}} \le \sigma_k \sqrt{2v_t \ln(1/u_k)} + 4\sigma_k \ln(1/u_k), \text{ and} \tag{C.3}$$

$$\sum_{k \in K} \frac{u_k}{\eta_k} \le c_{\boldsymbol{\sigma},\boldsymbol{\pi}} \sigma_{\min} \sqrt{2v_t} + 4\sigma_{\min} \tag{C.4}$$

with $c_{\boldsymbol{\sigma},\boldsymbol{\pi}} = \sum_{k \in K} \pi_k \left(\frac{1}{\sqrt{\ln(1/u_k)}} \right)$. Use Proposition 4.2.3 to conclude the first claim. For the additional claim, use Lemma C.7.2 with $\lambda = 0$. □

C.3. Algorithm Analysis

The only step in the algorithm that may be problematic is the definition of Δv_t at every round, which one might think can take infinite values. We show in Proposition C.7.1 that this is not the case and that consequently $t \mapsto v_t$ is well defined.

C.3.1. Untuned regret bound, proof of Proposition 4.2.2

We prove that the potential $t \mapsto \Phi_t$ is decreasing for optimistic MUSCADA. The result for the nonoptimistic version follows by setting the guesses $\boldsymbol{m}_t$ to $\mathbf{0}$. Recall from (4.4) in Section 4.2 that the potential Φ_t is defined by

$$\Phi_t \;=\; \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_t) \;=\; \max_{\boldsymbol{w} \in \mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{R}_t - \boldsymbol{\mu}_t \rangle - D_{\boldsymbol{\eta}_t}(\boldsymbol{w}, \boldsymbol{u}).$$

Proof of Lemma 4.2.1. We prove the result in the optimistic case. The nonoptimistic case is recovered for $\boldsymbol{m}_t = \mathbf{0}$ and replacing $4\sigma_k^2$, which is a bound on $|m_{t,k} - \ell_{t,k}|^2$, by σ_k^2, which bounds $|\ell_{t,k}|^2$. The result is a consequence of the following inequalities:

$$\begin{aligned}
\Phi_t &\le \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_{t-1}) && \boldsymbol{\eta} \mapsto D_{\boldsymbol{\eta}} \text{ decr.} && \text{(C.5)}\\
&= \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_{t-1} - 4\boldsymbol{\eta}_{t-1}\boldsymbol{\sigma}^2 \Delta v_t, \boldsymbol{\eta}_{t-1}) && \text{by def. of } \boldsymbol{\mu}_t \\
&= \Phi(\boldsymbol{R}_{t-1} + \langle \boldsymbol{w}_t, \boldsymbol{\mu}_t \rangle - \boldsymbol{m}_t - \boldsymbol{\mu}_{t-1}, \boldsymbol{\eta}_{t-1}) && \text{by def. of } \Delta v_t \\
&= \max_{\boldsymbol{w} \in \mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{R}_{t-1} + \langle \boldsymbol{w}_t, \boldsymbol{m}_t \rangle - \boldsymbol{m}_t - \boldsymbol{\mu}_{t-1} \rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}, \boldsymbol{u}) && \text{by def. of } \Phi \\
&= \langle \boldsymbol{w}_t, \boldsymbol{R}_{t-1} + \langle \boldsymbol{w}_t, \boldsymbol{m}_t \rangle - \boldsymbol{m}_t - \boldsymbol{\mu}_{t-1} \rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}_t, \boldsymbol{u}) && \text{by def. of } \boldsymbol{w}_t \\
&= \langle \boldsymbol{w}_t, \boldsymbol{R}_{t-1} - \boldsymbol{\mu}_{t-1} \rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}_t, \boldsymbol{u}) && \langle \boldsymbol{w}_t, \boldsymbol{m}_t \rangle \text{ cancels} \\
&\le \max_{\boldsymbol{w} \in \mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{R}_{t-1} - \boldsymbol{\mu}_{t-1} \rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}, \boldsymbol{u}) && \text{since } \boldsymbol{w}_t \in \mathcal{P}(K) \\
& && && \text{(C.6)}\\
&= \Phi(\boldsymbol{R}_{t-1} - \boldsymbol{\mu}_{t-1}, \boldsymbol{\eta}_{t-1}) \;=\; \Phi_{t-1} && \text{by def. of } \Phi,\ \Phi_t.
\end{aligned}$$

Hence, $\Phi_t \le \Phi_{t-1}$, as we were to show. □

Proof of Proposition 4.2.2. Lemma 4.2.1 shows that the potential $t \mapsto \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_t)$ is decreasing in t and that consequently $\Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_t) \le \Phi(\boldsymbol{R}_0 - \boldsymbol{\mu}_0, \boldsymbol{\eta}_0) = -D_{\boldsymbol{\eta}_0}(\boldsymbol{w}_1, \boldsymbol{u})$. The maximal nature of the definition of Φ implies that, for any probability distribution $\boldsymbol{p} \in \mathcal{P}(K)$,

$$\langle \boldsymbol{p}, \boldsymbol{R}_t \rangle \le \langle \boldsymbol{p}, \boldsymbol{\mu}_t \rangle + D_{\boldsymbol{\eta}_t}(\boldsymbol{p}, \boldsymbol{u}) - D_{\boldsymbol{\eta}_0}(\boldsymbol{w}_1, \boldsymbol{u}). \tag{C.7}$$

The second claim contained in (4.8) follows from the special case where $\boldsymbol{p} = \boldsymbol{\delta}_k$, the probability distribution that puts all of its mass on expert k, and by bounding the last term in (C.7) by zero. The last statement contained in (4.9) is proven in Lemma C.6.3. This is all that we had set ourselves to prove. □

C.3.2. Tuning, proof of Proposition 4.2.3

Proof of Proposition 4.2.3. The main tool that is employed here to derive the regret bounds is Proposition 4.2.2. The fact that the learning rates at hand are decreasing is a consequence of Lemma C.6.4; we give more details in the following. A slightly stronger result than what we claim could be obtained by replacing directly the learning rates in Proposition 4.2.2. However, the result is not amenable to an easy interpretation, and we use upper bounds on the learning rates and their reciprocals. Recall that $\gamma_k = 8\frac{\sigma_{\max}^2}{\sigma_k^2}\ln(1/u_k)$. The learning rate is of the form $\eta_{t,k} = \eta_{0,k} H_{1,k}(v_t) = \eta_{0,k} h(v_t/\gamma_k)$ with $h(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[\sqrt{\frac{x^2}{1+x}}\right] = \frac{x+2}{2(1+x)^{3/2}}$ and $\eta_{0,k} = 1/(2\sigma_{\max})$. That this choice of learning rate is indeed nondecreasing can be proven using Lemma C.6.4. We use the following two elementary inequalities in relation to this specific choice of function h.

Lemma C.3.1. Let $x \ge 0$. The function $h(x) = \frac{x+2}{2(1+x)^{3/2}}$ satisfies

$$\int_0^x h(x')\mathrm{d}x' \le \min\left\{x, \sqrt{x}\right\} \le \max\left\{1, \sqrt{x}\right\}, \quad \text{and} \tag{C.8}$$

$$\frac{1}{h(x)} \le \begin{Bmatrix} 1+x & \text{if } x \le 1 \\ 2\sqrt{x} & \text{if } x > 1 \end{Bmatrix} \le 2\max\left\{1, \sqrt{x}\right\}, \tag{C.9}$$

where the first minimum is equalized at $x = 1$.

Using these upper bounds and the choice $u_k = \pi_k \frac{\sigma_{\min}}{\sigma_k}$ in Proposition 4.2.2 gives the claimed result. Indeed, recall that Proposition 4.2.2 implies that

$$R_{t,k} \le \sigma_k^2 \eta_{0,k} \int_0^{v_t} h(x/\gamma_k)\mathrm{d}x + \frac{\ln(1/u_k)}{\eta_{t,k}} + \sum_{j \in K} \frac{u_j}{\eta_{t,j}} + \sigma_k^2 \eta_{0,k} \max_{s \le t} \Delta v_s. \tag{C.10}$$

We now focus on bounding each term. First,

$$\int_0^{v_t} h(x/\gamma_k)\mathrm{d}x = \gamma_k \int_0^{v_t/\gamma_k} h(x')\mathrm{d}x' \le \max\left\{\gamma_k, \sqrt{v_t \gamma_k}\right\}.$$

Consequently,

$$\sigma_k^2 \eta_{0,k} \int_0^{v_t} h(x/\gamma_k)\mathrm{d}x \le \sigma_k \sqrt{2 v_t \ln(1/u_k)} + 4\sigma_{\max}\ln(1/u_k). \tag{C.11}$$

Next,

$$\frac{1}{\eta_k} = \frac{2\sigma_{\max}}{h(v/\gamma_k)} \leq 4\sigma_{\max} \max\left\{1, \sqrt{\frac{v_t}{\gamma_k}}\right\} \leq 4\sigma_{\max} + \sigma_k \sqrt{\frac{2v_t}{\ln(1/u_k)}}.$$

With this at hand, the second and third term on the right hand side of (C.10) can be bounded by

$$\frac{\ln(1/u_k)}{\eta_{t,k}} \leq \sigma_k \sqrt{2v_t \ln(1/u_k)} + 4\sigma_{\max} \ln(1/u_k), \text{ and} \tag{C.12}$$

$$\sum_{j \in K} \frac{u_j}{\eta_j} \leq c_{\boldsymbol{\sigma},\boldsymbol{\pi}} \sigma_{\min} \sqrt{2v_t} + 4\sigma_{\max} \tag{C.13}$$

with $c_{\boldsymbol{\sigma},\boldsymbol{\pi}} = \sum_{k \in K} \pi_k \left(\frac{1}{\sqrt{\ln(1/u_k)}}\right)$. Replace (C.11), (C.12), and (C.13) in the the regret bound (C.10) to obtain the result. In order to prove the second claim we follow a similar path; we use Proposition 4.2.2 as our main tool. Recall that in this case the learning rate is of the form $\eta_{t,k} = \eta_{0,k} H_{2,k}(v_t)$ with $\eta_{0,k} = 1/(2\sigma_{\max})$ and

$$H_{2,k}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[\sqrt{\alpha_k^2 \left\{\left(1 + \frac{x}{\alpha_k}\right)\ln\left(1 + \frac{x}{\alpha_k}\right) - \frac{x}{\alpha_k}\right\} + \frac{x^2}{2(1 + x/(2\gamma_k))}}\right]$$

with $\alpha_k = 32\frac{\sigma_{\max}^2}{\sigma_k^2}$ and $\gamma_k = \alpha_k \ln(1/\pi_k)$. The fact that $k \mapsto H_{2,k}(x)$ is decreasing follows from Lemma C.6.4 after performing the change of variable $x' = x/\alpha_k$. We use the inequalities for $H_{2,k}$ that are proven in the following lemma.

Lemma C.3.2. Let $\beta_k = \ln(1/\pi_k)$. The function $H_{2,k}$ satisfies

$$\int_0^x H_{2,k}(x')\mathrm{d}x' \leq \sqrt{\alpha_k x(\ln(1 + x/\alpha_k) + \beta_k)}, \text{ and} \tag{C.14}$$

$$\frac{1}{H_{2,k}(x)} \leq 2\sqrt{\frac{x/\alpha_k}{\ln(1 + x/\alpha_k)}}\sqrt{1 + \frac{\min\{\beta_k, \frac{1}{2}\frac{x}{\alpha_k}\}}{\ln(1 + x/\alpha_k)}}. \tag{C.15}$$

We can now compute the analogs of (C.11), (C.12), and (C.13) to obtain that

$$\sigma_k^2 \eta_{0,k} \int_0^{v_t} H_{2,k}(x)\mathrm{d}x \leq 2\sigma_k \sqrt{2v_t\left(\ln\left(1 + \tfrac{\sigma_k^2}{32\sigma_{\max}^2} v_t\right) + \ln(1/\pi_k)\right)},$$

$$\frac{\ln(1/\pi_k)}{\eta_{t,k}} \leq \sigma_k \ln(1/\pi_k) \sqrt{\frac{v_t}{2\ln\left(1 + \frac{\sigma_k^2}{32\sigma_{\max}^2} v_t\right)}} \left(1 + \sqrt{\frac{\min\{\ln\left(\frac{1}{\pi_k}\right), \frac{\sigma_k^2}{16\sigma_{\max}^2} v_t\}}{\ln\left(1 + \frac{\sigma_k^2}{32\sigma_{\max}^2} v_t\right)}}\right),$$

$$\sum_{j \in K} \frac{u_j}{\eta_j} \leq \sum_{j \in K} \pi_j \left(\sigma_j \sqrt{\frac{v_t}{2\ln\left(1 + \frac{\sigma_j^2}{32\sigma_{\max}^2} v_t\right)}} \left(1 + \frac{\sqrt{\min\{\ln\left(\frac{1}{\pi_j}\right), \frac{\sigma_j^2}{16\sigma_{\max}^2} v_t\}}}{\sqrt{\ln\left(1 + \frac{\sigma_j^2}{32\sigma_{\max}^2} v_t\right)}}\right)\right),$$

and employ them in Proposition 4.2.2 to obtain the result. □

Proof of Lemma C.3.1. The relations are clear for $x = 0$. Let $x > 0$. Recall that $\int_0^x h(x')\mathrm{d}x' = \frac{x}{\sqrt{1+x}}$. We start by proving (C.8). The fact that $\frac{x}{\sqrt{1+x}} \le x$ is clear. The inequality $\frac{x}{\sqrt{1+x}} \le \sqrt{x}$ follows from dividing both sides of the inequality $x \le \sqrt{x^2+x}$ by $\sqrt{1+x}$. Thus, the first inequality in (C.8) follows, and the second is direct after observing that $x \le \sqrt{x} \le 1$ for $x \le 1$. We now turn to proving (C.9). Recall that $1/h(x) = \frac{2(1+x)^{3/2}}{2+x}$. We start by showing that $1/h(x) \le 1+x$ for all $x > 0$. Note that $\frac{2(1+x)^{3/2}}{2+x} = (1+x)\frac{2\sqrt{1+x}}{2+x}$. Thus, the claim holds if and only if $2\sqrt{1+x} \le 2+x$, which is easily checked to be the case. Now let $x > 1$. Observe that the second claim in the first inequality holds if and only if $2(1+x)^{3/2} \le 2\sqrt{x}(2+x)$. Square both members and rearrange to conclude that the sought relation holds if and only if $0 \le 4x^2 + 4x - 4$, which is the case as $x > 1$. The second inequality in (C.9) is clear. □

