Accepted Manuscript

Dynamic benchmark targeting

Karl H. Schlag, Andriy Zapechelnyuk

PII: S0022-0531(17)30021-2
DOI: http://dx.doi.org/10.1016/j.jet.2017.02.004
Reference: YJETH 4639

To appear in: *Journal of Economic Theory*

Received date: 4 August 2015
Revised date: 22 January 2017
Accepted date: 10 February 2017

Please cite this article in press as: Schlag, K.H., Zapechelnyuk, A. Dynamic benchmark targeting. *J. Econ. Theory* (2017), http://dx.doi.org/10.1016/j.jet.2017.02.004

Dynamic Benchmark Targeting*

Karl H. Schlag
University of Vienna‡

Andriy Zapechelnyuk
University of Glasgow§

February 14, 2017

Abstract

We study decision making in complex discrete-time dynamic environments where Bayesian optimization is intractable. A decision maker is equipped with a finite set of benchmark strategies. She aims to perform similarly to or better than each of these benchmarks. Furthermore, she cannot commit to any decision rule, hence she must satisfy this goal at all times and after every history. We find such a rule for a sufficiently patient decision maker and show that it necessitates not to rely too much on observations from distant past. In this sense we find that it can be optimal to forget.

Keywords: Dynamic consistency, experts, regret minimization, forecast combination, non-Bayesian decision making.

JEL classification numbers: C44, D81

*The authors thank Olivier Gossner, Sergiu Hart, Sebastian Koch, and Gábor Lugosi for valuable comments. Karl Schlag gratefully acknowledges financial support from the Department of Economics and Business of the Universitat Pompeu Fabra, Grant AL 12207, and from the Spanish Ministerio de Educacion y Ciencia, Grant MEC-SEJ2006-09993.

‡Department of Economics, University of Vienna, Hohenstaufengasse 9, 1010 Vienna, Austria. *E-mail:* karl.schlag@univie.ac.at.

§*Corresponding author.* Adam Smith Business School, University of Glasgow, University Avenue, Glasgow G12 8QQ, UK. *E-mail:* azapech@gmail.com

1 Introduction

We are concerned with decision making in discrete-time dynamic environments that are hard to predict and to model explicitly, due to complexity or lack of information.

How would a firm optimally choose its inventories if the demand for its product is stochastic and subject to unpredictable structural breaks?

How would a police department decide about the number of police cars and their patrol routes if crimes do not follow any stationary pattern?

How should patients in an emergency ward be assigned to doctors if there is no discernible system in arrival of patients with different urgency of medical attention?

In economics, the standard approach to dynamic decision making involves modelling the environment as a specific stochastic process and then optimizing within this model. Unknown parameters of the process are estimated by statistical methods. However, this approach typically comes with several problems. Different assumptions on the underlying stochastic process lead to different solutions, and the true environment is never known. Explicit and tractable solutions only exist for simplest scenarios. Complex models that include more realistic features, such as structural breaks at unknown time, easily make the problem intractable. Tractable models often cannot approximate the real environment, resulting in serious errors in decision making.

An alternative approach that is popular in machine learning considers decision making with expert advice and the well known no-regret problem.[1] It can deal with environments of arbitrary complexity—in fact, the modeller does not even need to know anything about the environment. In this approach, the decision maker is equipped with a finite set of benchmark strategies or experts that she uses as targets. Her objective is to perform similarly to or better than each of them, without making any specific assumptions about the environment. These benchmark strategies could be simple heuristic decision making rules, standard practices in the given situation, solutions to the problem under specific assumptions about the environment, or strategies of experts who know more about the environment than the decision maker does. However, this approach has two caveats from an economist's point of view. First, a decision maker is infinitely patient, there is no discounting of payoffs in practically all papers. Second,

[1] For a survey of this literature see Cesa-Bianchi and Lugosi (2006).

the decision maker has the power to commit to a decision rule, as the performance is only measured at the outset.

This paper addresses these two caveats by inserting a new pair of elements into decision making with expert advice: *discounting of future payoffs* and *dynamic consistency*. We refer to our methodology as *dynamic benchmark targeting*. We design a decision-making rule that dynamically combines benchmark strategies and achieves a similar or superior present-value performance to each of them in all environments, at each point of time, provided the decision maker is sufficiently patient.

Freedom of choice and the absence of legal instutions that hardwire the behavior make dynamic consistency a necessity. Dynamic consistency, while being a standard assumption for economists (see Strotz, 1956; Rubinstein, 1998), is a novel feature that our paper introduces to the literature on decision making with expert advice. A decision rule is dynamically consistent if it performs well at any point in time, not only ex-ante. The decision maker does not commit to any course of actions from the start. She asks herself in every period, after every history, whether the previously chosen strategy will continue to perform well enough relative to the set targets and whether she should continue using it. However, all the literature on decision making with expert advice assumes commitment to a particular strategy from the start, and thus ignores dynamic consistency.[2] All but two decision rules used in this literature are not dynamically consistent. The two exceptions are discussed at the end of this section.

Discounting of future payoffs is the paradigm in economic decision making in which one is forward-looking and considers tradeoffs over time. The literature on decision making with expert advice considers a backward-looking decision maker concerned with sums or simple averages of payoffs. The two exceptions are Fudenberg and Levine (1999) and Olszewski and Peski (2011) that study decision makers who care about future discounted payoffs, but assume commitment to a particular strategy from the start, thus ignoring dynamic consistency.

Having both features, dynamic consistency and discounting of future payoffs, makes Blackwell's Approachability Theorem (Blackwell, 1956) and its extensions (Lehrer, 2003; Lehrer and Solan, 2009) inapplicable. A different method has to be used in this case.

[2]Some papers consider infinitely patient decision makers who care about long-run average streams of payoffs, so the dynamic consistency issue does not arise (e.g., Hart and Mas-Colell, 2000, 2001).

We show that dynamic consistency is intimately linked to the ability to react to recent changes in the environment. When evaluating past performance, a dynamically consistent decision rule must not place the same weight on all past events, recent and distant alike. Instead, recent periods should carry a greater weight, as if one gives recent events more attention. There are many ways to accomplish that, such as assignment of exponentially decaying weights, equal wights with bounded recall, or equal weights with periodic restarts by forgetting all past information. It turns out that there is a strategy that provides better results, both numerically and asymptotically. It requires the decision maker to have a bounded recall, m, and in every period t to average out the payoffs over last k_t periods, where each k_t is an independent and uniformly random draw from $\{1, ..., m\}$. The recall length m is chosen by the decision maker. We derive an optimal length of recall and show that the achieved performance is arbitrarily close to that of the best benchmark strategy, provided the decision maker is sufficiently patient. Importantly, we provide an upper bound on how close the decision maker's performance is to the best benchmark, for discount factors bounded away from one.

Infinite patience is a simplifying model of someone who is very patient or who looks far ahead. In the context of this paper we discover that being very patient cannot be approximated by infinite patience. There is a discontinuity as the discount factor tends to one, as we further comment on in Section 4.

Related Literature. The problem of outperforming in hindsight a given set of benchmark strategies, or the no-regret problem, was first considered by Hannan (1957). A substantial literature revisited the problem and offered solutions for a variety of applications. This methodology is used, among other things, to combine competing forecasting models (Foster and Vohra, 1993, 1999; Littlestone and Warmuth, 1994),[3] to design investment portfolios and derive bounds on the prices of financial instruments (DeMarzo et al., 2006, Chen and Vaughan, 2010), to investigate learning in games (Fudenberg and Levine, 1995; Freund and Schapire, 1999; Hart and Mas-Colell, 2000, 2001) and to compute efficient algorithms for job scheduling (Mansour, 2010), online routing and shortest paths problems (Takimoto and Warmuth, 2003; Blum et al., 2006).[4] In the above literature, a decision maker commits to a decision rule in the initial period.

[3]For an overview of the forecast combination literature see Timmerman (2006) and Clemen and Winkler (2007).

[4]This is also related to how to aggregate opinions (Larrick and Soll, 2006; Jose et al., 2014).

In this paper we investigate what happens if the decision maker does not have the power to commit, and hence is tempted to change the rule at later points in time.

Strategies with diminishing weights on past observations have been introduced in the literature, but without concern neither for dynamic consistency, nor for optimization of the discounted sum of future payoffs. Cesa-Bianchi and Lugosi (2006, Ch. 2.11) evaluate the past regret as a sum of past single-period regrets with diminishing weights and show that the past regrets can vanish if and only if the sum of the weights diverges. Lehrer and Solan (2009) consider a decision maker who periodically erases her memory, and show that no regret can be achieved with a sequence of strategies with bounded recall. While Cesa-Bianchi and Lugosi (2006) and Lehrer and Solan (2009) only consider performance from the ex-ante perspective, we show that their strategies are dynamically consistent. However, the aim of our paper is not a mathematical proof of existence, but a formulation of methodology and design of a decision rule with good properties. Our rule provides a bound on the performance that is superior to those that we derive for the rules of Cesa-Bianchi and Lugosi (2006) and Lehrer and Solan (2009).

Our paper also connects to the psychology and experimental literature that documented the so-called *recency effect*, according to which more distant events are regarded as less relevant.[5] In the forecast combination literature, the recency effect, manifested as diminishing weights on past events, have been used as a heuristic or empirical performance improvement tool, e.g., Bates and Granger (1969), Winkler and Makridakis (1983), Timmerman (2006), Sánchez (2008), and Mallet et al. (2009). In this paper we identify a novel strategic reason for the recency effect (Section 4).

