Efficient tracking of a growing number of experts

Jaouad Mourtada & Odalric-ambrym Maillard

CMAP, École Polytechnique & Sequel, INRIA Lille – Nord Europe

ALT 2017, Kyoto University

Prediction with expert advice

- Well studied, standard framework for online learning (see [Cesa-Bianchi and Lugosi, 2006])
- **Aim:** combine the forecasts of several **experts** $\Longrightarrow$ **predict almost as well as the best** of them
- **Adversarial/worst case** setting (no stochasticity assumption on the signal)

Formal setting

- $\mathscr{X}$ prediction space, $\mathscr{Y}$ signal space, $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}$ loss function
- Experts $i = 1, \ldots, M$

Prediction with expert advice

At each time step $t = 1, 2, \ldots$

1. Experts $i = 1, \ldots, M$ output predictions $x_{i,t} \in \mathscr{X}$
2. Forecaster predicts $x_t \in \mathscr{X}$
3. Environment chooses signal value $y_t \in \mathscr{Y}$
4. Experts $i = 1, \ldots, M$ incur loss $\ell_{i,t} := \ell(x_{i,t}, y_t)$, forecaster gets loss $\ell_t := \ell(x_t, y_t)$

Formal setting

Prediction with expert advice

At each time step $t = 1, 2, \ldots$

1. Experts $i = 1, \ldots, M$ output predictions $x_{i,t} \in \mathscr{X}$
2. Forecaster predicts $x_t \in \mathscr{X}$
3. Environment chooses signal value $y_t \in \mathscr{Y}$
4. Experts $i = 1, \ldots, M$ incur loss $\ell_{i,t} := \ell(x_{i,t}, y_t)$, forecaster gets loss $\ell_t := \ell(x_t, y_t)$

Goal: strategy for the Forecaster with controlled worst-case **regret**

$$R_{i,T} = L_T - L_{i,T} = \sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i,t}$$

Assumption on the loss function

Assumption (η-Exp-concavity)

Loss function ℓ is η-**exp-concave** for some $\eta > 0$, *i.e.* for every $y \in \mathscr{Y}$, the function $\exp(-\eta\,\ell(\cdot, y)) : \mathscr{X} \to \mathbf{R}_+$ is concave.

Important examples:

- **Logarithmic**, or **self-information** loss: $\mathscr{X} = \mathscr{P}(\mathscr{Y})$, $\ell(p, y) = -\log p(\{y\})$
- **Square loss** on a **bounded** domain: $\mathscr{X} = \mathscr{Y} = [a, b]$, $\ell(x, y) = (x - y)^2$, $\eta = \frac{1}{2(b-a)^2}$
- **NOT** the absolute loss $\ell(x, y) = |x - y|$ on $[0, 1]^2$

The exponential weights algorithm

x_{it} : prediction of expert i at time t

Exponential weights/Hedge algorithm

$$x_t = \sum_{i=1}^{M} v_{i,t}\, x_{i,t} \qquad v_{i,t} = \frac{\pi_i e^{-\eta\, L_{i,t-1}}}{\sum_{j=1}^{M} \pi_j e^{-\eta\, L_{j,t-1}}}$$

with $\boldsymbol{\pi} = (\pi_i)_{1 \leqslant i \leqslant M}$ a **prior** probability distribution on the experts

- Start with $\boldsymbol{v}_1 = \boldsymbol{\pi}$.
- At end of round $t \geqslant 1$, after predicting and seeing losses $\ell_{i,t}$, **update** $\boldsymbol{v}_{t+1}$ by setting it to the **posterior distribution** $\boldsymbol{v}_t^m$:

$$v_{i,t+1} = v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^{M} v_{j,t}\, e^{-\eta\, \ell_{j,t}}}$$

Regret of the Hedge algorithm

Proposition (Vovk, Littlestone & Warmuth)

If ℓ is η-exp-concave, the Exponential Weights algorithm with prior $\boldsymbol{\pi}$ achieves the regret bound:

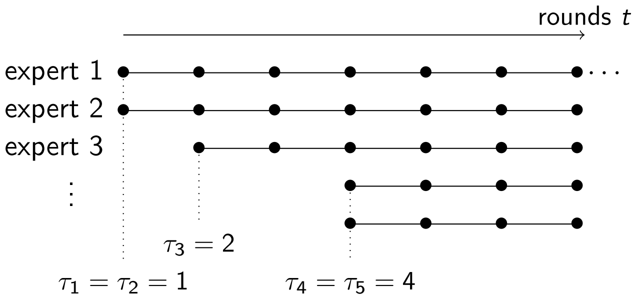
$$\forall i = 1, \ldots, M, \quad L_T - L_{i,T} \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i} \,. \tag{1}$$

In particular, if $\boldsymbol{\pi} = \frac{1}{M}\mathbf{1}$ is uniform,

$$L_T \leqslant \min_{1 \leqslant i \leqslant M} L_{i,T} + \frac{1}{\eta} \log M \,. \tag{2}$$

Sequentially incoming forecasters

- What if **new** experts (algorithms, methods, new data/variables...) become available over time ? How to **incorporate** them, with **formal regret guarantees** ?
- **Proposed setting: Growing set of experts.** M_t **increases** over time, and is **unknown in advance**; at time t, new experts $i = M_{t-1} + 1, \ldots, M_t$ start issuing predictions



Objective

Design **forecasting strategies** for the "Growing number of experts" setting, with emphasis on:

- **computationally inexpensive** strategies: ideal complexity $O(M_t)$ at step t
- **anytime** strategies: no fixed time horizon T
- no a priori knowledge of M_t
- no free parameters to tune
- **regret bounds** against **several classes of competitors**, that are **adaptive** to the parameters of the comparison class

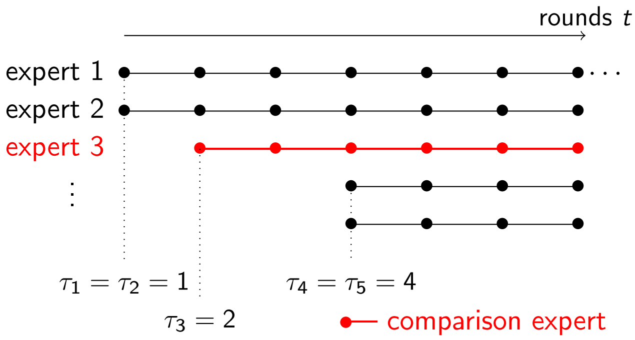
Growing experts

Recall the framework:

- At time t, experts $i = 1, \dots, M_t$ issue predictions; *i.e.* at time t, $m_t := M_t - M_{t-1}$ new experts $i = M_{t-1} + 1, \dots, M_t$ enter
- $\tau_i = \inf\{t \geqslant 1 \mid i \leqslant M_t\}$ **entry time** of expert i
- First notion of regret = **constant experts**: for each i,

$$R_{i,T} = \sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t})$$

$\longrightarrow$ "specialist trick"


rounds t
expert 1
expert 2
expert 3
$\tau_1 = \tau_2 = 1$
$\tau_3 = 2$
$\tau_4 = \tau_5 = 4$
comparison expert

The specialist setting

- Introduced by [Freund et al., 1997]
- **Specialists** $i = 1, \ldots, M$; at each time step t, only a subset $A_t \subset \{1, \ldots, M\}$ of **active** specialists output a prediction $x_{i,t}$
- **Goal**: minimize "regret" with respect to each specialist i

$$R_{i,T} = \sum_{t \leqslant T \,:\, i \in A_t} (\ell_t - \ell_{i,t})$$

The "specialist trick" [Chernov and Vovk, 2009]

- General method to turn an "expert" algorithm into a "specialist" algorithm
- **Idea**: "complete" specialists' predictions by making inactive specialists $i \notin A_t$ predict the same as the forecaster $x_{i,t} := x_t$

The "specialist trick" [Chernov and Vovk, 2009]

- General method to turn an "expert" algorithm into a "specialist" algorithm
- **Idea**: "complete" specialists' predictions by making inactive specialists $i \notin A_t$ predict the same as the forecaster $x_{i,t} := x_t$
- **Circular ?** $x_t = \sum_{i=1}^{M} v_{i,t} \, x_{i,t} \ldots$