Proof of Lemma C.3.2. The inequalities contained in (C.14) and (C.15) are a consequence of the fact that

$$\int_0^x H_2(x')\mathrm{d}x' = \sqrt{\alpha^2\left\{\left(1+\frac{x}{\alpha}\right)\ln\left(1+\frac{x}{\alpha}\right) - \frac{x}{\alpha}\right\} + \frac{x^2}{2(1+x/(2\gamma))}}$$

and the inequalities

$$(1+x')\ln(1+x') - x' \le x'\ln(1+x') \quad \text{and} \quad \frac{a^2x'^2}{2(1+x'/(2b))} \le \min\{bx', \tfrac{1}{2}a^2x'^2\},$$

that hold for $x', a, b \ge 0$. From this, (C.14) is immediate once we use the substitutions $x' = x/\alpha$, $a = \alpha$, and $b = \beta$. To prove (C.15), use the same substitution and estimate

$$\begin{aligned}
\frac{1}{H(x')} &= 2\frac{\sqrt{(1+x')\ln(1+x') - x' + \frac{x'^2}{2(1+x'/(2b))}}}{\ln(1+x') + \frac{1}{2}\frac{2x'+x'^2/(2b)}{(1+x'/(2b))^2}} \\
&\le 2\frac{\sqrt{x'\ln(1+x') + \min\{bx', \frac{1}{2}x'^2\}}}{\ln(1+x')} \\
&= 2\sqrt{\frac{x'}{\ln(1+x')}}\sqrt{1+\frac{\min\{b, \frac{1}{2}x'\}}{\ln(1+x')}} \\
&\le 2\sqrt{\frac{x'}{\ln(1+x')}}\left(1+\sqrt{\frac{\min\{b, \frac{1}{2}x'\}}{\ln(1+x')}}\right).
\end{aligned}$$

This is all we set ourselves to prove. □

C.4. Optimism, proof of Proposition 4.4.1

Proof of Proposition 4.4.1. In Lemma 4.2.1 we show that the potential $t \mapsto \Phi_t$ is decreasesing. The rest of the proof is identical to that of Proposition 4.2.3 after multi-

plying all scales by 2. The "furthermore" claim follows from a direct modification of Proposition C.7.1. □

C.5. Luckiness

This appendix contains the proofs of the luckiness results in Section 4.3.

C.5.1. Proof of Theorem 4.3.1

Proof of Theorem 4.3.1. Let $s_t = \sum_{s \le t} \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_s}(\ell_t)}{\langle \tilde{\boldsymbol{w}}_s, \boldsymbol{\sigma}^2 \rangle}$. It is shown in Proposition C.7.1 that v_t can be bounded in terms of s_t. Indeed, in any case $v_t \le 4$, and because the learning rates are low enough at the start of the protocol, namely $\eta_{t,k} \le 1/(2\sigma_{\max})$, the upper bound $v_t \le 4s_t$ also holds. A verification of the regret bound obtained in Proposition 4.2.2 shows that it is increasing in v_t, and consequently the same regret bound holds once we replace v_t with the larger quantity $4s_t$, and the proof of Proposition 4.2.3 can be repeated with no problems. Consequently the regret bounds in Proposition 4.2.3 are available with $4s_t$ occupying the place of v_t. The next step that we follow is to show that $\mathbf{E}_{\mathbf{P}}[s_t] \lesssim \mathbf{E}_{\mathbf{P}}[R_{t,k^*}]$, which is done in the following lemma.

Lemma C.5.1. Under Massart's condition (see Definition 4.1.3),

$$\mathbf{E}_{\mathbf{P}}[s_t] \le k_{\mathrm{M}} \mathbf{E}_{\mathbf{P}}[R_{t,k^*}],$$

where $k_{\mathrm{M}} = c_{\mathrm{M}} \max_{i,j \in K} \sup_{v \ge 0} \left\{ \frac{\eta_{0,i} H_i(v)}{\eta_{0,j} H_j(v) \sigma_j^2} \right\}$ satisfies

$$k_{\mathrm{M}} \le 2c_{\mathrm{M}} \max_{i,j \in K} \left\{ \frac{1}{\sigma_i \sigma_j} \frac{\ln(1/\pi_i) + \ln(\sigma_i/\sigma_{\min})}{\ln(1/\pi_j) + \ln(\sigma_j/\sigma_{\min})} \right\}.$$

From the previous discussion, a small modification of Proposition 4.2.3 shows that this tuning guarantees a regret bound of the form

$$R_{t,k^*} \le a' \sqrt{s_t} + b \tag{C.16}$$

with $a' = 4\sigma_{k^*} \sqrt{2 \ln(1/u_{k^*})} + 2\sqrt{2} c_{\boldsymbol{\sigma},\boldsymbol{\pi}} \sigma_{\min}$ and $b = 8\sigma_{\max} \ln(1/u_{k^*}) + 4\sigma_{\max} + 2\sigma_{k^*}$. Take $\mathbf{P}$-expectations in the last display, use the concavity of $x \mapsto \sqrt{x}$ to invoke Jensen's inequality, and use Lemma C.5.1 to obtain that

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \le a' \sqrt{k_{\mathrm{M}} \mathbf{E}_{\mathbf{P}}[R_{t,k^*}]} + b. \tag{C.17}$$

This implies that the expected regret satisfies $\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \lesssim 1$, that is, it is constant. Indeed, using Lemma C.5.2 yields that

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \le a'^2 k_{\mathrm{M}} + b. \tag{C.18}$$

The upper bound for k_{M} is contained in Lemma C.5.3. Given the definition of a in the claim, this is what we set ourselves to prove. □

Proof of Lemma C.5.1. Recall that $s_t = \sum_{s\le t} \Delta s_s = \sum_{s\le t} \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_s}(\ell_s)}{\langle \tilde{\boldsymbol{w}}_s, \boldsymbol{\sigma}^2\rangle}$ with the weights $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$. Define $\ell_s^* = \ell_{s,k^*}$ to be the loss of the best expert k^*, and use that the variance $\mathrm{Var}_{\tilde{\boldsymbol{w}}_s}(\ell_s)$ satisfies $\mathrm{Var}_{\tilde{\boldsymbol{w}}_s}(\ell_s) \le \langle \tilde{\boldsymbol{w}}_s, (\ell_s - \ell_s^*)^2\rangle$ to obtain the estimate

$$\Delta s_s \le \frac{\langle \tilde{\boldsymbol{w}}_s, (\ell_s - \ell_s^*)^2\rangle}{\langle \tilde{\boldsymbol{w}}_s, \boldsymbol{\sigma}^2\rangle}.$$

Recall that, under $\mathbf{P}$, the loss vector ℓ_s is assumed to be independent of ℓ_{s-1}. This implies that

$$\begin{aligned}\mathbf{E}_{\mathbf{P}}[\Delta s_s] &\le \sum_{k\in K}\left(\mathbf{E}_{\mathbf{P}}\left[\frac{\tilde{w}_{s,k}}{\langle \tilde{\boldsymbol{w}}_s, \boldsymbol{\sigma}^2\rangle}\right]\mathbf{E}_{\mathbf{P}}\left[(\ell_{s,k} - \ell_s^*)^2\right]\right)\\ &\le c_{\mathrm{M}} \sum_{k\in K}\left(\mathbf{E}_{\mathbf{P}}\left[\frac{\tilde{w}_{s,k}}{\langle \tilde{\boldsymbol{w}}_s, \boldsymbol{\sigma}^2\rangle}\right]\mathbf{E}_{\mathbf{P}}\left[\ell_{s,k} - \ell_s^*\right]\right).\end{aligned}$$

Sum the last display over rounds, and use the fact that the weights $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$ to deduce that

$$\mathbf{E}_{\mathbf{P}}\left[s_t\right] \le c_{\mathrm{M}} \left\| \max_{s\le t}\left\{\frac{\max_{k\in K}\eta_{s-1,k}}{\min_{k\in K}\eta_{s-1,k}\sigma_k^2}\right\}\right\|_\infty \mathbf{E}_{\mathbf{P}}\left[R_{t,k^*}\right],$$

where $\|\cdot\|_\infty$ is the infinity norm w.r.t. $\mathbf{P}$ (recall that $\eta_{t-1,k}$ depend on the random losses ℓ_{t-1}). Since, for any $s = 1, \dots,$ and $k \in K$, the learning rate $\eta_{s-1,k} = \eta_{0,k} H_k(v)$, we can deduce that $c_{\mathrm{M}} \left\|\max_{s\le t}\left\{\frac{\max_{k\in K}\eta_{s-1,k}}{\min_{k\in K}\eta_{s-1,k}\sigma_k^2}\right\}\right\|_\infty \le k_{\mathrm{M}}$, where k_{M} is as defined in the claim of the proposition. This implies what we set ourselves to prove. □

Lemma C.5.2. Let $y, a, b \ge 0$. If $y^2 \le ay + b$ then $y \le b + \sqrt{a}$.

Proof. The quadratic polynomial $y^2 - ay - b$ has a zero at $y^* = \frac{b+\sqrt{b^2+4a}}{2} \le b + \sqrt{a}$. Hence, if $y^2 \le ay + b$, then $y \le y^*$, and the result follows. □

C.5.2. Proof of Theorem 4.3.2

Proof of Theorem 4.3.2. Call $\Delta_{t,k} = L_{s,k} - L_{s,k^*}$, and $d_k = \mathbf{E}_{\mathbf{P}}[\Delta_{t,k}]$. Since ℓ_s and ℓ_{s-1} are independent, the expected value of the increment of the regret R_{t,k^*} is

$$\begin{aligned}\mathbf{E}_{\mathbf{P}}[\Delta R_{t,k^*}] &= \sum_{k\ne k^*} \mathbf{E}_{\mathbf{P}}[w_{t,k}]\mathbf{E}_{\mathbf{P}}[\ell_{t,k} - \ell_{t,k^*}] &\text{(C.19)}\\ &= \sum_{k\ne k^*} \mathbf{E}_{\mathbf{P}}[w_{t,k}] d_k. &\text{(C.20)}\end{aligned}$$

We seek to prove that for $k \ne k^*$, in an event $\Omega_{t,k}$ which we define next, the weight $w_{t,k}$ is small. Define, for each $k \ne k^*$ and $t \ge 1$, the event $\Omega_{t,k}$ by

$$\Omega_{t,k} = \left\{L_{t,k^*} - L_{t,k} \le \mu_{t,k} - \mu_{t,k^*} - \frac{1}{\eta_{t,k^*}}\ln\left(1/\pi_{k^*}\right) - \frac{1}{\eta_{t,k}}\ln\left(\frac{1}{\pi_k \varepsilon_t}\right)\right\},$$

for deterministic constants $\varepsilon_t = 1/t^2$. Recall that the weights have the form $w_{t,k} = \pi_k \mathrm{e}^{-\eta_{t,k}(L_{t,k}+\mu_{t,k}+a_t^*)}$, where a_t^* is such that $\sum_k w_{t,k} = 1$. Next, we show that, in each event $\Omega_{t,k}$, for carefully chosen $\tilde{a}_t = -\frac{1}{\eta_{t,k^*}} \ln(1/\pi_{k^*}) - L_{t,k^*} - \mu_{t,k^*}$, it holds that $a_t^* \geq \tilde{a}_t$. Indeed, this follows because, by design $\pi_{k^*}\mathrm{e}^{-\eta_{t,k}(L_{t,k}+\mu_{t,k}+\tilde{a}_t)} = 1$, and consequently,

$$\sum_{k\in K} \pi_k(\mathrm{e}^{-\eta_{t,k}(L_t+\mu_{t,k}+\tilde{a}_{t,k})}) \geq 1 = \sum_{k\in K} \pi_k(\mathrm{e}^{-\eta_{t,k}\left(L_t+\mu_{t,k}+a^*_{t,k}\right)}),$$

which implies $a_t^* \geq \tilde{a}_t$. We use this in the weight $w_{t,k}$ of expert k to conclude that

$$\mathbf{E}_{\mathbf{P}}[w_{t,k}\mathbf{1}\{\Omega_{t,k}\}] \leq \pi_k(\mathrm{e}^{-\eta_{t,k}(L_t+\mu_{t,k}+\tilde{a}_{t,k})}) = \pi_k\varepsilon_t,$$

hence

$$\mathbf{E}_{\mathbf{P}}[w_{t,k}] \leq \pi_k\varepsilon_t + \mathbf{P}\left\{\Omega_{t,k}^c\right\}.$$