2 Example

For illustration, let us consider the inventory control problem. Even though our example is simple and stylized, it has enough structure to demonstrate how difficult and complex it would be to solve it by standard economic methods. The example shows how our approach tackles the problem while disregarding its complexity.

[5]This goes back at least to Watson (1930) and Guthrie (1952). See, e.g., Roth and Erev (1995), Erev and Roth (1998), Camerer and Ho (1999), Ray and Wang (2001) and the references within.

A retailer decides what quantity of a product to hold in stock at the beginning of each day $t = 1, 2, \ldots$ The product can be restocked every morning for free, but overnight storage of unsold goods is costly. The daily demand for the product, q_t, follows a stochastic process that we make no assumptions about. The tradeoff is that a larger stock in the morning means more profit in a day with high demand, but more storage cost in a day with low demand. As usual in business decision making, the retailer's strategy can be adjusted any time. Therefore, it is necessary to consider dynamically consistent decision rules.

There are three parameters: a per-unit profit from sales π, a per-unit cost of overnight storage c, and an annual interest rate r. Denote by s_t the stock at the beginning of day t (after the restocking has taken place) and let $y_t = \min\{s_t, q_t\}$ be the daily sales. Then $s_t - y_t$ is the amount of goods left overnight. The retailer's profit in day t is thus

$$\pi y_t - c(s_t - y_t).$$

The *performance* evaluated at day t_0, measured as the normalized present value of future profits, is

$$\Pi_{t_0} = (1-\delta)\sum_{t=t_0}^{\infty} \delta^{t-t_0}\big(\pi y_t - c(s_t - y_t)\big),$$

where δ is the daily discount factor, $\delta = e^{-r/365}$.

The standard approach to this problem requires specific assumptions about the stochastic process $\{q_t\}$ that determines demand. Since nothing is known about the stochastic process, any specific assumptions may be unwarranted, resulting in an inadequate solution. Moreover, if the stochastic process is not ergodic, tractability can become a problem. For instance, this is the case when there are structural breaks in demand occurring at unknown points in time.

Our approach circumvents these difficulties. Instead of focusing on the environment, the retailer focuses on a few benchmark strategies that are candidates for being "good" decision rules. Her aim is to perform similar to or better than each of them.

One such benchmark could be, for example, the *fixed quantity system* that dictates to restock daily up to a fixed quantity $s_0 > 0$. Another such benchmark could be the *replenishment system* that dictates to restock up to a level $\bar{s}$ whenever the stock has fallen below $\underline{s}$. The three parameters, s_0, $\underline{s}$, and $\bar{s}$, for instance, can be updated daily

based on past data.

We design a decision rule that performs in any environment similarly to or better than each of these two benchmarks. In particular, if the true environment is i.i.d., in which case the fixed quantity system is optimal, then our rule will approximately follow the fixed quantity system. If instead the environment turns out to be Markov, where the replenishment system performs well, then the retailer will also performs well by using our rule.

In reality, benchmarks perform differently at different points of time. Our decision rule tracks which one performs better and keeps the retailer's performance within a small error ε of that at all times. It does so by balancing between the two benchmarks, combining their actions with dynamically adjusted weights that depend on the benchmarks' past performance, as described in Section 3.2 below. In particular, it is enough that one of the benchmarks performs well in order to guarantee good profits with our rule.

How well does our rule perform? Clearly we cannot expect such a benchmark combining rule to perform better than all the benchmarks. We show that our rule may perform worse than the best benchmark, but only by a fairly small amount called the *error bound*. For example, let the annual interest rate be 5%, so the daily discount factor is $\delta = e^{-0.05/365} \approx 0.999863$. According to our formula (3) in Section 3.2, the error bound of our rule is 4.3% of the daily profit range.[6] So, the present value of the retailer's future profits is guaranteed, at any time, to be at least as much as that of the best benchmark strategy minus 4.3% of the daily profit range.

3 Dynamic Benchmark Targeting

3.1 Model

We now introduce the formal model of benchmark targeting. A decision maker takes actions in discrete time periods $t = 1, 2, ...$ In each period t the decision maker chooses an action a_t from a set A of available actions. Then, a state of environment, $\omega_t \in \Omega$,

[6] Our decision rule's actions are convex combinations of actions dictated by the benchmarks, so the maximum stock s_t can never exceed $\bar{s}$. Hence, the daily profit is within $[-c\bar{s}, \pi\bar{s}]$, and the maximum profit variation is $(\pi - c)\bar{s}$.

is realized and observed by the decision maker. The decision maker's payoff in that period depends on both a_t and ω_t and is denoted by $u(a_t, \omega_t)$.

In each period t, before the decision maker makes her choice, she is provided with recommendations of n benchmark strategies. Each benchmark strategy i recommends an action $r_t(i) \in A$. Then, the decision maker chooses an action $a_t \in A$ as a function of the benchmark recommendations $r_t(1), ..., r_t(n)$, as well as all past states and recommendations. These benchmarks could be simple heuristic decision making rules, standard practices in the given situation, solutions to the problem under specific assumptions about the environment, or experts who know more about the environment than the decision maker does. The important assumption is that benchmark recommendations do not depend on choices of the decision maker.

We make the following assumptions. The set of actions A is a convex and compact subset of $\mathbb{R}^d$, $d \geq 1$. The payoff function u is uniformly bounded, w.l.o.g. $u(a, \omega) \in [0, 1]$. In addition, $u(a, \omega)$ is concave in a for every $\omega \in \Omega$.[7] The state space Ω is a compact space (finite or infinite). The sequence of states of environment, $\bar{\omega} = \{\omega_t\}_{t=1}^{\infty}$, is arbitrary. For example, it can be determined by a discrete-time stochastic process, which we make no assumptions about.

The profile of a realized state of the environment and the actions recommended by the n benchmarks in period t is denoted by $x_t = (\omega_t, r_{1,t}, \ldots, r_{n,t})$ and called an *event* in period t. Let $h_t = (x_1, \ldots, x_t)$ be the history of the events up to period t and let $\mathcal{H}$ be the set of all finite histories, including the empty history. For each $i = 1, \ldots, n$, a *benchmark strategy* is described by a map $p_i : \mathcal{H} \to A$ that associates with every history h_{t-1} an action $r_t(i)$ in A. A *decision rule* of the decision maker is a map $p : \mathcal{H} \times A^n \to A$ that associates with every history h_{t-1} and every profile of current recommendations $r_t = (r_t(1), \ldots, r_t(n))$ an action r_t in A to be chosen in period t.

In what follows, for a given set of benchmark strategies $p_1, ..., p_n$ and a given decision rule p, we write for each period t

$$r_t = \big(p_1(h_{t-1}), ..., p_n(h_{t-1})\big) \quad \text{and} \quad a_t = p(h_{t-1}, r_t).$$

These notations permit two interpretations of r_t that are equivalent for our purpose. Ei-

[7] The convexity of A and concavity of u in the action implies that every mixed action (lottery over actions) is dominated by a pure action in A. So there is no need to consider mixed actions. The extension of the model to finite and non-convex actions sets is discussed in Section 5.

ther the decision maker knows benchmark strategies $(p_1, ..., p_n)$, and hence can deduce their recommendations, or she directly observes the recommendations of the benchmarks and does not need to know their strategies.

For a given sequence of states, $\bar{\omega} = \{\omega_t\}_{t=1}^{\infty}$, the performance of a decision rule p from the perspective of period t_0 is measured as the (normalized) discounted sum of future payoffs that this decision rule delivers,

$$U_{t_0}(p, \bar{\omega}) = (1-\delta) \sum\nolimits_{t=t_0}^{\infty} \delta^{t-t_0} u(a_t, \omega_t), \tag{1}$$

where $\delta \in (0,1)$ is the decision maker's discount factor. The performance of each benchmark i from the perspective of period t_0 is the discounted sum of future payoffs that the decision maker can obtain by always following the recommendations of i,

$$U_{t_0}(p_i, \bar{\omega}) = (1-\delta) \sum\nolimits_{t=t_0}^{\infty} \delta^{t-t_0} u(r_t(i), \omega_t).$$

We now introduce "dynamic benchmark targeting." The decision maker wishes to guarantee the performance within a given error bound of, or better than, the performance of each benchmark strategy, under each possible sequence of states, and from perspective of each period of time. "Benchmark targeting" refers to the decision maker's goal to outperform all the benchmarks in a given set, allowing only a limited error margin. "Dynamic" refers to the dynamic consistency of the objective of being within the same error bound after every possible past history.[8]

Definition 1. A decision rule p *for dynamic benchmark targeting w.r.t. benchmark strategies* $p_1, \ldots, p_n$ *has error bound* ε if

$$U_{t_0}(p, \bar{\omega}) \geq \max_{i \in \{1,\ldots,n\}} U_{t_0}(p_i, \bar{\omega}) - \varepsilon \tag{2}$$

for all periods $t_0 = 1, 2, ...$ and for all sequences of states $\bar{\omega}$.