The "specialist trick" [Chernov and Vovk, 2009]

- General method to turn an "expert" algorithm into a "specialist" algorithm
- **Idea**: "complete" specialists' predictions by making inactive specialists $i \notin A_t$ predict the same as the forecaster $x_{i,t} := x_t$
- **Circular ?** $x_t = \sum_{i=1}^{M} v_{i,t} \, x_{i,t} \dots$
- **(Unique) Solution:** For $i \notin A_t$, define

$$x_{i,t} := \frac{\sum_{i \in A_t} v_{i,t} \, x_{i,t}}{\sum_{i \in A_t} v_{i,t}}$$

$$\Longrightarrow x_t = \frac{\sum_{i \in A_t} v_{i,t} \, x_{i,t}}{\sum_{i \in A_t} v_{i,t}} = x_{i,t} \quad \text{for each } i \notin A_t$$

The “specialist trick” [Chernov and Vovk, 2009]

- General method to turn an “expert” algorithm into a “specialist” algorithm
- **Idea**: “complete” specialists’ predictions by making inactive specialists $i \notin A_t$ predict the same as the forecaster $x_{i,t} := x_t$
- **(Unique) Solution:** For $i \notin A_t$, define

$$x_{i,t} := \frac{\sum_{i \in A_t} v_{i,t}\, x_{i,t}}{\sum_{i \in A_t} v_{i,t}}$$

- By construction $\sum_{t=1}^{T} (\ell_t - \ell_{i,t}) = \sum_{t \leqslant T \,:\, i \in A_t} (\ell_t - \ell_{i,t})$ + Hedge regret bound $\Longrightarrow$ regret for SpecialistHedge with prior π

$$\forall i = 1, \dots, M, \qquad \sum_{t \leqslant T \,:\, i \in A_t} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i} \,.$$

Growing experts and specialists

- Specialists can abstain from predicting $\Longrightarrow$ can handle experts who have not entered yet
- **Growing experts can be viewed as specialists**: $A_t = \{1, \ldots, M_t\}$
- SpecialistHedge gives a regret bound for $R_{i,T}$
- Exactly **which total set of specialists** ?

Which total set of specialists ?

- **Naive choice :** Both T and M_T known in advance $\Longrightarrow$ set of specialists $\{1, \ldots, M_T\}$
  - Prior $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{M_T})$;
  - SpecialistHedge with prior $\boldsymbol{\pi}$ yields regret $R_{i,T} \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i}$ for $i = 1, \ldots, M_T$ (e.g. $\frac{1}{\eta} \log M_T$)
  - **Problem:** not anytime + requires knowledge of M_T

Which total set of specialists ?

- **Naive choice** : Both T and M_T known in advance $\Longrightarrow$ set of specialists $\{1, \dots, M_T\}$
- **Better choice** : set of specialists $\mathbf{N}^*$
  - Prior = **probability distribution** $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$ on $\mathbf{N}^*$
  - Keeping track of $e^{-\eta L_t}$, we only need to maintain the weights of **entered** experts
  - Yields anytime algorithm GrowingHedge with $O(M_t)$ per-round complexity + regret bound $R_{i,T} \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i}$ $\forall i, \forall T$

Which total set of specialists ?

- **Naive choice :** Both T and M_T known in advance $\Longrightarrow$ set of specialists $\{1, \dots, M_T\}$
- **Better choice :** set of specialists $\mathbf{N}^*$, GrowingHedge with normalized prior $\boldsymbol{\pi}$
- **Slightly better :** GrowingHedge with unnormalized prior $\boldsymbol{\pi}$
  - Observation : $\forall T \geqslant 1$, GrowingHedge coincides up to time T with SpecialistHedge on $\{1, \dots, M_T\}$ with prior $(\frac{\pi_1}{\Pi_{M_T}}, \dots, \frac{\pi_{M_T}}{\Pi_{M_T}})$, where $\Pi_{M_T} := \sum_{i=1}^{M_T} \pi_i$
  - Remains true even for arbitrary (non-summable) prior $\boldsymbol{\pi} \in (\mathbf{R}_+^*)^{\mathbf{N}^*}$: **more flexibility + simpler bounds**