Consequently, using (C.20),

$$\begin{aligned}
\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] &= \sum_{s\leq t}\sum_{k\neq k^*} d_k \mathbf{E}_{\mathbf{P}}[w_{t,k}]d_k && \text{(C.21)}\\
&\leq \sum_{s\leq t}\sum_{k\neq k^*} \{\pi_k d_k \varepsilon_s + d_k \mathbf{P}\{\Omega_{s,k}\}\} && \text{(C.22)}\\
&\leq 2\sum_{k\in K} \pi_k d_k + \sum_{s\leq t}\sum_{k\neq k^*} d_k \mathbf{P}\{\Omega_{s,k}^c\}, && \text{(C.23)}
\end{aligned}$$

where we used that $\sum_s \varepsilon_s \leq \pi^2/6 \leq 2$. We now focus on bounding the probabilities $\mathbf{P}\{\Omega_{s,k}^c\}$. We use that $\Delta\mu_{t,k} \geq 0$ to deduce that

$$\begin{aligned}
\mathbf{P}\left\{\Omega_{t,k}^c\right\} &= \mathbf{P}\left\{L_{t,k^*} - L_{t,k} > \mu_{t,k} - \mu_{t,k^*} - \frac{1}{\eta_{t,k^*}}\ln(1/\pi_{k^*}) - \frac{1}{\eta_{t,k}}\ln(1/\varepsilon_t)\right\}\\
&\leq \mathbf{P}\left\{L_{t,k^*} - L_{t,k} > -\mu_{t,k^*} - \frac{1}{\eta_{t,k^*}}\ln(1/\pi_{k^*}) - \frac{1}{\eta_{t,k}}\ln(1/\varepsilon_t)\right\}.
\end{aligned}$$

In order to continue, we derive an upper bound on μ_{t,k^*}, and a lower bound on $\eta_{t,k}$, and η_{t,k^*} consisting of deterministic functions of time. Recall from Lemma C.7.1 that $v_t \leq 4t$, and that Lemma C.3.2 can be used to bound μ_{t,k^*} in terms of the integral of the function $x \mapsto H_{2,k^*}(x)$ (see proof of Proposition 4.2.3) to obtain that

$$\begin{aligned}
\mu_{t,k^*} &\leq \sigma_{k^*}^2\eta_{0,k^*}\int_0^{4t} H_{2,k^*}(v)\mathrm{d}v + 4\sigma_{k^*}^2\eta_{0,k^*}\\
&\leq 4\sigma_{k^*}\sqrt{2t(\ln(1+t/8)+\ln(1/\pi_*))} + 4\sigma_{k^*}
\end{aligned}$$

Now fix $k \in K$, and use again that $v_t \leq 4t$ and that $x \mapsto H_{2,k}(x)$ is decreasing (see Lemma C.6.4) to deduce that $\eta_{t,k} = \eta_{0,k}H_k(v_t) \geq \eta_{0,k}H_k(4t)$. From these observations, $\mathbf{P}\left\{\Omega_{t,k}^c\right\}$ can be further bounded by

$$\mathbf{P}\left\{\Omega_{t,k}^c\right\} \leq \mathbf{P}\left\{L_{t,k^*} - L_{t,k} > -F_k(t)\right\},$$

where $F_k(t) = 4\sigma_{k^*}\sqrt{2t(\ln(1+t/8)+\ln(1/\pi_{k^*}))} + 4\sigma_{k^*} + \frac{\ln(1/\pi_{k^*})}{\eta_{0,k^*}H_{2,k^*}(4t)} + \frac{\ln(1/\varepsilon_t)}{\eta_{0,k}H_{2,k}(4t)}$ is a deterministic function of time. Recall that the gap $d_{\min}$ was defined as $d_{\min} = \min_{k\neq k^*} d_k$ and that is assumed to be strictly positive. Recall that $\Delta_{t,k} = L_{t,k} - L_{t,k^*}$ is the gap in losses between expert k and the best expert k^*. Hoeffding's inequality implies that

$$\begin{aligned}\mathbf{P}\left\{L_{t,k^*} - L_{t,k} > F_k(t)\right\} &= \mathbf{P}\left\{td_k - \Delta_{t,k} > td_k - F_k(t)\right\}\\ &\leq \exp\left(-\frac{t}{2\sigma_{\max}^2}((d_k - F_k(t)/t)_+)^2\right)\\ &= \exp\left(-\frac{td_k^2}{2\sigma_{\max}^2}((1 - F_k(t)/(d_k t))_+)^2\right),\end{aligned}$$

where $x \mapsto (x)_+ = \max\{0, x\}$. We now seek a bound on the point $t_k^\star$ at which $F_k(t_k^\star)/t_k^\star = d_k/2$. For these values $t_k^\star$, we have, using (C.23), that

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \leq 2\sum_{k\in K}\pi_k d_k + \sum_{k\neq k^*}(t_k^\star d_k + \sum_{s\geq t^\star} d_k \mathbf{P}\{\Omega_{t,k}^c\}). \tag{C.24}$$

We now concentrate on bounding $t_k^\star$ and the probability of the event $\Omega_{t,k}^c$ for each k. In the limit that $d_k \to 0$, the time $t_k^\star \to \infty$. A quick computation shows that, as $t \to \infty$, $H_{2,k}(t) \sim \sqrt{\frac{2\ln t}{t}}$, and, in the same limit, $4\sigma_{k^*}\sqrt{2t(\ln(1+t/8)+\ln(1/\pi_*))} + 4\sigma_{k^*} \sim 4\sigma_{k^*}\sqrt{2t\ln t}$. Hence, as $t \to \infty$, the function F_k satisfies $F_k(t) \sim (4\sigma_{k^*} + 2\sigma_{\max})\sqrt{2t\ln t}$. We now give a bound on the solution $x_k^\star$ to the equation $xd_k/2 = (4\sigma_{k^*}+2\sigma_{\max})\sqrt{2x\ln x}$ that holds asymptotically as $d_k \to 0$. Call $c = d_k/(2\sqrt{2}(4\sigma_{k^*} + 2\sigma_{\max}))$. Our equation of interest can be rewritten as $xc^2 = \ln x$. Linearize $x\ln x$ around $x = 2/c^2$, and use its concavity to obtain that $\ln(x) \leq \ln(2/c^2) + (c^2/2)(x - 2/c^2)$. With this estimate at hand, the solution to the simpler, linear equation $xc^2 = \ln(2/c^2) + (c^2/2)(x - 2/c^2)$ is an upper bound on $x_k^\star$. From this discussion it follows that the point $t_k^\star$ of interest satisfies $t_k^\star \leq 2\frac{\ln(1/c^2)}{c^2} - \frac{2}{c^2}$. Hence, as $d_k \to 0$,

$$d_k t_k^\star \leq \frac{2(4\sigma_{k^*} + 2\sigma_{\max})^2}{d_k}\left\{\ln\left(\frac{8(4\sigma_{k^*} + 2\sigma_{\max})^2}{d_k^2}\right) - 1\right\} = O\left(\frac{\sigma_{\max}^2}{d_k}\ln\left(\frac{\sigma_{\max}^2}{d_k^2}\right)\right). \tag{C.25}$$

We deduce that, as $d_k \to 0$, for $t \geq t_k^\star$ and any $k \neq k^*$, the probability $\mathbf{P}\{\Omega_{t,k}^c\} \leq \exp\left(-\frac{t}{8\sigma_{\max}^2}d_k^2\right)$. We sum $\mathbf{P}\{\Omega_{t,k}^c\}$ over rounds to conclude that

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \leq 2\sum_{k\in K}\pi_k d_k + \sum_{k\neq k^*}(t_k^\star d_k + \sum_{s\geq t^\star} d_k \mathbf{P}\{\Omega_{t,k}^c\}). \tag{C.26}$$

We now concentrate on bounding $t_k^\star$ and the probability of the event $\Omega_{t,k}^c$ for each k. In the limit that $d_k \to 0$, the time $t_k \to \infty$. A quick computation shows that, as $t \to \infty$, $H_{2,k}(t) \sim \sqrt{\frac{2\ln t}{t}}$, and, in the same limit, $4\sigma_{k^*}\sqrt{2t(\ln(1+t/8)+\ln(1/\pi_*))} + 4\sigma_{k^*} \sim 4\sigma_{k^*}\sqrt{2t\ln t}$. Hence, as $t \to \infty$, the function F_k satisfies $F_k(t) \sim (4\sigma_{k^*} + 2\sigma_{\max})\sqrt{2t\ln t}$. We know give a bound on the solution $x_k^\star$ to the equation $xd_k/2 =$

$(4\sigma_{k^*} + 2\sigma_{\max})\sqrt{2x\ln x}$ that holds asymptotically as $d_k \to 0$. Call $c = d_k/(2\sqrt{2}(4\sigma_{k^*} + 2\sigma_{\max}))$. Our equation of interest can be rewritten as $xc^2 = \ln x$. Linearize $x\ln x$ around $x = 2/c^2$, and use its concavity to obtain that $\ln(x) \le \ln(2/c^2) + (c^2/2)(x - 2/c^2)$. With this estimate at hand, the solution to the simpler, linear equation $xc^2 = \ln(2/c^2) + (c^2/2)(x - 2/c^2)$ is an upper bound on $x_k^\star$. From this discussion it follows that the point $t_k^\star$ of interest satisfies

$$t_k^\star \le 2\frac{\ln(1/c^2)}{c^2} - \frac{2}{c^2} = O\left(\frac{\sigma_{\max}}{d_k^2}\ln\frac{\sigma_{\max}^2}{d_k^2}\right), \tag{C.27}$$

as $d_k \to 0$. Hence, again, as $d_k \to 0$,

$$\sum_{t\ge t^\star} d_k \mathbf{P}\left\{\Omega_{t,k}^c\right\} \le \sum_{t\ge t^\star} d_k \mathrm{e}^{-td_k^2/(8\sigma_{\max}^2)} \le \frac{d_k}{1 - \mathrm{e}^{-d_k^2/(8\sigma_{\max}^2)}}. \tag{C.28}$$

We use (C.27) and (C.28) in (C.26), and the fact that $d/(1 + \mathrm{e}^{d^2/\sigma^2}) = O(\sigma^2/d)$ as $d \to 0$ to conclude the proof. □

C.5.3. In Lemma C.5.1, k_{M} is bounded

Lemma C.5.3. In Lemma C.5.1, the constant k_{M} is bounded for Tuning 1, shown in Figure 4.2. More precisely,

$$k_{\mathrm{M}} \le 2\max_{i,j\in K}\left\{\frac{1}{\sigma_i\sigma_j}\frac{\ln(1/\pi_i) + \ln(\sigma_i/\sigma_{\min})}{\ln(1/\pi_j) + \ln(\sigma_j/\sigma_{\min})}\right\}.$$

Proof. Recall that in both tunings of the algorithm we use the starting learning rate $\eta_{0,k} = 1/(2\sigma_{\max})$, a constant over the experts. As long as this is the case, the constant of interest k_{M} can be bounded by

$$k_{\mathrm{M}} \le \max_{i,j\in K}\sup_v \frac{H_i(v)}{\sigma_j^2 H_j(v)}. \tag{C.29}$$

Recall from Figure 4.2 that $H_{1,k}(v)$ is defined as $H_{1,k}(v) = \frac{v/\gamma_k + 2}{2(1+v/\gamma_k)^{3/2}}$ with $\gamma_k = 8\frac{\sigma_{\max}^2}{\sigma_k^2}(\ln(1/\pi_k) + \ln(\sigma_k/\sigma_{\min}))$. We can estimate the ratio

$$\begin{aligned}
\frac{H_i(v)}{H_j(v)} &= \frac{v/\gamma_i + 2}{(1+v/\gamma_i)^{3/2}}\frac{(1+v/\gamma_j)^{3/2}}{v/\gamma_j + 2} \\
&\le \frac{2v/\gamma_i + 2}{(1+v/\gamma_i)^{3/2}}\frac{(1+v/\gamma_j)^{3/2}}{v/\gamma_j + 1} \\
&= 2\sqrt{\frac{1+v/\gamma_j}{1+v/\gamma_i}} \\
&\le 2\max\left\{1, \sqrt{\frac{\gamma_i}{\gamma_j}}\right\}.
\end{aligned}$$

Hence

$$k_{\mathrm{M}} \le 2 \max_{i,j \in K} \left\{ \frac{1}{\sigma_i \sigma_j} \frac{\ln(1/\pi_i) + \ln(\sigma_i/\sigma_{\min})}{\ln(1/\pi_j) + \ln(\sigma_j/\sigma_{\min})} \right\},$$

as it was to be shown. □

C.6. Technical Lemmas

In this appendix we gather technical results used in previous sections.

C.6.1. For showing that the potential decreases

Lemma C.6.1. For fixed $\boldsymbol{X}$, the function $\boldsymbol{\eta} \mapsto \Phi(\boldsymbol{X}, \boldsymbol{\eta})$ is increasing, that is, if $\eta_k \le \eta'_k$, then, for fixed X, it holds that $\Phi(\boldsymbol{X}, \boldsymbol{\eta}) \le \Phi(\boldsymbol{X}, \boldsymbol{\eta}')$.