Note that condition (2) is a sure inequality that has to hold for every realized sequence of states. Performance measures $U_{t_0}(\cdot, \bar{\omega})$ are not expected utilities, these are the

[8]This objective is analogous to the ε-sequential optimality notion, or ε-subgame perfect equilibrium in repeated games (see Radner (1980) and Mailath et al. (2005)), where ε is the tolerance level that keeps the decision maker from changing her behavior so long as the payoff is within ε of the optimum.

discounted sums of the future payoffs that will be realized under the given sequence of states $\bar{\omega}$.

We assume that past states are observable. Thus, the decision maker can calculate in retrospect for each period in the past what she would have achieved with each action. In fact, for our results to hold, observability of past states is unnecessary, so long as the decision maker can observe her (foregone) payoffs that she would have received if she had followed any particular benchmark.[9]

For clarity of exposition we have assumed that each benchmark i's recommendation is a deterministic function $p_i(h_t)$ of past states (and benchmark recommendations). Our model can deal with arbitrary sequences of states and recommendations, where a recommendation in each period is simply an action. No assumptions are necessary about how they are generated. In particular, such sequences can be realizations of a stochastic process where states and benchmark recommendations are interdependent. However, an important assumption is that these sequences of states and recommendations are exogenous to the decision maker's problem and do not depend on the choices made by the decision maker. Otherwise, a decision rule with a small error bound need not exist, because some actions of the decision maker may trigger an irreversible change in all future payoffs.[10]

3.2 Decision Rule

We now introduce a simple decision rule for dynamic benchmark targeting and then present an error bound that shows how close it can track the best benchmark at any point in time. According to this decision rule, in each round the decision maker chooses a convex combination of the recommendations of the n benchmark strategies. The benchmarks that made better past recommendations receive greater weights.

[9] Actually, it suffices to have unbiased estimates of payoffs of each benchmark in each period. Everything goes through after replacement of realized performance by expected performance. We use this insight in Section 5 to apply our methodology to the case where only payoffs from chosen actions are observed.

[10] For example, consider the problem with two states, $\Omega = \{0, 1\}$, where the decision maker aims to guess the state in each period, $u(a_t, \omega_t) = 1 - |\omega_t - a_t|$, $a_t \in A = [0, 1]$. The nature picks $\omega_1 \in \{0, 1\}$, equally likely. If the decision maker has guesses ω_1 correctly, then the nature repeats the same state forever. Otherwise the nature is i.i.d. uniformly random forever. Among the two constant benchmarks, one always guessing 0 and the other always guessing 1, one of them guarantees the maximum normalized discounted payoff, 1. But the decision maker can only guarantee $\frac{1}{2}$. This issue in the context of the standard no-regret problem is highlighted in Schlag and Zapechelnyuk (2012).

Fix a period t with a history h_{t-1}. For each benchmark i denote by $C_{t,k}(i)$ the aggregate payoff over the last $k \geq 1$ periods,

$$C_{t,k}(i) = \sum\nolimits_{s=t-k}^{t-1} u(r_s(i), \omega_s),$$

where we define $u(\cdot, \omega_s) = 0$ for all $s \leq 0$, i.e., all payoffs prior to the first period are set equal to zero. Let $C_{t,0}(i) = 0$. Define for $k \geq 0$ the *k-score* of each benchmark i as the logistic weight of these aggregate payoffs,

$$\lambda_{t,k}(i) = \frac{e^{\eta C_{t,k}(i)}}{\sum_{j=1}^{n} e^{\eta C_{t,k}(j)}},$$

where $\eta \geq 0$ is a parameter. Then compute the average of k-scores of each benchmark i for k from 0 to $m-1$, $\frac{1}{m}\sum_{k=0}^{m-1} \lambda_{t,k}(i)$, where m is an integer parameter. Note that the average scores of all benchmarks add up to one. The decision rule $p_{(m,\eta)}$ that depends on m and η combines the benchmark recommendations by assigning to each recommendation $r_t(i)$ the weight equal to i's average score,

$$p_{(m,\eta)}(h_{t-1}, r_t) = \sum_{i=1}^{n} \left(\frac{1}{m} \sum_{k=0}^{m-1} \lambda_{t,k}(i) \right) r_t(i).$$

In this way the agent chooses a convex combination of the recommendations.

The decision rule has two free parameters, m and η. The value $m-1$ is the maximal number of previous periods that are included in the performance evaluation, whereas η is a sensitivity coefficient used in the logistic formula. We choose these free parameters to optimize the order of convergence of the error bound as δ approaches 1.

Theorem 1. *For a discount factor $\delta \in (0,1)$ let*

$$\eta = 2^{\frac{4}{3}} (\ln n)^{\frac{1}{3}} (1-\delta)^{\frac{1}{3}} \quad \textit{and} \quad m = \frac{\eta}{2(1-\delta)} + x,$$

where $x \in (-1, 1]$ is the adjustment such that m is an even integer. Decision rule $p_{(m,\eta)}$ has error bound

$$\varepsilon = \frac{3}{4}\big(2(1-\delta)\ln n\big)^{\frac{1}{3}} + \frac{7}{96}\big(2(1-\delta)\ln n\big)^{\frac{2}{3}}. \tag{3}$$

The proof is in Appendix A.

Note that $\varepsilon \to 0$ as $\delta \to 1$. Hence, if the decision maker is sufficiently patient, or if the interval between two consecutive periods is small, then the decision maker can be guaranteed to perform arbitrarily close to, or better than, the best benchmark strategy, from the perspective of any period.

Periods per year	$\delta = e^{-0.05/T}$	n	m	η	ε
$T = 365$	0.999863	2	420	0.115	4.3%
		4	528	0.145	5.5%
		10	626	0.172	6.5%
$T = 52$	0.999039	2	114	0.220	8.3%
		4	144	0.277	10.5%
		10	170	0.328	12.5%
$T = 12$	0.995842	2	44	0.358	13.7%
		4	54	0.451	17.3%
		10	64	0.535	20.1%

Table 1: Numerical percentage values of the error bound.

Our error bound and the optimal parameters m and η are easily computed for specific values of δ and n. Table 1 demonstrates the numerical value of the error bound for the case of annual interest rate $r = 0.05$ and payoffs evaluated daily ($\delta = e^{-0.05/365} \approx 0.999863$), weekly ($\delta = e^{-0.05/52} \approx 0.999039$) and monthly ($\delta = e^{-0.05/12} \approx 0.995842$), with the number of benchmarks $n = 2$, 4, and 10. As evident from (3) and illustrated numerically by Table 1, the magnitude of the value of $1-\delta$ (or the frequency of periods T) plays an exponentially greater role on the size of the error bound, as compared to the number of benchmarks n. Doubling $(1-\delta)$ has the same effect on the error bound as making n squared.

3.3 Comparison to Other Decision Rules

The decision rule $p_{(m,\eta)}$ defined in the previous section performs well when δ is large. One wonders whether alternative, possibly even simpler rules perform similarly. In this section we present the error bounds of a few alternative rules. Then, in the next section, we proceed to derive general necessary properties of rules with low error bounds.

Let $B_t(i)$ denote the evaluation of the past performance of each benchmark i. We consider four rules that combine the recommendations of the benchmarks by the ex-

ponential weights of their past performances $B_t(i)$,

$$q(h_{t-1}, r_t) = \sum_{i=1}^{n} \left(\frac{e^{\eta B_t(i)}}{\sum_{j=1}^{n} e^{\eta B_t(j)}} \right) r_t(i). \tag{4}$$

These four rules use the same formula (4) to combine benchmarks, but differ in how they evaluate benchmark past performance, $B_t(i)$.

First, consider the *exponentially weighted average forecaster* rule introduced in Littlestone and Warmuth (1994). This rule aggregates all past payoffs from the start for each benchmark i,

$$B_t(i) = \sum\nolimits_{s=1}^{t-1} u(r_s(i), \omega_s). \tag{5}$$

This rule has the error bound at least $1/2$. In fact, in Section 4 we prove a more general result that any decision rule that relies "too much" on the distant past will have a large error bound (Theorem 2). Such decision rules include, among others, calibrated forecasting of Foster and Vohra (1993, 1999), smooth fictitious play of Fudenberg and Levine (1995), and regret matching of Hart and Mas-Colell (2000, 2001).

As a good decision rule must focus on the recent past, a natural candidate is the rule that aggregate payoffs only over the last m periods for a fixed parameter m,

$$B_t(i) = \sum\nolimits_{s=t-m}^{t-1} u(r_s(i), \omega_s).$$

This rule is not satisfactory either, since its error bound is bounded away from zero. It does not converge to zero as $\delta \to 1$. It is possible to construct an example, as in Zapechelnyuk (2008), where the decision maker's and benchmarks' performances are cyclical (the cycle length is a function of the length of recall m), so the decision maker underperforms relative to some benchmark by a constant that is independent of m.

Another simple possibility is to make the decision maker periodically "forget" the past and start anew. This periodic-restart rule, considered in Lehrer and Solan (2009), evaluates the past performance of each benchmark i by its aggregate payoff since the last restart,

$$B_t(i) = \sum\nolimits_{s=\rho(t)}^{t-1} u(r_s(i), \omega_s),$$

where restarts occur in periods m, $2m$, $3m$,..., and $\rho(t) = m\lfloor \frac{t-1}{m} \rfloor$ denotes the period of restart preceding t. Denote by $\bar{q}_{(m,\eta)}$ the periodic-restart rule with restart period m

defined by (4), where $B_t(i)$ is defined above. We now present an error bound of this rule.