Growing Hedge algorithm

Growing Hedge

- Set $w_{i,1} = \pi_i$ for $i = 1, \ldots, M_1$.
- For $t = 1, 2, \ldots$
  - Given predictions $x_{i,t}$ from experts $1 \leqslant i \leqslant M_t$, **predict**

$$x_t = \frac{\sum_{i=1}^{M_t} w_{i,t}\, x_{i,t}}{\sum_{i=1}^{M_t} w_{i,t}}$$

  - **Update** weights by $w_{i,t+1} = w_{i,t} e^{-\eta\, \ell_{i,t}}$ for $i = 1, \ldots, M_t$ and **introduce** $w_{i,t+1} = \pi_i e^{-\eta\, L_t}$ for $i = M_t + 1, \ldots, M_{t+1}$

Anytime, **efficient** algorithm, agnostic to M_t ; π_i **only used from time** τ_i

GrowingHedge: regret bound

Proposition

With arbitrary prior $\boldsymbol{\pi}$, GrowingHedge achieves regret bound

$$\forall T \geqslant 1,\ \forall i = 1, \dots, M_T, \quad \sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log \left(\frac{1}{\pi_i} \sum_{j=1}^{M_T} \pi_j \right)$$

- Prior $\pi_i = 1$ gives $R_{i,T} \leqslant \frac{1}{\eta} \log M_T$ (but now anytime).
- Prior $\pi_i = 1/(\tau_i m_{\tau_i})$: depends on entry time τ_i and number of new experts m_{τ_i}, both revealed at step $t = \tau_i$. Regret bound:

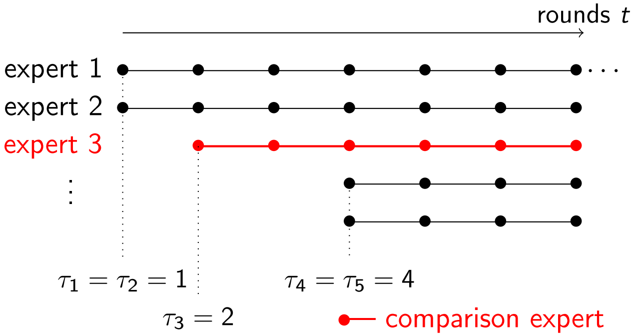
$$R_{i,T} \leqslant \frac{1}{\eta} \log m_{\tau_i} + \frac{1}{\eta} \log \tau_i + \frac{1}{\eta} \log(1 + \log T)\,.$$

Summary

- Regret against constant experts: naturally handled by the **specialist** setting
- Small subtlety in the choice of the set of specialists + extension to unnormalized prior (more general/flexible strategies with unified analysis)
- Using Hedge as base algorithm $\Longrightarrow$ simple and efficient: **only maintain weights for entered experts**
- But somewhat limited: does not work as seamlessly for more complex base algorithms/comparison classes

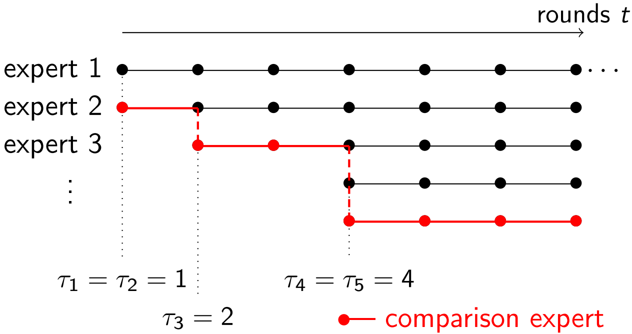
Different perspective on growing experts

- "Specialist" or "abstention trick" $\Longrightarrow$ GrowingHedge
- Controls regret w.r.t **constant experts** $R_{i,T} = \sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t})$



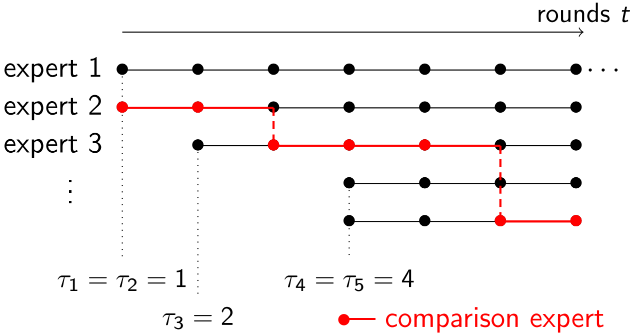
Different perspective on growing experts