Proof. It follows from the definition of Φ and the fact that, for all $x \ge 0$, the function $x \mapsto -\ln(x) - 1 + x$ is nonnegative. Indeed, for any $w \in \mathcal{P}(K)$, it holds that

$$\begin{aligned} D_{\boldsymbol{\eta}}(\boldsymbol{w}, \boldsymbol{u}) &= \sum_{k \in K} w_k \left(\frac{\ln(w_k/u_k) - (1 - u_k/w_k)}{\eta_k} \right) \\ &\ge \sum_{k \in K} w_k \left(\frac{\ln(w_k/u_k) - (1 - u_k/w_k)}{\eta'_k} \right) \\ &= D_{\boldsymbol{\eta}'}(\boldsymbol{w}, \boldsymbol{u}). \end{aligned}$$

The result follows from the definition of Φ contained in (4.4). □

Lemma C.6.2. Fix vectors $\boldsymbol{X}, \boldsymbol{m} \in \mathbb{R}^K$ and $\boldsymbol{u}, \boldsymbol{\eta} \in \mathbb{R}^K_+$. Let $\boldsymbol{w}$ be the optimum value $\boldsymbol{w} = \arg\max_{\boldsymbol{p} \in \mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X} + \boldsymbol{m} \rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u})$. Then,

$$\Phi(\boldsymbol{X} + \boldsymbol{m} - \langle \boldsymbol{w}, \boldsymbol{m} \rangle, \boldsymbol{\eta}) \le \Phi(\boldsymbol{X}, \boldsymbol{\eta})$$

Proof. The result follows from the chain of inequalities

$$\begin{aligned} \Phi(\boldsymbol{X} + \boldsymbol{m} - \langle \boldsymbol{w}, \boldsymbol{m} \rangle, \boldsymbol{u}) &= \langle \boldsymbol{w}, \boldsymbol{X} + \boldsymbol{m} - \langle \boldsymbol{w}, \boldsymbol{m} \rangle \rangle - D_{\boldsymbol{\eta}}(\boldsymbol{w}, \boldsymbol{u}) \\ &= \langle \boldsymbol{w}, \boldsymbol{X} \rangle - D_{\boldsymbol{\eta}}(\boldsymbol{w}, \boldsymbol{u}) \\ &\le \Phi(\boldsymbol{X}, \boldsymbol{\eta}). \end{aligned}$$

□

C.6.2. For bounding μ with v

The following is the consequence of a standard result in the theory of Riemann integration.

Lemma C.6.3. Let $x \mapsto H(x)$ be a decreasing, positive, real, and continuous function such that $H(x) < \infty$ on $0 \le x < \infty$. If $\Delta v_s \ge 0$ for $s = 1, 2, \ldots, t$ then

$$\sum_{s \le t} H(v_{s-1}) \Delta v_s \le \int_0^{v_t} H(x) \mathrm{d}x + (H(0) - H(v_t)) \max_{s \le t} \Delta v_t,$$

where $v_t = \sum_{s \le t} \Delta v_s$.

Proof. Because H is decreasing and $t \mapsto v_t = \sum_{s \le t} \Delta v_t$ is nondecreasing,

$$\int_0^{v_t} H(x) \mathrm{d}x \ge \sum_{s \le t} H(v_s) \Delta v_s.$$

Use this observation to deduce that

$$\begin{aligned} \sum_{s=1} H(v_s) \Delta v_s - \int_0^{v_t} H(x) \mathrm{d}x &\le \sum_{s \le t} (H(v_{s-1}) - H(v_s)) \Delta v_s \\ &\le (H(0) - H(v_s)) \max_{s \le t} \Delta v_s, \end{aligned}$$

which is what we set ourselves to prove. □

C.6.3. The learning rates decrease

Lemma C.6.4. The functions $f(x) = \frac{x+2}{2(1+x)^{3/2}}$ and $g(x) = \frac{\ln(1+x) + \frac{2x+x^2/a}{(1+x/a)^2}}{\sqrt{(1+x)\ln(1+x) - x + \frac{x^2}{2(1+x/a)}}}$ are decreasing in $x \ge 0$ for any fixed $a > 0$.

Proof. The function f is differentiable in $x \ge 0$, and its derivative is $f'(x) = -\frac{x+4}{(1+x)^{5/2}}$, a negative function. Thus, f is decreasing. We turn our attention to the function g. Let $h_1(x) = \ln(1+x)$, $h_2(x) = \frac{2x+x^2/a}{(1+x/a)^2}$, and let $H_1(x) = \int_0^x h_1(s)\mathrm{d}s = (1+x)\ln(1+x) - x$, and $H_2(x) = \int_0^x h_2(s)\mathrm{d}s = \frac{x^2}{2(1+x/a)}$. Then, the function g is of the form $h/(2\sqrt{H})$ with $h = h_1 + h_2$, and $H = H_1 + H_2$. Since $g(x)$ is differentiable in $x \ge 0$, it is enough to prove that $g' \le 0$. We compute the derivative $g' = \frac{h'(x)\sqrt{H(x)} - h^2(x)/(2\sqrt{H(x)})}{H(x)}$ and conclude that $g' \le 0$ if and only if

$$h'(x)H(x) \le \frac{1}{2}h^2(x). \tag{C.30}$$

Since $h_1/\sqrt{H_1} = \frac{\sqrt{2}}{2} f(x/a)$, the analog of the last display holds for the pair h_1, H_1. We will show that the same holds true for the pair h_2, H_2 at the end of the proof. For now, use that (C.30) holds for both pairs, replace the definition of h and H, and conclude that it is enough to show that

$$h_1' H_2 + h_2' H_1 \le h_1 h_2.$$

We now focus on showing that $\delta^\star = h_1 h_2 - h_1' H_2 - h_2' H_1$ is nonnegative. Define $\delta(x) = (1 + x/a)^3 (x+1) 2a^3 \delta^\star(x)$. It is clear that it is sufficient to our purposes to show that $\delta(x) \geq 0$ for $x \geq 0$. Computation shows that

$$\delta(x) = a^3 x^2 - 2\,a^2 x^3 - a x^4 + 2\,a^3 x + \\ \left((4\,a+1)x^4 + x^5 - 2\,a^3 x + 5\,a^2 x^2 + \left(5\,a^2 + 4\,a\right)x^3 - 2\,a^3\right) \ln(x+1)\,.$$

Since $\delta(0) = 0$, it is enough to show that its derivative is positive; that $\delta'(x) \geq 0$ for $x \geq 0$. Computation shows that

$$\delta'(x) = 2\,a^3 x - a^2 x^2 + x^4 + \\ \left(4\,(4\,a+1)x^3 + 5\,x^4 - 2\,a^3 + 10\,a^2 x + 3\left(5\,a^2 + 4\,a\right)x^2\right) \ln(x+1)\,.$$

We now pay attention to the first three summands of the previous display. We use that $2a^3 x - a^2 x^2 + x^4 = x(2a^3 - a^2 x + x^3) \geq \ln(1+x)(2a^3 - a^2 x + x^3)$, which follows from the fact that last factor of the last equation is a depressed cubic that is nonnegative for $x, a \geq 0$. This fact, the previous display, and a short computation together imply that

$$\frac{\delta'(x)}{\ln(1+x)} \geq (16\,a+5)x^3 + 5\,x^4 + 9\,a^2 x + 3\left(5\,a^2 + 4\,a\right)x^2,$$

which shows that $\delta'(x) \geq 0$ for $x \geq 0$. This in turn shows that the function δ is positive, that consequently the relation (C.30) holds, and finally, that the original function of interest g is decreasing. □

C.6.4. For bounding Δv in terms of Δs

Lemma C.6.5. Let $y, x, b \in \mathbb{R}$ be such that $b \geq 0$, $x \leq b$, and $y > 0$. Let $\varphi = \frac{\mathrm{e}^b - 1 - b}{\frac{1}{2}b^2} \geq 1$. Then the following statements hold.

1. For $g(y) = \frac{\varphi - 1 - \sqrt{(\varphi-1)^2 + 2\varphi y}}{\varphi} - \ln\left(\varphi - \sqrt{(\varphi-1)^2 + 2\varphi y}\right)$, we have

$$\mathrm{e}^{x - g(y)} - 1 - x \leq \frac{1}{2}\varphi x^2 - y$$

   any time that $y \leq \frac{2\varphi - 1}{\varphi}$.

2. Let $c = \varphi/(\varphi - 1)$. For any $0 < s < 1/c$ it holds that

$$\mathrm{e}^{x - s - h(cs)} - 1 - x \leq \frac{1}{2}\varphi x^2 - s,$$

   where

$$h(u) = -u - \ln(1-u) \leq \frac{1}{2}\frac{u^2}{1-u}$$

   for $0 < u < 1$.

Proof. Proving our claim is equivalent to proving that

$$g(z) \geq x - \ln\left(1 - z + x + \frac{1}{2}\varphi x^2\right).$$

The condition that $z < \frac{2\varphi-1}{2\varphi}$ ensures that the logarithm is well defined. The first claim follows because g was chosen as the maximizer over $x \leq b$ of the right hand side of the previous display. Indeed, the maximizer is $-x^\star(z)$ with $x^\star(z) = -\frac{\varphi-1-\sqrt{(\varphi-1)^2+2\varphi z}}{\varphi} \geq 0$. Now we turn to proving the second claim, which will follow from a series of rewritings of the first claim. The previous display can be rewritten as

$$g(z) = -x^\star(z) - \ln(1 - \varphi x^\star(z)).$$

Let $s' = x^\star(z)$ so that $z = \frac{1}{2}\varphi s'^2 + (\varphi - 1)s'$. If we let $h(u) = -u - \ln(1-u)$, the previous display can be rewritten as

$$g(z) = (\varphi - 1)s' + h(\varphi s').$$

In these terms, the first claim that we already proved takes the shape

$$\mathrm{e}^{x-(\varphi-1)s'-h(\varphi s')} - 1 - x \leq \frac{1}{2}\varphi x^2 - (\varphi - 1)s' - \frac{1}{2}\varphi s'^2$$

any time that $s' \leq 1/\varphi$. Define $s = (\varphi-1)s'$. Replace this in the last display and bound the last, negative term by 0 to obtain that, as long as $s \leq \frac{\varphi-1}{\varphi}$,

$$\mathrm{e}^{x-s-h(cs)} - 1 - x \leq \frac{1}{2}\varphi x^2 - s.$$

This is our claim. The additional bound on h is well known and can be proven with a term-wise bound on the Taylor expansion of $u \mapsto -u - \ln(1-u)$. □

C.6.5. Dual formulation of $\Delta\Phi$

Recall from the definitions in Section 4.2 that the Bregman divergence $D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u})$ between $\boldsymbol{p}$ and $\boldsymbol{u}$, two vectors in $\mathbb{R}_+^K$, was defined in (4.3) as

$$D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}) = \sum_{k \in K} p_k \left(\frac{\ln(p_k/u_k) - (p_k - u_k)}{\eta_k}\right);$$

and the corresponding potential Φ, in (4.4) as

$$\Phi(\boldsymbol{X}, \boldsymbol{\eta}) = \sup_{\boldsymbol{p} \in \mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X} \rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}).$$

In the implementation of the algorithm, we rely on the dual formulation of the potential Φ and its change $\Delta\Phi$ between rounds. We compute these in the following two lemmas.

Lemma C.6.6 (Potential difference in dual form). Let $\boldsymbol{X}, \Delta\boldsymbol{X} \in \mathbb{R}^K$ and $\boldsymbol{u}, \boldsymbol{\eta} \in \mathbb{R}_+^K$, $\Delta\Phi = \Phi(\boldsymbol{X} + \Delta\boldsymbol{X}, \boldsymbol{\eta}) - \Phi(\boldsymbol{X}, \boldsymbol{\eta})$, and $\boldsymbol{w} = \arg\max_{\boldsymbol{p}\in\mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X}\rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u})$. Then

$$\Delta\Phi = \inf_{\Delta a\in\mathbb{R}} \sum_{k\in K} w_k \left(\frac{\mathrm{e}^{\eta_k(\Delta X_k - \Delta a)} + \eta_k \Delta a - 1}{\eta_k} \right).$$

Proof. From Lemma C.6.7 we know that

$$w_k = u_k \mathrm{e}^{\eta_k(X_k - a^*)},$$

where a^* is such that $\sum_{k\in K} w_k = 1$, and that

$$\Phi(\boldsymbol{X}, \boldsymbol{\eta}) = a^* + \sum_{k\in K} u_k \left(\frac{\mathrm{e}^{\eta_k(X_k - a^*)} - 1}{\eta_k} \right).$$

Use the same lemma and the change of variable $a = a^* + \Delta a$ to obtain that

$$\Phi(\boldsymbol{X} + \Delta\boldsymbol{X}, \boldsymbol{\eta}) = \inf_{\Delta a\in\mathbb{R}} \left\{ a^* + \Delta a + \sum_{k\in K} u_k \left(\frac{\mathrm{e}^{\eta_k(X_k - a^* + \Delta X_k - \Delta a)} - 1}{\eta_k} \right) \right\}.$$

Substract these two displays and use the explicit expresion for w. In this way, we obtain the result. □

Lemma C.6.7 (Potential Dual). Let $\boldsymbol{X} \in \mathbb{R}^K$ be a vector, and let $\boldsymbol{u}, \boldsymbol{\eta} \in \mathbb{R}_+^K$ be positive vectors. Then

1. The potential Φ satisfies

$$\Phi(\boldsymbol{X}, \boldsymbol{\eta}) = \langle \boldsymbol{p}^*, \boldsymbol{X}\rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}^*, \boldsymbol{u}),$$

   where $p_k^* = u_k \mathrm{e}^{\eta_k(X_k - a^*)}$, and a^* is such that $\sum_{k\in K} p_k^* = 1$.