Proposition 1. *For every $\delta \in (0,1)$ there exists (m,η) such that the periodic-restart rule $\bar{q}_{(m,\eta)}$ has the error bound*

$$\varepsilon = \left(\tfrac{3}{2}\right)^{4/3} (\ln n)^{1/3}(1-\delta)^{1/3}. \tag{6}$$

The proof is in Appendix B.

Lastly, we consider the decision rule that places exponentially decaying weights to more distant periods, referred to as the exponential-decay rule,

$$B_t(i) = \sum_{s=1}^{t-1} \alpha^{t-s} u(r_s(i), \omega_s) \tag{7}$$

for some $\alpha \in (0,1)$. This is a special case of Cesa-Bianchi and Lugosi's (2006, Ch. 2.11) rule of aggregation of the past performance with diminishing weights.

Denote by $\tilde{q}_{(\alpha,\eta)}$ the exponential-decay rule defined by (4) with the above choice of $B_t(i)$. We now determine the error bound of this rule.

Proposition 2. *For every $\delta \in (0,1)$ there exists (α,η) such that the exponential-decay rule $\tilde{q}_{(\alpha,\eta)}$ has the error bound*

$$\varepsilon = \tfrac{3}{2}(\ln n)^{1/3}(1-\delta)^{1/3} + \tfrac{1}{2}(\ln n)^{2/3}(1-\delta)^{2/3}. \tag{8}$$

The proof is in Appendix B.

The rates of convergence of the error bounds in Propositions 1 and 2 are the same as that of our rule $p_{(m,\eta)}$, but their leading constants are substantially larger. For the periodic-restart rule the leading constant is $\left(\frac{3}{2}\right)^{4/3} \approx 1.717$ and for the exponential-decay rule the leading constant is $\frac{3}{2} = 1.5$, while the leading constant for our rule is $\frac{3}{4}2^{1/3} \approx 0.945$, where $1.717 > 1.5 > 0.945$. The error bounds of these two rules are also compared to the error bound of our rule numerically in Table 2.

Intuitively, the periodic-restart rule performs worse than our rule because of fixed restart periods. The "adverse" nature can exploit the knowledge of restart periods by changing which benchmark is best half way to the next restart to make the rule

Periods per year	$\delta = e^{-0.05/T}$	n	Our Rule	Periodic Restart	Exponential Decay
$T = 365$	0.999863	2	4.3%	7.8%	7.0%
		4	5.5%	9.9%	8.8%
		10	6.5%	11.7%	10.4%
$T = 52$	0.999039	2	8.3%	15.0%	13.5%
		4	10.5%	18.9%	17.1%
		10	12.5%	22.4%	20.4%
$T = 12$	0.995842	2	13.7%	24.4%	22.4%
		4	17.3%	30.8%	28.5%
		10	20.1%	36.5%	34.1%

Table 2: Numerical comparison of error bounds of three decision rules.

perform badly when evaluated from the perspective of this period. The idea to avoid this vulnerability by concealing the periods of restart led to the construction of our rule.

The reason why our rule performs better than the exponential-decay rule roots in our method of proof. Our derivation of the error bounds relies on Cesa-Bianchi and Lugosi's (2006, Theorem 2.3) tight bound on simple (unweighted) sums of single-period losses. The uniform distribution of past windows used in our rule translates nicely into simple sums of losses, whereas it is more difficult to translate the sum of exponentially weighted losses into weighted simple sums. Another intuitive reason for a better performance of our rule is its restriction of the number of recent periods involved in making the next choice. The intuition brought forward in the next section is that sufficiently old observations should simply be ignored, not even included with exponentially small weights.

4 The Role of Adaptation

In this section we identify necessary conditions for a decision rule to have a low error bound. We will argue that a key issue in the design of such decision rules is the appropriate choice of the weights on past information. The error bound ε remains bounded away from zero as $\delta \to 1$ if the decision rule adapts to new information too fast or too slow. Too much weight on the recent past makes the rule susceptible to noise and prevents learning which benchmark is best. Too much weight on the distant

past makes the rule sluggish and unable to track recent changes the performance of the benchmarks, and hence of which benchmark performs best.

We consider a subclass of decision rules $\mathcal{P}$ described as follows. Every rule $p \in \mathcal{P}$ chooses an action at each period t equal to the convex combination of the benchmark's recommendations $(r_t(1), ..., r_t(n))$,

$$p(h_{t-1}, r_t) = \sum_{i=1}^{n} \mu_t(i) r_t(i),$$

with weights $(\mu_t(1), ..., \mu_t(n))$ satisfying the following two conditions.

Monotonicity. For each period t and each benchmark i, weight $\mu_t(i)$ is weakly increasing in the past performance of benchmark i, ceteris paribus. Formally, for any two sequences of states, $\bar{\omega}$ and $\bar{\omega}'$, that differ only in the payoff of benchmark i in some period $s < t$, if $u_s(r_s(i), \omega_s) > u_s(r_s(i), \omega'_s)$, then weight $\mu_t(i)$ is weakly greater under $\bar{\omega}$ than under $\bar{\omega}'$.

Anonymity. Names of benchmarks do not matter. The weights $(\mu_t(1), ..., \mu_t(n))$ are invariant under permutation of indices $(1, ..., n)$.

For each rule in class $\mathcal{P}$ we define a measure of adaptivity, that is, the degree to which the rule adapts to new information, and then show how adaptive a rule has to be in order to generate a low error bound. Our measure of adaptivity is defined by looking at sequences in which each benchmark recently only generated the extreme payoffs 0 or 1. We call a rule *at most k-adaptive* for $k \in \mathbb{N}$ if it puts a weakly greater weight on benchmark i whenever in the last k periods benchmark i received 1 while all other benchmarks received 0. We call a rule *k-adaptive* if it is not at most $k-1$ adaptive. If no such k exists, then the decision rule is called *unadaptive*. Formally, we say that a decision rule $p \in \mathcal{P}$ is *k-adaptive* if k is the smallest integer that satisfies for each period $t \geq k+1$,

> if $u(r_s(i), \omega_s) = 1$ and $u(r_s(j), \omega_s) = 0$ for all $j \neq i$ and all $s \in \{t-k, ..., t-1\}$
> then $\mu_t(i) \geq \mu_t(j)$ for all $j \neq i$.

Obviously, every monotonic and anonymous decision rule with weights that depend only the recent m periods is at most m-adaptive. This applies to our decision rule $p_{(m,\eta)}$ defined in the previous section, as well as Lehrer and Solan's (2009) rule with periodic

restarts and Zapechelnyuk's (2008) rule with a bounded recall window. The decision rule with exponentially decaying weights (7) is k-adaptive, where k is the median of the exponential distribution, the smallest integer satisfying $\sum_{s=1}^{k} \alpha^s \geq \sum_{s=k+1}^{\infty} \alpha^s$. The exponentially weighted average forecaster rule (5), as well as any rule based on the sum or simple average of all past payoffs, is unadaptive.

Theorem 2. *Every k-adaptive decision rule in $\mathcal{P}$ has error bound*

$$\varepsilon \geq \max\left\{\frac{1}{2^{(k+1)\log(k+1)}}, \frac{1-\delta^{k-1}}{2}\right\}.$$

Every unadaptive decision rule in $\mathcal{P}$ has error bound $\varepsilon \geq \frac{1}{2}$.

The proof is in Appendix A.

Theorem 2 shows that a decision rule whose error bound approaches 0 as δ tends to 1 must necessarily be increasingly adaptive w.r.t. δ, but not too adaptive. The adaptivity parameter $k = k(\delta)$ must diverge as $\delta \to 1$, but it must grow slower than $(\ln \delta)^{-1}$, so that both bounds, $\frac{1}{2^{(k+1)\log(k+1)}}$ and $\frac{1-\delta^{k-1}}{2}$, approach zero.

In particular, we uncover a discontinuity at $\delta = 1$. The unadaptive strategies used in the literature on no-regret and decision making with expert advice that are known to perform well for an infinitely patient decision maker, such as Littlestone and Warmuth's (1994) exponentially weighted average forecaster rule, Hart and Mas-Colell's (2000) regret matching, $l_{\mathbf{p}}$-norm strategies of Hart and Mas-Colell (2001) and Cesa-Bianchi and Lugosi (2003), as well as the smooth fictitious play (Fudenberg and Levine, 1995), perform very badly when δ is less than, but arbitrarily close to 1.

To obtain these lower bounds, we test a rule against specific environments. First we explain what can go wrong if too little weight is given on the distant past. The corresponding bound is $\varepsilon \geq \frac{1}{2^{(k+1)\log(k+1)}}$. We obtain this bound by testing a rule in an i.i.d. environment and evaluating its performance in expectation. Note that any bound on the expected performance is also a lower bound on the realized performance. When a rule is k-adaptive, then unlikely sequences of events will be too influential on the decisions and can steer the rule away the best benchmark.