- "Specialist" or "abstention trick" $\Longrightarrow$ GrowingHedge
- Controls regret w.r.t **constant experts**
- Implies (by summing) regret bound against **sequences of "fresh" experts** $(i_1, \ldots, i_T)$, *i.e.* sequences that only switch to new experts $R_{i,T} = \sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t})$



Different perspective on growing experts

- "Specialist" or "abstention trick" $\Longrightarrow$ GrowingHedge
- Controls regret w.r.t **constant experts**
- Implies regret bound against **sequences of "fresh" experts** $(i_1, \ldots, i_T)$, *i.e.* sequences that only switch to new experts
- What about **other** comparison sequences $(i_1, \ldots, i_T)$?



Comparing to sequences of experts (fixed M)

- **Tracking the best expert** [Herbster and Warmuth, 1998]: comparing to sequences of experts $(i_1, \ldots, i_T)$ ($k \ll T$ shifts)

$$R_T(i_1, \ldots, i_T) := \sum_{1 \leqslant t \leqslant T} (\ell_t - \ell_{i_t,t})$$

- Inefficient solution : aggregate over M^T sequences of experts $\Longrightarrow$ oracle regret bound of $\approx \frac{1}{\eta}(k+1)\log M + \frac{1}{\eta} k \log \frac{T}{k}$
- Efficient *Fixed Share* algorithm [Herbster and Warmuth, 1998] $\Longrightarrow$ optimal regret bound with $O(M)$ per-round complexity
- Can be seen as aggregation of **sequences** under **Markov chain prior** [Vovk, 1999]

Aggregating sequences of experts

- **Key fact.** When the prior π on sequences of experts is Markovian with transition probabilities $\theta_t(i_t \mid i_{t-1})$

$$\pi(i_1, \ldots, i_T) = \theta_1(i_1)\, \theta_2(i_2 \mid i_1) \cdots \theta_T(i_T \mid i_{T-1})$$

  Hedge collapses to efficient algorithm MarkovHedge with update

$$v_{i,t+1} = \sum_{j=1}^{M} \theta_{t+1}(i \mid j)\, v_{j,t}^{m}$$

The "muting trick"

- In order to transport to the "growing experts" setting, we need

$$x_t = \sum_i v_{i,t} x_{i,t}$$

  to be well-defined
- **Trick:** since θ_t **can be chosen at time** t, take θ_t to **only transition to entered experts**
- Prior π (defined recursively) puts all mass to **admissible** sequences of experts with $i_t \leqslant M_t$ for all t
- Amounts to set $v_{i,t} = 0$ ("muting") for experts i that have not entered yet
- "**Dual**" to "specialist trick", but **more versatile**

FreshMarkovHedge

- Only switch to **new** experts (all mass to sequences of fresh experts)
- Turns out to be **equivalent to GrowingHedge under unnormalized prior**
- Using proper transition probabilities, regret e.g. of the form

$$\frac{1}{\eta}\left((k+1)\log \max_{1\leqslant t\leqslant T} m_t + (k+1)\log T\right)$$

for sequences of **fresh** experts with k shifts

GrowingMarkovHedge

- Transition to both **new** and **incumbent** experts
- Again **anytime**, with $O(M_t)$ per-round complexity
- With proper choice of transition probabilities, regret

$$\frac{1}{\eta}\left((k+1)\log\max_{1\leqslant t\leqslant T} m_t + (k+k_1+2)\log T\right)$$