2. The potential Φ satisfies the identity

$$\Phi(\boldsymbol{X}, \boldsymbol{\eta}) = \inf_{a\in\mathbb{R}} \left\{ a + \sum_{k\in K} u_k \left(\frac{\mathrm{e}^{\eta(X_k - a)} - 1}{\eta_k} \right) \right\}.$$

Proof. Consider the optimization problem

$$\sup_{p\in\mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X}\rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}).$$

Its Lagrangian function is

$$\mathcal{L}(a, \boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{X}\rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}) - a\left(\sum_{k\in K} p_k - 1 \right).$$

The strong duality relation

$$\sup_{\boldsymbol{p}\in\mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X}\rangle + D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}) = \inf_{a\in\mathbb{R}} \sup_{\boldsymbol{p}\in\mathbb{R}^K} \mathcal{L}(a, \boldsymbol{p}) \tag{C.31}$$

holds, and the maximum on the right hand side can be computed by diferentiation. The gradient with respect to $\boldsymbol{p}$ is

$$\nabla_{\boldsymbol{p}} \mathcal{L}_k = X_k - a - \frac{\ln(p_k/u_k)}{\eta_k},$$

which is zero at

$$p_k^* = u_k \mathrm{e}^{\eta_k (X_k - a)}.$$

Replace $\boldsymbol{p}^*$ in the Lagrangian $\mathcal{L}$ to conclude that

$$\mathcal{L}(a, \boldsymbol{p}^*) = a + \sum_{k \in K} u_k \left(\frac{\mathrm{e}^{\eta_k (X_k - a)} - 1}{\eta_k} \right).$$

Replace this in (C.31) to obtain the second claim. For the first claim, differentiate $\inf_{a \in \mathbb{R}} \mathcal{L}(a, p^*)$ with respect to a and equate to 0. □

C.7. Proof of Theorem 4.1.2

Recall that Δv_t is implicitly specified in the definition of MUSCADA, in Figure 4.1. The main intuition driving the result contained in Theorem 4.1.2 stems from a Taylor approximation of the increment of the potential function at round t for small learning rates. The duality computation for the potential increment $\Delta\Phi$ of Lemma C.6.6 implies that, at round t, Δv_t is the value of Δv that satisfies

$$\inf_{\lambda \in \mathbb{R}} \sum_{k \in K} w_{t,k} \left(\frac{\mathrm{e}^{-\eta_{t-1,k}(\ell_{t,k} - \lambda) - \eta_{t-1,k}^2 \sigma_k^2 \Delta v} + \eta_{t-1,k}(\ell_{t,k} - \lambda) - 1}{\eta_{t-1,k}} \right) = 0, \tag{C.32}$$

where, in the notation of Lemma C.6.6, we used $\Delta \boldsymbol{X} = \Delta \boldsymbol{R}_t$ and reparametrized by $\lambda = \langle \boldsymbol{w}_t, \ell_t \rangle - \Delta a$. For small values of η, the Taylor approximation $\mathrm{e}^{\eta x - \eta^2 b} = 1 + \eta x + \frac{1}{2}\eta^2 (x^2 - 2b) + O(\eta^3)$ gives that, if all the learning rates are small, the quantity being minimized in the previous display can be approximated as

$$\begin{aligned} \sum_{k \in K} w_{t,k} \left(\frac{\mathrm{e}^{-\eta_{t-1,k}(\ell_{t,k} - \lambda) - \eta_{t-1,k}^2 \sigma_k^2 \Delta v} + \eta_{t-1,k}(\ell_{t,k} - \lambda) - 1}{\eta_{t-1,k}} \right) \approx \\ \frac{1}{2} \sum_{k \in K} w_{t,k} \eta_{t-1,k} (\ell_{t,k} - \lambda)^2 - \Delta v \sum_{k \in K} w_{t,k} \eta_{t-1,k} \sigma_k^2. \end{aligned} \tag{C.33}$$

If this approximate expression could be plugged into (C.32), we could solve the infimum and obtain that

$$\Delta v_t \approx \frac{1}{2} \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}}(\ell_t)}{\langle \tilde{\boldsymbol{w}}, \boldsymbol{\sigma}^2 \rangle}$$

with $\tilde{w}_{t,k} \propto w_{t,k} \eta_{t-1,k}$. However, this approximation is only valid under range restrictions in the values of λ. This is the subject of Lemma C.7.2, whose main technical ingredient is the inequality obtained in Lemma C.6.5, which contains an estimate that

makes (C.33) precise. We gather theses results in the following proposition. Used with $b = 1$, it implies Theorem 4.1.2 because the learning rates from Figure 4.2 are all smaller than $1/(2\sigma_{\max})$.

Proposition C.7.1. *Fix $t \geq 1$. Let $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$, where $\boldsymbol{w}_t$ are the weights played by* Muscada *at round t, and $\boldsymbol{\eta}_{t-1}$ its learning rates. The following statements hold.*

1. *If $\max_k 2\eta_{t-1,k}\sigma_k \leq b$ and $b \leq 1$, then*

$$\Delta v_t \leq c_0 \frac{\langle \tilde{\boldsymbol{w}}_t, \ell_t^2 \rangle}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \leq c_0, \tag{C.34}$$

   where the constant c_0 satisfies $c_0 \leq 3.1$ and depends only on b.

2. *If $\max_k 2\eta_{t-1,k}\sigma_{\max} \leq b$ for some $b \leq 1$, and*

$$\Delta s_t = \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_t}(\ell_t)}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle},$$

   then

$$\Delta v_t \leq c_1 \Delta s_t + c_2 \Delta s_t^2, \tag{C.35}$$

   and consequently

$$v_t \leq c_3 s_t,$$

   where $c_1 \leq 0.72$, $c_2 \leq 2.4$, and $c_3 = c_1 + c_2 \leq 3.1$ depend on b only.

Proof of Proposition C.7.1. First, we prove 1. Assume that $\max_k 2\eta_{t-1,k}\sigma_k \leq b'$ and that $b' \leq 1$. Our objective is to use Lemma C.7.2 with $\lambda = 0$. To this end, let $\varphi' = \frac{\mathrm{e}^{b'} - b' - 1}{\frac{1}{2}b'^2} \geq 1$, $c_1' = \frac{b'^2\varphi'^2}{8(\varphi'-1)}$, and $c_2' = \frac{\varphi'^4 b'^2}{8(\varphi'-1)^2} - \frac{\varphi'^3 b'^2}{8(\varphi'-1)}$ be as in Lemma C.7.2. Since we assumed that $b \leq 1$, we have that $c_1' \leq 1/2$, and we can conclude that

$$\Delta v_t \leq \frac{\varphi'}{2}\Delta s_{t,0} + \frac{1}{2}\frac{c_2'\Delta s_{t,0}^2}{1 - c_1'\Delta s_{t,0}}$$

with $\Delta s_{t,0} = \frac{\langle \tilde{\boldsymbol{w}}_t, \ell_t^2 \rangle}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \leq 1$. Use this to conclude that

$$\Delta v_t \leq \frac{\varphi'}{2} + \frac{1}{2}\frac{c_2'}{1 - c_1'}.$$

This last display is exactly our first claim once we set $c_0 = \frac{\varphi'}{2} + \frac{1}{2}\frac{c_2'}{1-c_1'}$. The value of c_0' depends monotonically on that of b'. Compute the value of c_0' for $b' = 1$ to confirm that $c_0' \leq 3.1$.

We now turn our attention to the second claim. We proceed in a similar fashion as before. Assume that $\max_k 2\eta_{t-1,k}\sigma_{\max} \leq b$ for some $b \geq 1$. Let φ, c_1, c_2 be defined as before but now in terms of b. Use Lemma C.7.2 to obtain that

$$\Delta v_t \leq \frac{\varphi}{2}\Delta s_t + \frac{1}{2}\frac{c_2\Delta s_t^2}{1 - c_1\Delta s_t}$$

with $\Delta s_t = \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_t}(\ell_t)}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \leq 1$. Use this to conclude that

$$\Delta v_t \leq \frac{\varphi}{2} \Delta s_t + \frac{1}{2} \frac{c_2}{1 - c_1} \Delta s_t^2.$$

This is exactly the second claim up to a redefinition of constants. The "consequenlty" part of the claim follows from the observation that $\Delta s_t^2 \leq \Delta s_t$ and a summation over time. The computation of the upper bound on the constants is similar as before. □

Lemma C.7.2. Let $t \geq 1$, $\lambda \in \mathbb{R}$, and let

$$\Delta s_t = \Delta s_t(\lambda) = \frac{\langle \tilde{\boldsymbol{w}}_t, (\ell_t - \lambda)^2 \rangle}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \tag{C.36}$$

with $\tilde{w}_{t,k} \propto w_{t,k} \eta_{t-1,k}$. Then, if $\max_k \eta_{t-1,k}(\ell_{k,t} - \lambda) \leq b$ and $\max_k (2\eta_{t-1,k}\sigma_k) \leq b$ for some $b \geq 0$, we have that

$$\Delta v_t \leq \frac{\varphi}{2} \Delta s_t + c_1 \Delta v_t \Delta s_t + \frac{1}{2} c_2 \Delta s_t^2, \tag{C.37}$$

where $\varphi = \frac{\mathrm{e}^b - b - 1}{\frac{1}{2} b^2} \geq 1$, $c_1 = \frac{b^2 \varphi^2}{8(\varphi - 1)}$, and $c_2 = \frac{\varphi^4 b^2}{8(\varphi-1)^2} - \frac{\varphi^3 b^2}{8(\varphi-1)}$. If additionally $c_1 \Delta v_t < 1$, then

$$\Delta v_t \leq \frac{\varphi}{2} \Delta s_t + \frac{1}{2} \frac{c_2 \Delta s^2}{1 - c_1 \Delta s_t}. \tag{C.38}$$

Proof. Let $t \geq 1$. First note that if $c_1 \Delta s_t \geq 1$, our claim becomes trivial. We can safely assume that that $c_1 \Delta s_t < 1$. We proceed in the following steps. Use Lemma C.6.6 to express the increase in the potential function $\Delta\Phi_t(\Delta v) = \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_{t-1} - \boldsymbol{\eta}\boldsymbol{\sigma}^2 \Delta v, \boldsymbol{\eta}_{t-1}) - \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_{t-1})$ in dual form as

$$\Delta\Phi_t(\Delta v) = \inf_{\lambda \in \mathbb{R}} \sum_{k \in K} w_{t,k} \left(\frac{\mathrm{e}^{-\eta_{t-1,k}(\ell_{t,k} - \lambda) - \eta_{t-1,k}^2 \sigma_k^2 \Delta v} + \eta_{t-1,k}(\ell_{t,k} - \lambda) - 1}{\eta_{t-1,k}} \right).$$

From now and until the end of the proof, omit the time indexes for readability.

Because of our assumption that $\eta_k |\ell_k - \lambda| \leq b$, Lemma C.6.5 can be used to obtain that

$$\Delta\Phi(\Delta v) \leq \frac{1}{2} \varphi \sum_{k \in K} w_k [\eta_k (\ell_k - \lambda)^2] - \sum_{k \in K} w_k \left(\frac{g^{-1}(\eta_k^2 \sigma_k^2 \Delta v)}{\eta_k} \right)$$

where $g(x) = x + h(cx)$, $h(u) = \frac{1}{2} \frac{u^2}{1-u}$ and $c = \varphi/(\varphi - 1)$. Use the concavity of $x \mapsto g^{-1}(\Delta v x)/x$ and Jensen's inequality to deduce that

$$\begin{aligned} \sum_{k \in K} w_k \left(\frac{g^{-1}(\eta_k^2 \sigma_k^2 \Delta v)}{\eta_k} \right) &= \sum_{k \in K} w_k \left(\eta_k \sigma_k^2 \frac{g^{-1}(\eta_k^2 \sigma_k^2 \Delta v)}{\eta_k^2 \sigma_k^2} \right) \\ &\geq \langle \boldsymbol{w}, \boldsymbol{\eta}\boldsymbol{\sigma}^2 \rangle \frac{g^{-1}\left(\Delta v \langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle \right)}{\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle}, \end{aligned}$$

where we defined $\hat{w}_k \propto w_k \eta_k \sigma_k^2$. This is useful for obtaining the bound

$$\Delta\Phi(\Delta v) \leq \frac{1}{2}\varphi \sum_{k\in K} w_k \left(\eta_k(\ell_k - \lambda)^2\right) - \langle \boldsymbol{w}, \boldsymbol{\eta}\boldsymbol{\sigma}^2\rangle \frac{g^{-1}\left(\Delta v \langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2\boldsymbol{\sigma}^2\rangle\right)}{\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2\boldsymbol{\sigma}^2\rangle}.$$