Consider the following example. There are two possible states of the environment, *Rain* and *Sun*. In each period the decision maker is asked to forecast the likelihood of *Rain*, denoted by a. If *Rain* occurs she receives payoff a, if *Sun* occurs she receives payoff

$1-a$. There are two constant benchmarks: one always forecasts *Rain*, the other always forecasts *Sun*. Suppose that states *Rain* and *Sun* occur with probability $1-\sigma$ and σ, respectively, independently in every period, $\sigma > 1/2$. After k consecutive periods of *Rain* a k-adaptive rule will assign the weight at least $1/2$ on *Rain*. The event that such a sequence occurs has a probability exponentially decreasing in k. Yet, this probability is strictly positive, thus preventing the decision maker from forecasting *Sun*, which is the best benchmark in expectation.

Next, we argue what can go wrong if too much weight is given on distant past. The correspondent bound is $\varepsilon \geq \frac{1-\delta^{k-1}}{2}$ for a k-adaptive rule and $\varepsilon \geq \frac{1}{2}$ for an unadaptive rule. We explain the intuition by illustrating what can happen with an unadaptive rule that equally weighs all past information. If some benchmark that has been the best for a long time becomes inferior, then it may take a very long time for the decision maker to adjust the weights towards different benchmarks. The longer the history, the longer it will take to adapt to changes. No matter how patient the decision maker is, she risks to get stuck with a wrong benchmark for an arbitrarily long period of time. Thus, the problem of dynamic consistency arises. After some time and some histories the decision maker will prefer to "forget" the past and to restart her decision rule from the empty history.

For illustration, let us consider the payoffs of the previous example. We now consider sequences of states and evaluate realized payoffs. Assume that *Sun* occurs in the first T periods and *Rain* occurs ever after. Then, in every period $t = T+1, \ldots, 2T$, the decision maker will assign a weight at most $1/2$ on *Rain*, even though *Rain* occurs in each of these periods. The payoffs in periods $T+1$ to $2T$ are thus at most $1/2$, far from the best. So, for any given discount factor $\delta < 1$ and a sufficiently large T, the decision rule's performance evaluated at period $T+1$ is substantially worse than that of the best benchmark (in this example, the constant benchmark that forecasts *Rain*).

5 Extensions

Within our methodology we can allow for certain extensions of our model.

Non-convex and finite action sets. We show why our results extend to a more general setting where the set of actions, A, need not be convex and payoff function

$u(a, \omega)$ need not be concave in a. The model where there are only finitely many different actions is a special case. The challenge that we face here is that for a given vector of benchmarks' actions, $(r_t(1), ..., r_t(n))$, the decision rule $p_{(m,\eta)}$ stipulates to choose an action a_t equal to some linear combination of $(r_t(1), ..., r_t(n))$. But a_t may not belong to A, since the latter need not be convex. As in Hannan (1957) or Hart and Mas-Colell (2001), we deal with this problem by letting the decision maker play a mixed strategy, a lottery over benchmark recommendations, which themselves are elements of A by definition. Accordingly, the decision maker follows the recommendation of benchmark i with probability equal to the weight $\lambda_t(i)$ assigned on this benchmark in each period t. All our results then hold *in expectation* w.r.t. the decision maker's own mixed strategy.

The multi-armed bandit setting. Consider learning under partial information where the decision maker observes only payoffs from *chosen* actions. Payoffs of the benchmarks whose actions have not been adopted are not observed. Here we explain how to extend our algorithm to derive the result analogous to Theorem 1.

Since the foregone payoffs are not observed, we use the trick of Auer et al. (1995) to construct their unbiased estimates. Define the estimate $\hat{u}_t(i)$ of a payoff of each benchmark i in every period t as $u(a_t, \omega_t)/r_t(i)$ if benchmark i's action is chosen by the decision maker in period t, and $\hat{u}_t(i) = 0$ otherwise. Then, in each period with probability $1 - \nu$ use our decision rule $p_{(m,\eta)}$ w.r.t. the estimated past performances of the benchmarks, and with probability ν follow the action of a random benchmark, choosing each benchmark equally likely. These adjustments can be easily accounted for in our proofs to yield a result as in Theorem 1, the existence of a simple decision rule for dynamic benchmark targeting. Note that each benchmark is followed with probability greater or equal to ν, hence all estimates are bounded from above by $1/\nu$. The parameter $\nu > 0$ is called the rate of experimentation, its value can be fine-tuned for the best performance. Naturally, the new error bound will be greater, as now the decision maker conditions her decisions on much less information.

Decision makers with bounded horizon. Suppose that a decision maker does not discount future payoffs, but instead is concerned in each period t with average payoffs over $t+1, \ldots, t+T$ for a fixed horizon T. Here the same simple rule can be used. Some work is needed to derive a new error bound and then to choose the free parameters m and η that minimize this bound.

We hasten to point out that if a decision maker faces a finitely repeated decision problem in periods $t = 1, 2, \ldots, \overline{T}$, then dynamic benchmark targeting strategies with error bound $\varepsilon < 1/2$ fail to exist, regardless of how past information is used. The intuition is simple. After facing $\overline{T} - 1$ periods, the decision maker is only concerned with her payoff in the final period $\overline{T}$. However, the state of the environment in the last period need not depend on the past realizations. Thus, the decision maker can guarantee only the maxmin payoff, in our *Rain & Sun* example in Section 4 this is 1/2, while the payoff of the best benchmark in the final round is equal to 1.

6 Conclusion

In this paper we introduce a methodology for dynamic decision making in which at each point in time the decision maker compares own performance to a given set of benchmark algorithms, rules of thumb, or advices of experts. The novelty of this paper is in the addition of a new pair of elements into decision making with expert advice: discounting of future payoffs and dynamic consistency. We present a decision rule that guarantees to perform, in terms of discounted present values, nearly as well as or better than each of these benchmarks at any point in time. Using our rule, the decision maker need not model the environment, as she would under the Bayesian paradigm, and hence does not use complicated optimization routines and need not be worried about misspecifying the environment. Choices are time consistent, hence if the best benchmark changes, then the decision maker will track this change.

Within our introduced methodology the notion of optimality is well defined, as we search for a decision rule with the smallest error bound. The bound presented for our rule (Theorem 1) and for the two alternative rules (Propositions 1 and 2) are not known to be tight. A topic for future research is to improve these bounds, to find new rules with better bounds or to design rules and establish bounds for more restricted environments. The natural first step in this direction is to establish a lower bound on the error bound of any rule.

Notice that the error bound of our rule has been derived for the worst case and depends only on the *number* of benchmarks, but not their properties. For a specific choice of benchmarks and for a specific environment the error bound can be much lower. How much lower it will be depends on the additional assumptions. This question is left for

future research.

A separate question that this paper does not address is the choice of benchmarks. An additional benchmark can substantially improve performance if it turns out to perform much better than the others in the given environment. At the same time, adding a benchmark potentially increases the error bound, as it is more difficult to outperform more benchmarks. So the decision maker has the tradeoff between a potentially higher absolute performance and a potentially larger gap in performance to the best benchmark. This is another avenue for future research.

Appendix A. Proofs

A.1 Proof of Theorem 1

Proof. Consider the rule $p_{(m,\eta)}$ with given parameters $m \in \mathbb{N}$ and $\eta > 0$. Recall that $C_{t,0}(i) = 0$, $C_{t,k}(i) = \sum_{s=t-k}^{t-1} u(a_s(i), \omega_s)$ for $k \geq 1$, and $\lambda_{t,k}(i) = \frac{e^{\eta C_{t,k}(i)}}{\sum_{j=1}^{n} e^{\eta C_{t,k}(j)}}$. The values for $t - k \leq 0$ are well defined by the convention that all payoffs in nonpositive rounds are zero. The actions of the rule $p_{(m,\eta)}$ for all $t \in \mathbb{N}$ are

$$a_t = p_{(m,\eta)}(h_{t-1}, r_t) = \sum_{i=1}^{n} \left(\frac{1}{m} \sum_{k=0}^{m-1} \lambda_{t,k}(i) \right) r_t(i).$$

Fix a benchmark $i \in \{1, ..., n\}$, a sequence of states $\bar{\omega}$, and a round t_0. We now bound the loss from not following that benchmark, $U_{t_0}(r_t(i), \bar{\omega}) - U_{t_0}(a_t, \bar{\omega})$.

For every $k = 0, 1, ..., m-1$ define the rule that combines the benchmarks based on their performance in the recent k periods,

$$b_{t,k} = \sum_{j=1}^{n} \lambda_{t,k}(j) r_t(j).$$

Note that for $k = 0$ the past is ignored and the weights are assigned uniformly to all benchmarks, $\lambda_{t,0}(j) = \frac{1}{n}$, $j = 1, ..., n$.

By concavity of $u(a,\omega)$ in a and Jensen's inequality we have

$$u(a_t,\omega_t) \geq \frac{1}{m}\sum_{k=0}^{m-1} u(b_{t,k},\omega_t). \tag{9}$$

For each $k = 0,1,...,m-1$ denote by $D_{t,t+k}(i)$ the loss from not following the action of benchmark i in round $t+k$ when using the rule $b_{t+k,k}$ based on the recent observations over rounds in $\{t,t+1,...,t+k-1\}$,

$$D_{t,t+k}(i) = u(r_{t+k}(i),\omega_{t+k}) - u(b_{t+k,k},\omega_{t+k}).$$

We now derive a bound on the sum $\sum_{k=s}^{m-1} D_{t,t+k}(i)$ using the technique of Cesa-Bianchi and Lugosi (2006, Theorem 2.2) based on Hoeffding inequality (Hoeffding, 1963).