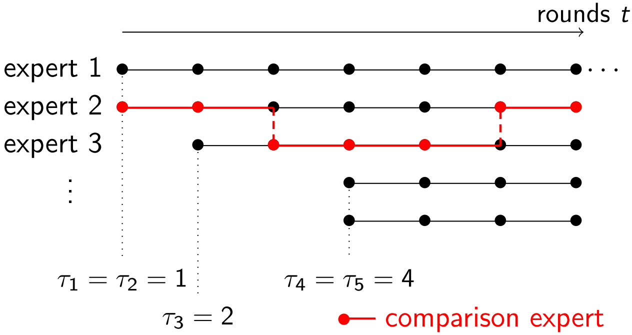
w.r.t. sequences with k switches, among which k_1 to **incumbent** experts, for **all** (*i.e.* **adaptive to**) k and k_1

Tracking a small pool of good experts

- GrowingMarkovHedge covers all **admissible** sequences, with essentially optimal regret bound
- But regret bound can be quite large: of order $\log M_t$ at each switch. Can we do better for **some** sequences ?
- **Tracking a small subset of good experts** (Freund, [Bousquet and Warmuth, 2002]): $n \ll M$ good experts, **sparse sequences** with $k \ll T$ shifts among these n experts
- **Particularly important for a growing number of experts**, as $M_T \to \infty$

Tracking a small pool of good experts

- **Tracking a small subset of good experts** (Freund, [Bousquet and Warmuth, 2002]): $n \ll M$ good experts, **sparse sequences** with $k \ll T$ shifts among these n experts



- **Particularly important for a growing number of experts**, as $M_T \to \infty$

The "sparse" case: fixed M

- Ad-hoc **Mixing Past Posterior** (MPP) algorithm [Bousquet and Warmuth, 2002], with (up to some tuning) regret bound $\approx n \log \frac{M}{n} + k \log n + 2k \log T$
  (log n regret per switch instead of log M)
- Interpreted by [Koolen et al., 2012] as an **aggregation of a structured class of specialists** + new algorithm (no tuning)

Small pool of experts in the growing experts setting

- Slight **reformulation** of [Koolen et al., 2012]'s algorithm: aggregation of **sequences of specialists** (i, a) with i expert and $a \in \{0, 1\}$; $(i, 0)$ always **inactive**, $(i, 1)$ always **active**
- More **flexibility**, necessary to **extend to the "growing" setting**
- Markov prior: transitions only occur between $(i, 0)$ and $(i, 1)$ (both ways) + "**muting trick**": zero mass to $(i, 1)$ as long as $i > M_t$
- Combines the "specialist" and "sequences of experts" viewpoints
- GrowingSleepingHedge: **Anytime** and **efficient** + regret bound (up to time T, k shifts among n base experts) of

$$\approx \frac{1}{\eta}\left(n \log \frac{\max_{1 \leqslant t \leqslant T} m_t}{n} + 2k \log T\right)$$

Conclusion

- **Specialist setting/trick**: most **natural** approach
- But can be somewhat **less appealing/seamless** beyond constant experts
- **Sequences of experts** = **more flexible approach** (recovers "Growing Hedge" as a particular case)
- **Generic algorithms** (esp. **efficient aggregation of structured classes of experts**) + **encode the "growing" structure in the prior** (can be done **on the fly**)
- Leads to **efficient and simple anytime algorithms** with **adaptive** regret bounds for various comparison classes + **conceptually transparent proofs**

Thank you !

References I

Bousquet, O. and Warmuth, M. K. (2002).
Tracking a small set of experts by mixing past posteriors.
The Journal of Machine Learning Research, 3:363–396.

Cesa-Bianchi, N. and Lugosi, G. (2006).
Prediction, Learning, and Games.
Cambridge University Press, Cambridge, New York, USA.

Chernov, A. and Vovk, V. (2009).
Prediction with expert evaluators' advice.
In *Algorithmic Learning Theory (ALT)*, pages 8–22.

References II

Freund, Y., Schapire, R. E., Singer, Y., and Warmuth, M. K. (1997).
Using and combining predictors that specialize.
In *ACM Symposium on Theory of Computing (STOC)*, pages 334–343.

Herbster, M. and Warmuth, M. K. (1998).
Tracking the best expert.
Machine Learning, 32(2):151–178.

Koolen, W. M., Adamskiy, D., and Warmuth, M. K. (2012).
Putting bayes to sleep.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 135–143.

References III

Vovk, V. (1999).
Derandomizing stochastic prediction strategies.
Machine Learning, 35(3):247–282.