Consequenlty, $\Delta\Phi(\Delta v^\star) \leq 0$ for

$$\Delta v^\star = \frac{1}{\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2\boldsymbol{\sigma}^2\rangle} g\left(\frac{1}{2}\varphi \langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2\boldsymbol{\sigma}^2\rangle \frac{\langle \tilde{\boldsymbol{w}}, (\ell-\lambda)^2\rangle}{\langle \tilde{\boldsymbol{w}}, \boldsymbol{\sigma}^2\rangle}\right),$$

where $\tilde{w}_k \propto w_k \eta_k$. Use the definition of Δv and the continuity of $\Delta\Phi$ to conclude that $\Delta v \leq \Delta v^\star$. Unpack the definition of g to obtain that

$$\Delta v \leq \frac{1}{2}\varphi\Delta s + \frac{1}{2}\frac{\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2\boldsymbol{\sigma}^2\rangle (c'\Delta s)^2}{1 - c'\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2\boldsymbol{\sigma}^2\rangle \Delta s}$$

with $c' = \frac{1}{2}\frac{\varphi^2}{\varphi-1}$. Next, we will use use that $\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2\boldsymbol{\sigma}^2\rangle \leq \frac{1}{4}b^2$ to bound further Δv. Use this observation and the definition of c_1 to deduce the inequality $\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2\boldsymbol{\sigma}^2\rangle c'\Delta s \leq c_1\Delta s < 1$. Plug this in the previous display and rearrange to obtain the result:

$$\Delta v \leq \frac{1}{2}\varphi\Delta s + \frac{1}{2}\Delta s^2\left(\frac{1}{4}c'^2b^2 - \frac{1}{4}\varphi c' b^2\right) + \frac{1}{4}c'b^2\Delta v\Delta s,$$

exactly what we claimed. □

D. Appendix to Chapter 5

D.1. Proofs for Section 5.2

of Proposition 5.2.5. Since $\exp(\eta Z) \leq \exp(\eta Z_+)$, we have for each $0 < \eta < b$, using also Fubini's theorem and the tail condition on Z

$$\mathbf{E}[\mathrm{e}^{\eta Z} - 1] \leq \mathbf{E}[\mathrm{e}^{\eta Z_+} - 1] = \mathbf{E}[\int_0^{Z_+} \eta \mathrm{e}^{\eta Z} \mathrm{d}z] = \eta \int_0^\infty \mathbf{P}\{Z \geq z\} \mathrm{e}^{\eta z} \mathrm{d}z \leq$$
$$a\eta \int_0^\infty \mathrm{e}^{-(b-\eta)z} \mathrm{d}z = \frac{a\eta}{b-\eta},$$

which means that

$$Z \trianglelefteq_\eta \frac{1}{\eta} \ln\left(1 + \frac{a\eta}{b-\eta}\right) \tag{D.1}$$

For the first claim, pick $0 \leq \eta^* < b$ and call c the right hand side of (D.1) when evaluated at $\eta = \eta^*$. The result follows by item 3 in Proposition 5.2.4, ahead. The converse follows from

$$\mathbf{P}\{X \geq Y + \epsilon\} \leq \mathrm{e}^{\eta(\mathbf{A}^\eta[X-Y]-\epsilon)} \leq \mathrm{e}^{\eta(c-\epsilon)}$$

with $a = \mathrm{e}^{\eta c}$, and $b = \eta$. □

of Proposition 5.2.11. We can write

$$\mathbf{E}[X] = \mathbf{E}[X \mid X \geq 0]\mathbf{P}\{X \geq 0\} + \mathbf{E}[X \mid X < 0]\mathbf{P}\{X < 0\}.$$

We will bound both terms on the right hand side from bellow. For the first one, use Markov's inequality and the definition of conditional expectation to obtain that

$$\mathbf{E}[X \mid X \geq 0]\mathbf{P}\{X \geq 0\} = \mathbf{E}[[X]_+] \geq a\mathbf{P}\{X \geq a\}. \tag{D.2}$$

For the second term, use the conditional version of Jensen's inequality to obtain that

$$\begin{aligned} \mathbf{E}[X \mid X < 0] &\geq -\frac{1}{\eta} \ln \mathbf{E}[e^{-\eta X} \mid X < 0], \\ &\geq -\frac{1}{\eta} \ln \frac{1}{\mathbf{P}\{X < 0\}}. \end{aligned} \tag{D.3}$$

where the last inequality holds because by the assumption that $0 \trianglelefteq_\eta X$, which implies

$$\begin{aligned} 1 \geq \mathbf{E}[\mathrm{e}^{-\eta X}] &= \mathbf{E}[\mathrm{e}^{-\eta X} \mid X \geq 0]\mathbf{P}\{X \geq 0\} + \mathbf{E}[\mathrm{e}^{-\eta X} \mid X < 0]\mathbf{P}\{X < 0\}, \\ &\geq \mathbf{E}[\mathrm{e}^{-\eta X} \mid X < 0]\mathbf{P}\{X < 0\}. \end{aligned}$$

Gathering (D.2) and (D.3) together implies

$$\mathbf{E}[X] \geq a\mathbf{P}\{X \geq a\} - \frac{1}{\eta}\mathbf{P}\{X < 0\} \ln \frac{1}{\mathbf{P}\{X < 0\}},$$

which after rearrangement implies the first inequality. The second inequality follows from maximizing the function $x \mapsto x \ln(1/x)$, which is a concave function that attains its maximum value 1/e at $x^* = 1/\mathrm{e}$. □

D.2. Proofs for Section 5.3.1

Proposition D.2.1 below strictly strengthens Proposition 5.3.1. To see how, note that (1) ⇒ (2) in Proposition 5.3.1 is implied by 1. below, noting that in Proposition 5.3.1 we assume that $\{X_f\}_{f\in\mathcal{F}}$ is regular, which implies the condition of (1). (2) ⇒ (3) in Proposition 5.3.1 is implied by 2. below, again since in Proposition 5.3.1 we assume that $\{X_f\}_{f\in\mathcal{F}}$ is regular, together with the fact that (2) in Proposition 5.3.1 already implies that for all $f \in \mathcal{F}$, the X_f are subcentered. (3) ⇒ (4) in Proposition 5.3.1 is directly implied by 3. below and again the fact that (3) in Proposition 5.3.1 already implies that for all $f \in \mathcal{F}$, the X_f are subcentered, so that $X_f \leq X - f - \mathbf{E}[X_f]$ for all $f \in \mathcal{F}$. (4) ⇒ (1) in Proposition 5.3.1 is directly implied by 4. below. (3) ⇒ (5) is implied by 5. below and (5) ⇒ (3) is implied by 6. below.

Proposition D.2.1. *1. Let $\{X_f\}_{f\in\mathcal{F}}$ be a family of random variables such that $\inf_{f\in\mathcal{F}} \mathbf{E}[X_f] > -\infty$. Suppose there is an ESI function u such that for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$. Then there are constants $C^* > 0$ and $\eta^* > 0$ such that uniformly for all $f \in \mathcal{F}$, $X_f \leq X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C^*$ (in particular, for all $f \in \mathcal{F}$, X_f is subcentered, i.e. $\mathbf{E}[X_f] \leq 0$).*

2. Suppose $\sup_{f\in\mathcal{F}} \mathrm{Var}(X_f) < \infty$. Suppose there is a constant $C^* > 0$ and a constant $\eta^* > 0$ such that uniformly for all $f \in \mathcal{F}$, $X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C^*$. Then there exists $c, v > 0$ such that for all $f \in \mathcal{F}$, the X_f are (c, v)-subgamma on the right, i.e. they satisfy (5.24).

3. Suppose there exists $c, v > 0$ such that for all $f \in \mathcal{F}$, the X_f are (c, v)-subgamma on the right. Then there is an ESI function h such that for all $f \in \mathcal{F}$, we have $X_f - \mathbf{E}[X_f] \trianglelefteq_h 0$ where h is of the form $h(\epsilon) = C\epsilon \wedge \eta^*$ for some constants $C > 0, \eta^* > 0$.

4. Suppose that for all $f \in \mathcal{F}$, we have $X_f \leq X_f - \mathbf{E}[X_f] \trianglelefteq_h 0$ where $h(\epsilon) = C\epsilon \wedge \eta^*$ for some $C, \eta^* > 0$. Then there is an ESI function u such that for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$.

5. Suppose there exists $c, v > 0$ such that for all $f \in \mathcal{F}$, the X_f are (c, v)-subgamma on the right. Then for all $0 < \delta \leq 1$, all $f \in \mathcal{F}$, (5.25) holds with probability at least $1 - \delta$.

6. *Suppose there exists $c, v > 0$ such that for all $f \in \mathcal{F}$, the X_f are subcentered and, for each $f \in \mathcal{F}$, for each $0 < \delta \leq 1$, with probability at least $1 - \delta$, (5.25) holds. Then:*
there exists $a > 0$ and a differentiable function $h : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ with $h(\epsilon) > 0$ and $h'(\epsilon) \geq 0$ for $\epsilon > 0$, such that for all $f \in \mathcal{F}$, the X_f are subcentered and $\mathbf{P}\{X \geq \epsilon\} \leq a \exp(-h(\epsilon))$.

7. *Suppose the condition above holds. Then there is $C^* > 0, \eta^* > 0$ such that uniformly for all $f \in \mathcal{F}$, $X_f \leq X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C^*$.*

Proof. Part 1. Let $C' := -\inf_{f \in \mathcal{F}} \mathbf{E}[X_f] < \infty$. Take $C'' > 0$ such that $\eta^* := u(C'') > 0$. Then $X_f \trianglelefteq_{\eta^*} C''$ and $X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C'' + C' := C^*$. Also $\mathbf{E}[X_f] \leq \epsilon$ for all $\epsilon > 0$ and hence $\mathbf{E}[X_f] \leq 0$, which implies subcenteredness.

Part 2. see Theorem 5.3.2 and its proof below.

Part 3. Let $U_f = X_f - \mathbf{E}[X_f]$. The assumption of right subgammaness implies that for all $0 < \eta \leq \eta^*$ with $\eta^* = 1/(2c)$ and $C' = 2v$, we have

$$\mathbf{E}[e^{\eta U_f}] \leq \exp\left(\eta^2 \cdot \frac{v}{1 - c\eta}\right) \leq \exp\left(\eta^2 \cdot \frac{v}{1 - (1/2)}\right) = \exp\left(\eta^2 \cdot C'\right).$$

Now take $h(\epsilon) = \epsilon/C'$ if $\epsilon \leq C'/2c$ and $h(\epsilon) = 1/2c$ for $\epsilon > C'/2c$. Then the above display implies that $\mathbf{E}[U_f] \trianglelefteq_h 0$, and h is seen to be equal to the h in the proposition statement.

Part 4. The premise implies that $\mathbf{E}[X_f] \leq 0$ for all $f \in \mathcal{F}$ and $X_f \trianglelefteq_h 0$, so we can (trivially) take $u = h$.

Part 5. (5.25) be rewritten as: for every $t > 0$,

$$\mathbf{P}\{X_f > \sqrt{2vt} + ct\} \leq \exp(-t), \tag{D.4}$$

which is shown to be implied by (c, v)-subgammaness on the right in [Boucheron et al., 2013, Section 2.4].

Part 6. [Boucheron et al., 2013, Section 2.4]. shows that (D.4) is equivalent to $\mathbf{P}\{X_f > t\} \leq \exp(-h(t))$ where $h(t) = (v/c^2)h_1(ct/v))$, with $h_1(u) = 1 + u - \sqrt{1 + 2u}$; this is of the required form.

Part 7. If $h(0) > 0$, then we can simply apply Proposition 5.2.5 to get the desired result; if $h(0) = 0$, apply the proposition with a set in the proposition to $2a$. □

of Theorem 5.3.2. Suppose that $U - \mathbf{E}[U] \trianglelefteq_{\eta^*} C$ holds and suppose without loss of generality that U is centered. Let $M = e^{\eta^* \mathbf{A}^{\eta^*}[U]} = e^{\eta^* C}$, which is finite by assumption.

We bound the moments of the right part of U.

$$\begin{aligned}\mathbf{E}[(U_f)_+^n] &= \mathbf{E}[\int_0^{(U_f)_+} nu^{n-1}\mathrm{d}u] \\ &= \mathbf{E}[\int_0^{\infty} \mathbf{P}\{U_f \geq u\} nu^{n-1}\mathrm{d}u] \\ &\leq \mathrm{e}^{\eta^* \mathbf{A}^{\eta^*}[U_f]} n \int_0^{\infty} \mathrm{e}^{-\eta^* u} u^{n-1}\mathrm{d}u \\ &= n! \frac{M}{\eta^{*n}}\end{aligned}$$

Now let $v = \mathrm{Var}(U) + \frac{2M}{\eta^{*2}}$ and $c = \frac{1}{\eta^*}$. This means that $\mathbf{E}[(U)_+^n] \leq n! \frac{v}{2} c^{n-2}$ for $n \geq 3$ and $\mathbf{E}[U^2] \leq v$ uniformly over $\mathcal{F}$. The rest of the proof follows that of Boucheron et al. [2013, Theorem 2.10]: note that $\mathrm{e}^x \leq 1 + x + \frac{1}{2}x^2$ for $x \leq 0$ so that

$$\mathrm{e}^x \leq 1 + x + \frac{1}{2}x^2 + \sum_{n\geq 3} \frac{x_+^n}{n!}$$

for all x. With this in mind, we obtain a bound on the moment generating function of U:

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{\eta U}] &\leq 1 + \mathbf{E}[U] + \frac{1}{2}\eta^2 \mathrm{Var}(U) + \sum_{n\geq 3} \frac{\eta^n \mathbf{E}[(U)_+^n]}{n!} \\ &\leq 1 + \frac{1}{2} v\eta^2 \sum_{n\geq 2} (c\eta)^{n-2} \\ &= 1 + \frac{1}{2} \frac{v\eta^2}{1 - c\eta}\end{aligned}$$

for $0 < c\eta < 1$. Taking logarithms on both sides, using that $\ln(1 + x) \leq x$ for $x \geq 0$ and rewriting leads to

$$\mathbf{A}^{\eta}[U] \leq \frac{1}{2} \frac{v\eta}{1 - c\eta},$$

which is exactly what we were after. □

Proof of the Claim in Example 5.3.3 Let $\eta < 1$. Using $\mathbf{P}\{-1/\eta \leq U < -1\} = \int_{-1/\eta}^{-1} p(u)du = (1 - \eta^{v-1})/(v - 1)$ and $\mathbf{E}[U^2 \cdot \mathbf{1}\{-1/\eta < U < 0\}] = \int_{-1/\eta}^{-1} u^2 p(u)du =$

$(1-\eta^{\nu-3})/(\nu-3)$ we find:

$$\begin{aligned}
\mathbf{E}[\exp(\eta U)] &\geq \mathbf{E}[\exp(\eta U)\cdot \mathbf{1}\{-1/\eta < U < 0\}] + \mathbf{E}[\exp(\eta U)\cdot \mathbf{1}\{U \geq 0\}] \\
&\geq \mathbf{E}[(1+\eta U + U^2\eta^2/4)\cdot \mathbf{1}\{-1/\eta < U < 0\}] + \mathbf{E}[(1+\eta U)\cdot \mathbf{1}\{U \geq 0\}] \\
&\geq \mathbf{P}\{U > -1/\eta\} + \eta\mathbf{E}[U] + (\eta^2\cdot\exp(-1)\cdot \mathbf{E}[U^2\cdot \mathbf{1}\{-1/\eta < U < 0\}] \\
&= \mathbf{P}\{-1/\eta \leq U < -1\} + (1-\mathbf{P}\{U \leq -1\}) + \\
&\quad \eta\mathbf{E}[U] + \eta^2\cdot\exp(-1)\cdot\mathbf{E}[U^2\cdot\mathbf{1}\{-1/\eta < U < 0\}] \\
&= \frac{1-\eta^{\nu-1}}{\nu-1} + (1-\frac{1}{\nu-1}) + \eta\mathbf{E}[U] + \eta^2\cdot\exp(-1)\cdot\frac{\eta^{\nu-3}-1}{3-\nu} \\
&= 1 - \frac{1}{\nu-1}\eta^{\nu-1} + \frac{\exp(-1)}{(3-\nu)}\cdot(\eta^{\nu-1}-\eta^2).
\end{aligned}$$

Since for $5/2 < \nu < 3$, we have $\exp(-1)/(3-\nu) > 1/(\nu-1)$, we find that, as $\eta \downarrow 0$, $(\mathbf{E}[\exp(\eta U)]-1)/\eta^2 \to \infty$, showing that right subgamma-ness is violated.