Lemma 1.

$$\sum_{k=s}^{m-1} D_{t,t+k}(i) \leq T(s,m) := \min\left\{m-s, \frac{\ln n}{\eta} + s + \frac{\eta}{8}(m-s)\right\}.$$

Proof. Since $D_{t,t+k}(i) \leq 1$, we obtain $\sum_{k=s}^{m-1} D_{t,t+k}(i) \leq m-s$. For the second bound we generalize Theorem 2.2 of Cesa-Bianchi and Lugosi (2006). For $i \in \{1,...,n\}$ let

$$w_s(i) = e^{-\eta\sum_{k=0}^{s-1} D_{t,t+k}(i)}$$

and let $W_s = \sum_{i=1}^{n} w_s(i)$. Note that $e^{-\eta s} \leq w_s(i) \leq 1$ for all s and all i, so $W_s \leq n$. Thus, for every $i \in \{1,...,n\}$ we have

$$\begin{aligned}\ln\frac{W_m}{W_s} &= \ln\left(\sum_{j=1}^{n} w_s(j)e^{-\eta\sum_{k=s}^{m-1} D_{t,t+k}(j)}\right) - \ln W_s \geq \ln\left(w_s(i)e^{-\eta\sum_{k=s}^{m-1} D_{t,t+k}(i)}\right) - \ln n \\ &= \ln w_s(i) - \eta\sum_{k=s}^{m-1} D_{t,t+k}(i) - \ln n \geq -\eta s - \eta\sum_{k=s}^{m-1} D_{t,t+k}(i) - \ln n.\end{aligned}$$

Using the following inequality (Cesa-Bianchi and Lugosi, 2006, p. 17)

$$\ln\frac{W_m}{W_s} \leq \frac{\eta^2}{8}(m-s),$$

we obtain

$$\sum_{k=s}^{m-1} D_{t,t+k}(i) \le \frac{\ln n}{\eta} + s + \frac{\eta}{8}(m-s).$$

□

Next, by (9) we have

$$\begin{aligned} U_{t_0}(r_t(i),\bar{\omega}) - U_{t_0}(a_t,\bar{\omega}) &= (1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}(u(r_t(i),\omega_t) - u(a_t,\omega_t) \\ &\le \Delta := (1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\frac{1}{m}\sum_{k=0}^{m-1} D_{t-k,t}(i). \end{aligned}$$

We can rewrite Δ as follows,

$$\Delta = \frac{1-\delta}{m}\sum_{t=t_0-m+1}^{t_0-1}\delta^{t-t_0}\sum_{s=t_0-t}^{m-1}\delta^s D_{t,t+s}(i) + \frac{1-\delta}{m}\sum_{t=t_0}^{\infty}\delta^{t-t_0}\sum_{s=0}^{m-1}\delta^s D_{t,t+s}(i). \tag{10}$$

Let us bound the second term in the right-hand side of (10). By Lemma 1,

$$\sum_{l=0}^{m-1} D_{t,t+l}(i) \le \frac{\ln n}{\eta} + \frac{\eta m}{8}.$$

Thus we have

$$\begin{aligned} \sum_{s=0}^{m-1}\delta^s D_{t,t+s}(i) &= (1-\delta)\sum_{s=0}^{m-2}\delta^s\sum_{l=0}^{s-1} D_{t,t+l}(i) + \delta^{m-1}\sum_{l=0}^{m-1} D_{t,t+l}(i) \\ &\le (1-\delta)\sum_{s=0}^{m-2}\delta^s\left(\frac{\ln n}{\eta} + \frac{\eta s}{8}\right) + \delta^{m-1}\left(\frac{\ln n}{\eta} + \frac{\eta m}{8}\right) \\ &= \frac{\ln n}{\eta} + \frac{\eta}{8(1-\delta)}(1-\delta^m). \end{aligned} \tag{11}$$

Next, let us deal with the first term in the right-hand side of (10). By Lemma 1,

$$\sum_{l=k}^{m-1} D_{t,t+l}(i) \le T(k,m).$$

For $t \in \{t_0 - m + 1, ..., t_0 - 1\}$ set $k = t_0 - t - 1$. Observe that $0 \leq k \leq m - 2$. We have

$$\begin{aligned}\delta^{t-t_0} \sum_{s=t_0-t}^{m-1} \delta^s D_{t,t+s}(i) &= (1-\delta) \sum_{s=k}^{m-2} \delta^{s-k} \sum_{l=k}^{s-1} D_{t,t+l}(i) + \delta^{m-1-k} \sum_{l=k}^{m-1} D_{t,t+l}(i) \\ &\leq (1-\delta) \sum_{s=k}^{m-2} \delta^{s-k} T(k,s) + \delta^{m-1-k} T(k,m). \end{aligned} \tag{12}$$

By (11) and (12) we obtain

$$\begin{aligned}\Delta \leq & \Phi_{(m,\eta)} := \frac{1-\delta}{m} \sum_{k=0}^{m-2} \left((1-\delta) \sum_{s=k}^{m-2} \delta^{s-k} T(k,s) + \delta^{m-1-k} T(k,m) \right) \\ & + \frac{1}{m} \left(\frac{\ln n}{\eta} + \frac{\eta}{8(1-\delta)} (1-\delta^m) \right). \end{aligned} \tag{13}$$

Since $U_{t_0}(r_t(i), \bar{\omega}) - U_{t_0}(a_t, \bar{\omega}) \leq \Phi_{(m,\eta)}$ for all benchmarks i, all rounds t_0, and all sequences of states $\bar{\omega}$, the term $\Phi_{(m,\eta)}$ is an error bound for rule $p_{(m,\eta)}$.

Next, we make the error bound $\Phi_{(m,\eta)}$ small by choosing the free parameters m and η. The values that approximately minimize $\Phi_{(m,\eta)}$ are

$$\eta^* = 2^{\frac{4}{3}} (\ln n)^{\frac{1}{3}} (1-\delta)^{\frac{1}{3}} \quad \text{and} \quad m^* = \frac{\eta^*}{2(1-\delta)} + x. \tag{14}$$

where $x \in (-1, 1]$ is the adjustment such that m is an even integer. For the proof we do not need to show how these optimal parameters are derived, we only need to prove that $\Phi_{(m^*,\eta^*)}$ has the stated error bound,

$$\Phi_{(m^*,\eta^*)} \leq \frac{3}{4} \big(2(1-\delta) \ln n\big)^{\frac{1}{3}} + \frac{7}{96} \big(2(1-\delta) \ln n\big)^{\frac{2}{3}}. \tag{15}$$

In order to deal with the inconvenient, nondifferentiable term $T(k,s)$ in (13), we use

$$T(k,s) = \min \left\{ s-k, \frac{\ln n}{\eta^*} + \frac{\eta^* s}{8} \right\} \leq \tilde{T}(k,s) := \begin{cases} s-k, & k \leq \frac{m^*}{2}, \\ \frac{\ln n}{\eta} + \frac{\eta s}{8}, & k > \frac{m^*}{2}. \end{cases}$$

The summations then split into two differentiable parts,

$$\sum_{s=k}^{m^*-2} \delta^{s-k} T(k,s) \leq \sum_{s=k}^{m^*-2} \delta^{s-k} \tilde{T}(k,s) = \sum_{s=k}^{m^*/2} \delta^{s-k}(s-k) + \sum_{s=m^*/2}^{m^*-2} \delta^{s-k} \left(\frac{\ln n}{\eta^*} + \frac{\eta^* s}{8} \right).$$

Replacing T by $\tilde{T}$ in the right-hand side of (13) yields a differentiable expression. Using the Taylor expansion of this expression w.r.t. $(1-\delta)$ up to the third term yields (15), where the third term of the expansion is nonpositive and bounded by zero. ■

A.2 Proof of Theorem 2

The theorem is proved by example. Consider two states 0 and 1, set of actions $A = [0,1]$, and payoffs given by $u(a,\omega) = 1 - |a-\omega|$, $a \in A = [0,1]$, $\omega \in \Omega = \{0,1\}$. There are two benchmarks, labeled 0 and 1, that recommend the respective extreme constant actions, $r_t(0) = 0$ and $r_t(1) = 1$ for all t.

To prove that the error bound of a k-adaptive decision rule satisfies $\varepsilon \geq \frac{1}{2^{(k+1)\log(k+1)}}$, we consider an i.i.d. environment and compare the expected performance of the benchmark and a given decision rule. Note that a lower bound on the difference in the expected performance is also a lower bound on the realized performance, for some sequence of realized events.

the following environment. The state equals 0 and 1 with probability $1-\sigma$ and σ, respectively, independently in all periods, $\sigma \in (\frac{1}{2}, 1)$. In this setting, benchmark 1 is the better of the two as it is correct with probability $\sigma > \frac{1}{2}$ in every period and yields the expected payoff $\mathbb{E}[u(1,\omega)] = \sigma$.

For each period $t > k$ let E_t be the event that $\omega_{t-s} = 0$ for every $s = 1, \ldots, k$. Since we have assumed $u(a,\omega) = 1 - |a-\omega|$, under event E_t we have $u(0,\omega_{t-s}) = 1$ and $u(1,\omega_{t-s}) = 0$ for each $s = 1, \ldots, k$, and hence $\mu_t(0) \geq \mu_t(1)$ by k-adaptivity. The expected payoff of the decision maker conditional on E_t is

$$\begin{aligned} \mathbb{E}[u(p,\omega_t)|E_t] &= \sigma(\mu_t(1)\cdot 1 + \mu_t(0)\cdot 0) + (1-\sigma)(\mu_t(1)\cdot 0 + \mu_t(0)\cdot 1) \\ &= \sigma - \mu_t(0)(2\sigma - 1) \leq \sigma - \tfrac{1}{2}(2\sigma - 1) = \tfrac{1}{2}, \end{aligned}$$

where we used $\mu_t(0) + \mu_t(1) = 1$ and $\mu_t(0) \geq \mu_t(1)$. Since $\Pr[E_t] = (1-\sigma)^k$ and the

upper bound on the expected stage payoff is σ, it follows that

$$\begin{aligned}\mathbb{E}[u(p,\omega_t)] &= \mathbb{E}[u(p,\omega_t)|E_t]\Pr[E_t] + \mathbb{E}[u(p,\omega_t)|\text{not } E_t](1-\Pr[E_t]) \\ &\leq \tfrac{1}{2}\Pr[E_t] + \sigma(1-\Pr[E_t]) = \sigma - (\sigma - \tfrac{1}{2})\Pr[E_t] = \sigma - \tfrac{2\sigma-1}{2}(1-\sigma)^k.\end{aligned}$$

As the expected payoff of benchmark 1 is σ, the difference is

$$\mathbb{E}[u(1,\omega_t) - u(p,\omega_t)] \geq \tfrac{2\sigma-1}{2}(1-\sigma)^k.$$

Since the choice of σ is arbitrary, maximizing the right-hand side w.r.t. $\sigma \in [\frac{1}{2}, 1]$ yields

$$\max_{\sigma\in[\frac{1}{2},1]} \tfrac{2\sigma-1}{2}(1-\sigma)^k = \left(\frac{k}{k+1}\right)^k \frac{1}{2^k(k+1)} \geq \frac{1}{2^{(k+1)\log(k+1)}}.$$

Since the state is i.i.d., the expected discounted sum of future payoffs for the decision maker in every period $t > k$ is also less than benchmark 1's payoff by at least $\frac{1}{2^{(k+1)\log(k+1)}}$, independently of the discount factor. It is immediate that the same statement is true for some realized path of the events. Consequently, the error bound satisfies

$$\varepsilon \geq \frac{1}{2^{(k+1)\log(k+1)}}.$$

Next, to prove that the error bound of a k-adaptive decision rule satisfies $\varepsilon \geq \frac{1-\delta^{k-1}}{2}$, consider the following environment. Let T be an integer and consider the sequence of states $\bar{\omega}$ where $\omega_t = 1$ for all $t \geq T$.

Then, for every period $t = T, T+1, ..., T+k-2$, in the recent $t - T < k$ periods benchmark 1 has payoff one and benchmark 0 has payoff zero. Hence, by k-adaptivity, there exists a large enough T and a history of states preceding T such that $\mu_t(1) < \mu_t(0)$. Moreover, by monotonicity, this history is such that $\omega_s = 0$ for all $s < T$, so that benchmark 1 is worst and benchmark 0 is best in all periods before T. Under this history,

$$\mu_t(1) < \mu_t(0) \quad \text{for all } t = T, T+1, ..., T+k-2. \tag{16}$$

Thus,

$$U_T(p,\bar{\omega}) < (1-\delta)\sum_{s=0}^{k-2}\delta^s(0\cdot\tfrac{1}{2} + 1\cdot\tfrac{1}{2}) + \delta^{k-1}U_{T+k-1}(p,\bar{\omega}) \leq \frac{1}{2}(1-\delta^{k-1}) + \delta^{k-1}.$$

The discounted sum of payoffs of benchmark 1 in period T is $U_T(p_1, \bar{\omega}) = 1$, since in all periods from T on benchmark 1's payoff is constantly one. Hence the error bound must satisfy

$$\varepsilon \geq U_T(p_1, \bar{\omega}) - U_T(p, \bar{\omega}) \geq 1 - \frac{1}{2}(1 - \delta^{k-1}) - \delta^{k-1} = \frac{1}{2}(1 - \delta^{k-1}).$$

Finally, we prove that the error bound of an unadaptive decision rule satisfies $\varepsilon \geq \frac{1}{2}$. Within the same environment considered above, if a decision rule is unadaptive, then for every k there exists $T = T(k)$ such that (16) holds, and hence

$$\varepsilon \geq U_{T(k)}(p_1, \bar{\omega}) - U_{T(k)}(p, \bar{\omega}) \geq \frac{1}{2}(1 - \delta^{k-1}).$$

Since the error bound must satisfy the above for all periods, we have

$$\varepsilon \geq \sup_{k \in \mathbb{N}} \left\{ \frac{1}{2}(1 - \delta^{k-1}) \right\} = \frac{1}{2}.$$

■

Appendix B. Proofs (Online Appendix)

B.1 Proof of Proposition 1

Consider a rule $\bar{q}_{(m,\eta)}$ for some parameters $m \in \mathbb{N}$ and $\eta > 0$, and fix a sequence of states $\bar{\omega}$.

Define $Z_t(i) = u(r_t(i), \omega_t) - u(a_t, \omega_t)$ for every $i = 1, \ldots, n$ and every t. We shall also simplify notations for the sum of the future discounted payoffs, writing $U_{t_0}(0)$ for $U_{t_0}(\bar{q}_{(m,\eta)}, \bar{\omega})$ and $U_{t_0}(i)$ for $U_{t_0}(p_i, \bar{\omega})$.

Fix a benchmark i and consider the starting period t_0 just after the restart, so $t_0 = mk_0$

for some integer k_0. We have

$$J(i) = U_{t_0}(i) - U_{t_0}(0) = (1-\delta)\sum_{t=mk_0}^{\infty}\delta^{(t-mk_0)}Z_t(i) = (1-\delta)\sum_{k=k_0}^{\infty}\delta^{m(k-k_0)}\sum_{s=0}^{m-1}\delta^s Z_{mk+s}(i)$$
$$= (1-\delta)\sum_{k=k_0}^{\infty}\delta^{m(k-k_0)}\left(\sum_{s=0}^{m-1}(\delta^s-\delta^{m-1})Z_{mk+s}(i) + \sum_{s=0}^{m-1}\delta^{m-1}Z_{mk+s}(i)\right).$$

Now, since $|Z_t(i)| \le 1$,

$$\sum_{s=0}^{m-1}(\delta^s-\delta^{m-1})Z_{mk+s}(i) \le \sum_{s=0}^{m-1}(\delta^s-\delta^{m-1}) = \frac{1-\delta^m}{1-\delta} - m\delta^{m-1}.$$

Also, by Theorem 2.2 in Cesa-Bianchi and Lugosi (2006),

$$\sum_{s=0}^{m-1}Z_{mk+s}(i) \le \frac{\ln n}{\eta} + \frac{m\eta}{8} \le \sqrt{\frac{m\ln n}{2}},$$

where we choose $\eta = \sqrt{(8\ln n)/m}$. Hence,

$$J(i) \le (1-\delta)\sum_{k=k_0}^{\infty}\delta^{m(k-k_0)}\left(\frac{1-\delta^m}{1-\delta} - m\delta^{m-1} + \delta^{m-1}\sqrt{\frac{m\ln n}{2}}\right)$$
$$= \frac{1-\delta}{1-\delta^m}\left(\frac{1-\delta^m}{1-\delta} - m\delta^{m-1} + \delta^{m-1}\sqrt{\frac{m\ln n}{2}}\right)$$
$$= 1 - \frac{1-\delta}{1-\delta^m}\delta^{m-1}\left(m - \sqrt{\frac{m\ln n}{2}}\right).$$

Next, consider any t_0 and denote by $z \in \{0,1,...,m-1\}$ the number of periods that remain until the next restart, so the integer t_0+z is a multiple of m. Using $|Z_t(i)| \le 1$ and that the sum from the period of restart on is $J(i)$, we have

$$U_{t_0}(i) - U_{t_0}(0) = (1-\delta)\sum_{t=t_0}^{t_0+z-1}\delta^{t-t_0}Z_t(i) + (1-\delta)\delta^z\sum_{t=t_0+z}^{\infty}\delta^{t-t_0}Z_t(i)$$
$$= (1-\delta)\sum_{t=t_0}^{t_0+z-1}\delta^{t-t_0} + \delta^z J(i) = 1 - \delta^z + \delta^z J(i).$$

Since $J(i) \leq 1$, this expression is increasing in z, so the worst case is $z = m-1$. Substituting the bound for $J(i)$, we have

$$\begin{aligned} U_{t_0}(i) - U_{t_0}(0) &\leq 1 - \delta^{m-1} + \delta^{m-1}\left(1 - \frac{1-\delta}{1-\delta^m}\delta^{m-1}\left(m - \sqrt{\frac{m\ln n}{2}}\right)\right) \\ &= 1 - \delta^{2(m-1)}\frac{1-\delta}{1-\delta^m}\left(m - \sqrt{\frac{m\ln n}{2}}\right). \end{aligned}$$

Substituting $m = m(\delta) = c/(1-\delta)^{2/3}$ with a parameter $c > 0$ into the above expression, using Taylor expansion up to the second term and upper-bounding that term by zero yields

$$U_{t_0}(i) - U_{t_0}(0) \leq \left(\frac{3}{2}c + \frac{1}{\sqrt{c}}\sqrt{\frac{\ln n}{2}}\right)(1-\delta)^{1/3}.$$

Choosing c to minimize the leading constant, $c = 2^{-1/3}3^{-2/3}(\ln n)^{1/3}$, yields

$$U_{t_0}(i) - U_{t_0}(0) \leq \left(\tfrac{3}{2}\right)^{4/3}(\ln n)^{1/3}(1-\delta)^{1/3}.$$

Since the above holds for each benchmark i and for each starting period t_0, the statement of the proposition follows immediately.