D.3. Proofs for Section 5.3.2

Below we first state Theorem D.3.1 and then Lemma D.3.2. We then show how, taken together, these two results imply the result in the main text, Theorem 5.3.11, as an almost direct corollary. After that, we provide first the proof of Lemma D.3.2 and then the proof of Theorem D.3.1 itself, followed by the statement and proof of Lemma D.3.3, a slight extension of the standard second-order Taylor approximation of the moment generating function that is crucial for proving Lemma D.3.2.

Theorem D.3.1. *1. Suppose $\{-X_f : f \in \mathcal{F}\}$ satisfies the $[0,b)$-Bernstein condition for some $0 < b \leq 1$ and the conclusion **C1** of Lemma D.3.2, Part 1 holds. Then for all $\beta \in [0,b)$, all $c \geq 0$, all $0 < c^* < 1$, there exists $\eta^\circ > 0$ and $C^\circ > 0$ such that for all $f \in \mathcal{F}$, all $0 < \eta \leq \eta^*$,*

$$X_f + c\eta X_f^2 - c^*\mathbf{E}[X_f] \trianglelefteq_\eta C^\circ\eta^{1/(1-\beta)}.$$

*2. Suppose there exists $\eta^\circ > 0, C^\circ > 0$ such that, for some $0 < \beta \leq 1$, for all $f \in \mathcal{F}$, $X_f \trianglelefteq_{\eta^\circ} C^\circ\eta^{1/(1-\beta)}$ and the conclusion **C2** of Lemma D.3.2, Part 2 holds. Then, (a) $\{-X_f : f \in \mathcal{F}\}$ satisfies the β-Bernstein condition. If furthermore $\sup_{f\in\mathcal{F}}\mathbf{E}[-X_f] < \infty$ then (b), $\{-X_f : f \in \mathcal{F}\}$ also satisfies the $[0,\beta]$-Bernstein condition and also $\sup_{f\in\mathcal{F}}\mathbf{E}[X_f^2]$ is bounded, so that the family is regular.*

Lemma D.3.2. Let $\{X_f : f \in \mathcal{F}\}$ be an ESI family.

1. Suppose that the family is regular. Then **C1** holds, with:
   C1: for each $k \geq 0$, for all $0 < \delta \leq 1$, there exist $C^*, C^\circ > 0$ and $\eta^\circ > 0$ (that may depend on δ) such that for all $0 < \eta \leq \eta^\circ$, all $f \in \mathcal{F}$, $X_f \trianglelefteq_{\eta^\circ} C^*$ and

$$\mathbf{E}[|X_f|^{2+k}\exp(\eta X_f)] \leq C^\circ(\mathbf{E}[X_f^2])^{1-\delta} \tag{D.5}$$

2. Suppose that the witness-type condition (5.30) holds for the family $\{X_f : f \in \mathcal{F}\}$ or for the family $\{(X_f)_- : f \in \mathcal{F}\}$. Then **C2** holds, with:
   C2: there exist $C > 0$ and an $\eta^\circ > 0$ such that for all $0 < \eta \le \eta^\circ$, all $f \in \mathcal{F}$,

$$\mathbf{E}[X_f^2] \le C \cdot \mathbf{E}[X_f^2 \exp(\eta X_f)]. \tag{D.6}$$

To see how the above two results together imply Theorem 5.3.11 in the main text, note that the implication (1) ⇒ (2) in that theorem is a direct consequence of the fact that **C1** in Lemma D.3.2 holds for regular ESI families for $\delta = 1 - \beta$, any $0 \le \beta < 1$, as expressed by Part 1 of that lemma, combined with Theorem D.3.1, Part 1.

Implication (2) ⇒ (3) of Theorem 5.3.11 is trivial. Implication (3) ⇒ (1) follows by the fact that **C2** in Lemma D.3.2 holds for families satisfying the witness-type condition, as expressed by Part 2 of that lemma, combined with Theorem D.3.1, Part 2.

of Lemma D.3.2. Part 1. Let $s = \sup_{f\in\mathcal{F}} \mathbf{E}[X_f^2]$ and set $X = X_f$ for arbitrary $f \in \mathcal{F}$. Since we assume the family is regular, we can use Proposition 5.3.1 to infer that there exists $\eta^* > 0$, $C^* > 0$ such that $X \trianglelefteq_\eta C^*$ for all $0 < \eta \le \eta^*$.

For all $\eta' > 0, \eta > 0$, all $0 < \gamma < 2$ there must be constants $C, C' > 0$, such that for all $p > 0, q > 0$ with $1/p + 1/q = 1$, it holds that :

$$\begin{aligned}
&\mathbf{E}[|X|^{2+k} \exp(\eta X)] \\
&= \mathbf{E}[\mathbf{1}\{X \le -1\} |X|^2 \cdot \left(|X|^k \exp(\eta X)\right)] + \\
&\qquad \mathbf{E}[\mathbf{1}\{-1 < X < 1\} |X|^{2+k} \exp(\eta X)] + \mathbf{E}[\mathbf{1}\{X \ge 1\} |X|^{2+k} \exp(\eta X)] \\
&\le C' \mathbf{E}[\mathbf{1}\{X \le -1\} X^2] + \exp(\eta) \mathbf{E}[\mathbf{1}\{-1 < X < 1\} X^2] + \\
&\qquad\qquad \mathbf{E}[\mathbf{1}\{X \ge 1\} |X|^{2-\gamma} \cdot \left(|X|^{k+\gamma} \exp(\eta X)\right)] \\
&\le (C' + \exp(\eta)) s (\mathbf{E}[X^2]/s)^{1-\delta} + C \mathbf{E}[\mathbf{1}\{X \ge 1\} X^{2-\gamma} \exp((\eta' + \eta) X)] \\
&\le (C' + \exp(\eta) s^\delta \cdot (\mathbf{E}[X^2])^{1-\delta} + \\
&\qquad C \left(\mathbf{E}[(\mathbf{1}\{X \ge 1\} X^{2-\gamma})^p]\right)^{1/p} \left(\mathbf{E}[\exp(q(\eta' + \eta) X)]\right)^{1/q},
\end{aligned}$$

where we in the first inequality we used that $|X|^k \exp(\eta X)$ is bounded on $X \le -1$ and in the second we used that $|X|^{k+\gamma} \exp(\eta X)$ is bounded by a constant times $\exp((\eta' + \eta) X)$ on $X \ge 1$. Then we used Hölder's inequality and the fact that $\mathbf{E}[X^2]/s \le 1$. We now take $0 < \delta \le 1$ as in the theorem statement and bound the second term further setting $1/p = 1 - \delta$, $1/q = \delta$ and $\gamma = 2\delta$ (so that $1/p + 1/q = 1$ as required and $(2 - \gamma)p = 2$):

$$\begin{aligned}
&\left(\mathbf{E}[(\mathbf{1}\{X \ge 1\} X^{2-\gamma})^p]\right)^{1/p} \left(\mathbf{E}[\exp(q(\eta' + \eta) X)]\right)^{1/q} \\
&= \left(\mathbf{E}[(\mathbf{1}\{X \ge 1\} X^{2-\gamma})^p]\right)^{1-\delta} \left(\mathbf{E}[\exp(\delta^{-1}(\eta' + \eta) X)]\right)^{\delta} \\
&\le \left(\mathbf{E}[\mathbf{1}\{X \ge 1\} X^2]\right)^{1-\delta} \left(\mathbf{E}[\exp(\delta^{-1}(\eta' + \eta) X)]\right)^{\delta} \le \left(\mathbf{E}[X^2]\right)^{1-\delta} C^*
\end{aligned}$$

where the final equation follows for the specific choice $\eta' = \eta^*/(2\delta)$ and any η with $0 < \eta \le \eta^\circ := \eta^*/(2\delta)$. Combining the two equations we find that for any such η, using

that the constants C and C^* do not depend on the f with $X = X_f$, the result (D.5) follows.

Part 2. Let $X = X_f$ for arbitrary $f \in \mathcal{F}$. First assume (5.30) for the family $\{X_f : f \in \mathcal{F}\}$. We have:

$$\mathbf{E}[X^2] = \mathbf{E}[X^2\mathbf{1}\{X \geq 0\}] + \mathbf{E}[X^2\mathbf{1}\{X < 0; X^2 \leq C\}] + \mathbf{E}[X^2\mathbf{1}\{X < 0; X^2 > C\}] \tag{D.7}$$

$$\leq \mathbf{E}[X^2\exp(\eta X)\mathbf{1}\{X \geq 0\}] + \mathbf{E}[X^2\mathbf{1}\{X < 0; X^2 \leq C\}] + \mathbf{E}[X^2\mathbf{1}\{X^2 > C\}] \tag{D.8}$$

$$\leq \mathbf{E}[X^2\exp(\eta X)\mathbf{1}\{X \geq 0\}] + \exp(\eta^\circ\sqrt{C})\mathbf{E}[X^2\mathbf{1}\{X < 0; X^2 \leq C\}e^{\eta X}] + c\mathbf{E}[X^2]$$

for some $0 < c < 1$. Moving the rightmost term to the left-hand side and dividing both sides by $1 - c$, we find that

$$\mathbf{E}[X^2] \leq \frac{1}{1-c}\left(\exp(\eta^\circ\sqrt{C})\mathbf{E}[X^2\exp(\eta X)]\right),$$

which is what we had to prove. In case that (5.30) holds the family $\{(X_f)_- : f \in \mathcal{F}\}$, the result follows by a minor variation of the above argument: the rightmost term in (D.7) can then be bounded by $c\mathbf{E}[X_-^2]$, so the rightmost term in (D.8) can be replaced by $c\mathbf{E}[X_-^2]$, and then the final inequality still holds. □

of Theorem D.3.1. Part 1. A short calculation using Jensen's inequality shows that for all $c, c^*, \eta > 0$, we have

$$(X_f - c^* \cdot \mathbf{E}[X_f] + \eta c X_f^2)^2 \leq 3X_f^2 + 3c^{*2}(\mathbf{E}[X_f])^2 + 3\eta^2 c^2 X_f^4. \tag{D.9}$$

Take $0 < c^* < 1$ and $\beta \in [0, b)$ as in the theorem statement. Set $\delta := 1 - \beta$ and choose η° for this δ as in Lemma D.3.2, Part 1 (which we can use because $0 < \delta \leq 1$). Without loss of generality let $\eta^\circ \leq 1$. We find for all $0 < \eta < \eta^\circ$, that for some η' with $0 < \eta' < \eta^\circ$, and for some constants $C_0, C_1, C_2, C_3, C_4, C^\circ > 0$, for $\delta' = 2\delta - \delta^2$ that

$$\begin{aligned}
&\mathbf{E}[\exp(\eta(X_f - c^* \cdot \mathbf{E}[X_f] + \eta c X_f^2)] \\
&\leq 1 + \eta\mathbf{E}[X_f - c^* \cdot \mathbf{E}[X_f] + \eta c X_f^2] + \frac{1}{2}\eta^2\mathbf{E}[(X_f - c^* \cdot \mathbf{E}[X_f] + \eta c X_f^2)^2 e^{\eta' X_f}] \\
&\leq 1 + \eta(1 - c^*) \cdot \mathbf{E}[X_f] + \eta^2 c\mathbf{E}[X_f^2] + \\
&\quad \frac{3}{2}\eta^2\left(c^* 2\mathbf{E}[X_f^2]\mathbf{E}[e^{\eta' X_f}] + \mathbf{E}[X_f^2 e^{\eta' X_f}] + 3c^2\mathbf{E}[X_f^4 e^{\eta' X_f}]\right) \\
&\leq 1 + \eta(1 - c^*) \cdot \mathbf{E}[X_f] + \eta^2 c\mathbf{E}[X_f^2] + \frac{3}{2}c^{*2}\eta^2 C^*\mathbf{E}[X_f^2] + \frac{3}{2}(1 + c^2)\eta^2 C^\circ(\mathbf{E}[X_f^2])^{1-\delta} \\
&\leq 1 + \eta(1 - c^*) \cdot \mathbf{E}[X_f] + \eta^2 C_1 B\mathbf{E}[-X_f]^{1-\delta} + \eta^2 C_2 B^{1-\delta}(\mathbf{E}[-X_f])^{(1-\delta)^2} \\
&\leq 1 + \eta(1 - c^*) \cdot \mathbf{E}[X_f] + \eta^2 C_3\mathbf{E}[-X_f]^{1-\delta'} + \eta^2 C_4(\mathbf{E}[-X_f])^{(1-\delta')} \\
&\leq 1 + \eta\left((1 - c^*) \cdot \mathbf{E}[X_f] + \eta\frac{C_3 + C_4}{(1 - c^*)^{1-\delta'}}\left((1 - c^*)\mathbf{E}[-X_f]\right)^{1-\delta'}\right) \\
&\leq 1 + \eta\left((1 - c^*) \cdot \mathbf{E}[X_f] + C^\circ\eta^{1/\delta} + (1 - c^*)\mathbf{E}[-X_f]\right) \leq 1 + \eta \cdot C^\circ\eta^{1/\delta} \leq e^{\eta C^\circ \eta^{1/\delta}}.
\end{aligned}$$

Here the first inequality is our extended Taylor approximation from Lemma D.3.3, stated and proved further below. For the second we used (D.9). The third follows by Lemma D.3.2, Part 1, applied with $k = 0$ (for the first and second term within the rightmost brackets, and for determining C^*) and $k = 2$, for the third term within those brackets, and with constant C_0 taken to be the maximum of the two corresponding constants C° in that lemma obtained with $k = 0$ and $k = 2$. The fourth follows from the β-Bernstein condition (applied to the two final terms) for $\beta = 1 - \delta$, which holds by assumption. The fifth follows because by Cauchy-Schwarz,

$$\mathbf{E}[-X_f]^{1-\delta} = \mathbf{E}[-X_f]^{1-\delta'} \cdot \mathbf{E}[X_f]^{\delta'-\delta} \leq \mathbf{E}[-X_f]^{1-\delta'} \cdot \mathbf{E}[X_f^2]^{(\delta'-\delta)/2}$$

and the latter factor is bounded since we assume the $\{X_f\}$ to be regular. The sixth inequality above is just rearranging, and the seventh inequality is a 'linearization' step. To see how it follows, note first that, for $p, q > 0$ with $1/p + 1/q = 1$, Young's inequality, $xy \leq |x|^p/p + |y|^q/q$, implies that for $0 < \beta < 1$,

$$ab^\beta \leq \frac{1-\beta}{\beta}(\beta a)^{\frac{1}{1-\beta}} + b \tag{D.10}$$

which follows for $a, b > 0$ by taking $\beta = 1/p$, $a = x^p$, and $b = y$. We apply this with $\beta = 1 - \delta'$, $b = (1 - c^*)\mathbf{E}[-X_f]$, $a = \eta(C_3 + C_4)/(1 - c^*)^{1-\delta'}$. The final inequality then gives the desired result.

Part 2 is similar to 1 but much easier; we omit the details. □

We end the section with the statement and proof of the validity of the second order Taylor approximation of the moment generating function, as used in the proof of Lemma D.3.2.

Lemma D.3.3 ("Extended Taylor"). Suppose that $\mathbf{E}[X^2] < \infty$ and also $\mathbf{E}[e^{\eta X}] < \infty$ for all $\eta \in [0, \eta_{\max}]$ and let η^* be any number with $0 < \eta^* < \eta_{\max}$. Then for all $0 < \eta < \eta^*$ we have:

$$\mathbf{E}[e^{\eta X}] = 1 + \eta\mathbf{E}[X] + \frac{1}{2}\eta^2\mathbf{E}[X^2 e^{\eta' X}] \tag{D.11}$$

for some η' with $\eta \leq \eta' < \eta^*$.

This is just the standard Taylor approximation of the moment generating function for random variable X at $\eta = 0$. However, in this chapter we need this approximation also for the case that $\mathbf{E}[e^{\eta X}] = \infty$ for all $\eta < 0$ (e.g. if X has polynomial left tail), in which the standard Taylor's theorem does not apply any more, since the standard (two-sided) derivative at $\eta = 0$ is undefined. The lemma shows that nevertheless, everything still works as one would expect.

Proof. Fix $\eta_0 \in (0, \eta^*)$. Then all derivatives of $\mathbf{E}[\exp(\eta X)]$ exist at η with $\eta_0 \leq \eta \leq \eta^*$, so that:

$$\mathbf{E}[e^{\eta X}] = \mathbf{E}[e^{\eta_0 X}] + (\eta - \eta_0)\mathbf{E}[X e^{\eta_0 X}] + \frac{1}{2}(\eta - \eta_0)^2\mathbf{E}[X^2 e^{s(\eta_0,\eta)X}] \tag{D.12}$$

with s some function $s : [0, \eta^*]^2 \to [\eta_0, \eta]$. Since

$$|\mathrm{e}^{\eta X} X| \le |\mathrm{e}^{\eta X_+} X| \le |\mathrm{e}^{\eta^* X_+} X| \le |X_-| + |\mathbf{1}\{X \ge 0\}\, \mathrm{e}^{\eta^* X_+} X| \le |X_-| + |\mathrm{e}^{\eta^* X} X|$$

and we know $\mathbf{E}[|\mathrm{e}^{\eta^* X} X|] < \infty$, we can use the dominated convergence theorem to conclude that $\lim_{\eta_0 \downarrow 0} \mathbf{E}[X \mathrm{e}^{\eta_0 X}] = \mathbf{E}[X]$. Analogously one shows that $\lim_{\eta_0 \downarrow 0} \mathbf{E}[X^2 \mathrm{e}^{s(\eta_0, \eta) X}] = \mathbf{E}[X^2 \mathrm{e}^{\eta' X}]$ for an $\eta' \in [\eta_0, \eta]$. The result now follows by using these two limiting results in taking the limit for $\eta_0 \downarrow 0$ in (D.12). □

D.4. Proofs for Section 5.5

of Theorem 5.5.2. Inequality (5.37) follows from Proposition 5.2.3, applied with $X = \hat{\eta} X_{\hat{\eta}}, Y = \hat{\eta} Y_{\hat{\eta}}, \eta = 1$. (5.39) follows from Proposition 5.2.5, with X, Y and η set in the same way (see the remark at the end of the proposition statement).

Now we prove (5.38). Define the random variable $W_{\hat{\eta}} \coloneqq X_{\hat{\eta}} - Z_{\hat{\eta}}$, and for $(a, k) \in (0, \infty) \times \mathbb{N}$ define the event $\mathcal{E}_{a,k} \coloneqq \{a \cdot (k-1) \le \hat{\eta} W_{\hat{\eta}} \le ak\}$. With $Z = \hat{\eta} W_{\hat{\eta}}$, we will first show that

$$\limsup_{a \downarrow 0} \sum_{k=1}^{\infty} ak \cdot \mathbf{P}\{\mathcal{E}_{a,k}\} \le \int_0^{\infty} \mathbf{P}\{Z \ge z\} \mathrm{d}z. \tag{D.13}$$

For $a, b > 0$ and $k_{a,b} \coloneqq \lfloor b/a \rfloor$, we have

$$\begin{aligned}
\sum_{k=1}^{k_{a,b}} ak \cdot \mathbf{P}\{\mathcal{E}_{a,k}\} &= \sum_{k=1}^{k_{a,b}} ak \cdot \left(\mathbf{P}\{Z > a \cdot (k-1)\} - \mathbf{P}\{Z \ge ak\}\right), \\
&\le \sum_{k=1}^{k_{a,b}} ak \cdot \left(\mathbf{P}\{Z \ge a \cdot (k-1)\} - \mathbf{P}\{Z \ge ak\}\right), \\
&\le \sum_{k=0}^{k_{a,b}-1} a \cdot \mathbf{P}\{Z \ge ak\}, \\
&\le a\mathbf{P}\{Z \ge 0\} + \int_0^{k_{a,b}} a\mathbf{P}\{Z \ge at\} \mathrm{d}t, \quad (t \mapsto a \cdot \mathbf{P}\{Z \ge at\} \text{ nonincr.}) \\
&= a\mathbf{P}\{Z \ge 0\} + \int_0^{ak_{a,b}} \mathbf{P}\{Z \ge z\} \mathrm{d}z, \quad (\text{change of variable } y = at) \\
&\le a + \int_0^{b} \mathbf{P}\{Z \ge z\} \mathrm{d}z.
\end{aligned} \tag{D.14}$$

Since (D.14) holds for all $a, b > 0$, we have

$$\limsup_{a \downarrow 0} \sum_{k=1}^{\infty} ak \cdot \mathbf{P}\{\mathcal{E}_{a,k}\} = \limsup_{a \downarrow 0} \left\{ \sup_{b > 0} \sum_{k=1}^{k_{a,b}} ak \cdot \mathbf{P}\{\mathcal{E}_{a,k}\} \right\} \overset{\text{(D.14)}}{\le} \int_0^{\infty} \mathbf{P}\{Z \ge z\} \mathrm{d}z, \tag{D.15}$$

and thus, the desired inequality (D.13) follows. We now have, for all $a > 0$,

$$\begin{aligned}\mathbf{E}[W_{\hat{\eta}}] &\leq \sum_{k=1}^{\infty} \mathbf{P}\{\mathcal{E}_{a,k}\} \cdot \mathbf{E}[W_{\hat{\eta}} | \mathcal{E}_{a,k}], \\ &\leq \sum_{k=1}^{\infty} \mathbf{P}\{\mathcal{E}_{a,k}\} \cdot \mathbf{E}\left[\frac{ak}{\hat{\eta}}\right] \quad \text{(by the definition of } \mathcal{E}_{a,k}) \\ &= \left(\sum_{k=1}^{\infty} \mathbf{P}\{\mathcal{E}_{a,k}\} \cdot ak\right) \cdot \mathbf{E}\left[\frac{1}{\hat{\eta}}\right].\end{aligned}$$

Since this inequality holds for all $a > 0$, using (D.13) implies that

$$\begin{aligned}\mathbf{E}[W_{\hat{\eta}}] &\leq \mathbf{E}\left[\frac{1}{\hat{\eta}}\right] \cdot \int_0^{\infty} \mathbf{P}\{\hat{\eta} W_{\hat{\eta}} \geq t\} \mathrm{d}t \\ &\leq \mathbf{E}\left[\frac{1}{\hat{\eta}}\right] \cdot \mathbf{E}\left[\mathrm{e}^{\hat{\eta} W_{\hat{\eta}}}\right] \int_0^{\infty} \mathrm{e}^{-t} \mathrm{d}t, \text{(Markov's Inequality)} \\ &\leq \mathbf{E}\left[\frac{1}{\hat{\eta}}\right], \end{aligned} \tag{D.16}$$

where the last step follows from the fact that $W_{\hat{\eta}} \trianglelefteq_{\hat{\eta}} 0$. □

of Proposition 5.5.4. Let's denote $X_{i,\eta}(Z; z^{n \setminus i}) \coloneqq X_{i,\eta}(z_1, \ldots, z_{i-1}, Z, z_{i+1}, \ldots, z_n)$, for all $\eta \in \mathcal{G}$, $i \in [n]$, $z^{n \setminus i} \in \mathcal{Z}^{n-1}$, and $Z \in \mathcal{Z}$. In this way, (5.40) can be written as

$$X_{i,\eta}(Z_i; z^{n \setminus i}) \trianglelefteq_{\eta} 0, \quad \text{for all } i \in [n] \text{ and } z^{n \setminus i} \in \mathcal{Z}^{n-1}. \tag{D.17}$$

In particular, since this holds for all $z^{n \setminus i} \in \mathcal{Z}^{n-1}$, we can also write

$$X_{i,\eta}(Z_i; Z^{n \setminus i}) \trianglelefteq_{\eta} 0. \tag{D.18}$$

Now let $Z_1^n, \ldots, Z_n^n \in \mathcal{Z}^n$ be n i.i.d copies of Z^n. From (D.18), we have, for each $i \in [n]$,

$$Y_{i,\eta} \coloneqq X_{i,\eta}(Z_{i,i}; Z_i^{n \setminus i}) \trianglelefteq_{\eta} 0. \tag{D.19}$$

Since $(Y_{i,\eta})_{i \in [n]}$ are i.i.d, we can chain (D.19), for $i = 1, \ldots, n$, using Proposition 5.2.6 to get

$$\sum_{i=1}^{n} Y_{i,\eta} \trianglelefteq_{\eta} 0. \tag{D.20}$$

By applying Proposition 5.5.5, we have, for any random $\hat{\eta}$ in $\mathcal{G}$,

$$\begin{aligned}\mathbf{E}\left[\frac{\ln|\mathcal{G}| + 1}{\hat{\eta}}\right] &\geq \mathbf{E}\left[\sum_{i=1}^{n} Y_{i,\hat{\eta}}\right], \\ &= \mathbf{E}\left[\sum_{i=1}^{n} X_{i,\hat{\eta}}(Z_{i,i}, Z_i^{n \setminus i})\right], \\ &= \mathbf{E}\left[\sum_{i=1}^{n} X_{i,\hat{\eta}}(Z^n)\right], \end{aligned} \tag{D.21}$$

the last equality follows from the fact that $(Z_i^n)_{i \in [n]}$ are i.i.d copies of Z^n. □