B.2 Proof of Proposition 2

Consider a rule $\tilde{q}_{(\alpha,\eta)}$ for some parameters $\alpha \in (0,1)$ and $\eta > 0$. Fix a sequence of states $\bar{\omega}$ and a round t_0.

Define $X_t(0) = u(a_t, \omega_t)$ and $X_t(i) = u(r_t(i), \omega_t)$ for every $i = 1, \ldots, n$ and every t. In these notations, the performance of every benchmark $i = 0, 1, \ldots, n$ is evaluated by

$$C_{\alpha,t}(i) = X_t(i) + \alpha C_{\alpha,t-1}(i), \quad t \geq 1,$$

with $C_{\alpha,0}(i) = 0$. We shall also simplify notations for the sum of the future discounted payoffs, writing $U_{t_0}(0)$ for $U_{t_0}(\tilde{q}_{(\alpha,\eta)}, \bar{\omega})$ and $U_{t_0}(i)$ for $U_{t_0}(p_i, \bar{\omega})$.

To begin with, let us show that

$$U_{t_0}(i) - U_{t_0}(0) \leq \alpha\frac{1-\delta}{1-\alpha} + \frac{(1-\delta\alpha)\eta}{8\alpha(1-\alpha)} + \frac{1-\delta\alpha}{\eta}\ln n. \tag{17}$$

Let $w_t(i) = e^{\eta C_{\alpha,t-1}(i)}$, so $w_1(i) = 1$ and

$$w_{t+1}(i) = w_t^{\alpha}(i) e^{\eta X_t(i)}, \quad t \geq 2 \ .$$

Also, let $W_t = \sum_{j=1}^n w_t(j)$ and $v_t(i) = \frac{w_t(i)}{W_t}$ for all $i = 1, \ldots, n$ and all $t \geq 1$. Note that decision rule $\tilde{q}_{(\alpha,\eta)}$ stipulates to play in every period t the weighted average of the benchmarks' recommended actions, with weight $v_t(i)$ assigned to the action recommended by benchmark $i = 1 \ldots, n$,

$$a_t = \sum_{i=1}^n v_t(i) r_t(i).$$

By concavity of $u(a, \omega)$ in a and Jensen's inequality,

$$X_t(0) \geq \sum_{j=1}^n v_t(j) X_t(j). \tag{18}$$

First, we find a bound on $X_t(0)$. Using Jensen's inequality again, we obtain

$$\begin{aligned}
\ln \frac{W_{t+1}}{W_t^{\alpha}} &= \ln \sum_{j=1}^n \frac{w_{t+1}(j)}{W_t^{\alpha}} = \ln \sum_{j=1}^n \frac{w_t^{\alpha}(j)}{W_t^{\alpha}} e^{\eta X_t(j)} = \ln \sum_{j=1}^n v_t^{\alpha}(j) e^{\eta X_t(j)} \\
&= \ln \left[\left(\sum_{j=1}^n v_t^{\alpha}(j) \frac{e^{\eta X_t(j)}}{\sum_{k=1}^n e^{\eta X_t(k)}} \right) \left(\sum_{k=1}^n e^{\eta X_t(k)} \right) \right] \\
&\leq \ln \left[\left(\sum_{j=1}^n v_t(j) \frac{e^{\eta X_t(j)}}{\sum_{k=1}^n e^{\eta X_t(k)}} \right)^{\alpha} \left(\sum_{k=1}^n e^{\eta X_t(k)} \right) \right] \\
&= \alpha \ln \sum_{j=1}^n v_t(j) e^{\eta X_t(j)} + (1-\alpha) \ln \sum_{j=1}^n e^{\eta X_t(j)} \\
&= \alpha \ln \sum_{j=1}^n v_t(j) e^{\eta X_t(j)} + (1-\alpha) \ln \left(\frac{1}{n} \sum_{j=1}^n e^{\eta X_t(j)} \right) + (1-\alpha) \ln n.
\end{aligned}$$

We will need the following generalization of the Hoeffding inequality.

Lemma 2 (Cesa-Bianchi and Lugosi 2006, Lemma 2.2)**.** *Let Z be a random variable with $a \leq Z \leq b$. Then for every $s \in \mathbb{R}$,*

$$\ln \mathbb{E}\left[e^{sZ}\right] \leq s\mathbb{E}Z + \frac{s^2(b-a)^2}{8}.$$

By Lemma 2, inequality (18) and the assumption that $X_t(j) \in [0,1]$,

$$\ln \sum_{j=1}^{n} v_t(j) e^{\eta X_t(j)} \leq \eta \sum_{j=1}^{n} v_t(j) X_t(j) + \frac{\eta^2}{8} \leq \eta X_t(0) + \frac{\eta^2}{8}.$$

Again, by Lemma 2,

$$\ln \left(\frac{1}{n} \sum_{j=1}^{n} e^{\eta X_t(j)} \right) \leq \frac{\eta}{n} \sum_{j=1}^{n} X_t(j) + \frac{\eta^2}{8} = \eta \theta_t + \frac{\eta^2}{8},$$

where $\theta_t = \frac{1}{n} \sum_{j=1}^{n} X_t(j)$. Consequently,

$$\ln \frac{W_{t+1}}{W_t^{\alpha}} \leq \alpha \eta X_t(0) + (1-\alpha)\eta \theta_t + \frac{\eta^2}{8} + (1-\alpha) \ln n.$$

Thus, we have derived

$$X_t(0) \geq \frac{1}{\alpha \eta} \ln \frac{W_{t+1}}{W_t^{\alpha}} - \frac{1-\alpha}{\alpha} \theta_t - \frac{\eta}{8\alpha} - \frac{1-\alpha}{\alpha \eta} \ln n. \tag{19}$$

Second, we find a bound on

$$C_{\alpha,t}(0) = \sum_{k=1}^{t} \alpha^{t-k} X_k(0).$$

Following (19),

$$\begin{aligned}
C_{\alpha,t}(0) &\geq \frac{1}{\alpha \eta} \left(\ln \frac{W_{t+1}}{W_t^{\alpha}} + \alpha \ln \frac{W_t}{W_{t-1}^{\alpha}} + \ldots + \alpha^{t-1} \ln \frac{W_2}{W_1^{\alpha}} \right) \\
&\quad - \frac{1-\alpha}{\alpha} \sum_{k=1}^{t} \alpha^{t-k} \theta_t - \left(\frac{\eta}{8\alpha} + \frac{1-\alpha}{\alpha \eta} \ln n \right) \sum_{k=1}^{t} \alpha^{t-k} \\
&= \frac{1}{\alpha \eta} \ln \frac{W_{t+1}}{W_t^{\alpha}} \frac{W_t^{\alpha}}{W_{t-1}^{\alpha^2}} \cdots \frac{W_2^{\alpha^{t-1}}}{W_1^{\alpha^t}} - \frac{1-\alpha}{\alpha} \sum_{k=1}^{t} \alpha^{t-k} \theta_t - \left(\frac{\eta}{8\alpha} + \frac{1-\alpha}{\alpha \eta} \ln n \right) \frac{1-\alpha^t}{1-\alpha} \\
&= \frac{1}{\alpha \eta} \ln \frac{W_{t+1}}{W_1^{\alpha^t}} - \frac{1-\alpha}{\alpha} \sum_{k=1}^{t} \alpha^{t-k} \theta_t - \frac{\eta}{8\alpha} \frac{1-\alpha^t}{1-\alpha} - \frac{1-\alpha^t}{\alpha \eta} \ln n \\
&= \frac{1}{\alpha \eta} \ln W_{t+1} - \frac{1-\alpha}{\alpha} \sum_{k=1}^{t} \alpha^{t-k} \theta_t - \frac{\eta}{8\alpha} \frac{1-\alpha^t}{1-\alpha} - \frac{\ln n}{\alpha \eta},
\end{aligned}$$

where we used $W_1 = \sum_{j=1}^n w_1(j) = n$, so $\ln W_1^{\alpha^t} = \alpha^t \ln n$.

Fix any $j = 1, \ldots, n$. Using $W_{t+1} = \sum_k w_{t+1}(k) \geq w_{t+1}(j) = e^{\eta C_t(j)}$, we obtain

$$\begin{aligned} C_{\alpha,t}(0) &\geq \frac{1}{\eta} \ln W_{t+1} + \frac{1-\alpha}{\alpha\eta} \ln W_{t+1} - \frac{1-\alpha}{\alpha} \sum_{k=1}^t \alpha^{t-k} \theta_t - \frac{\eta}{8\alpha} \frac{1-\alpha^t}{1-\alpha} - \frac{1}{\alpha\eta} \ln n \\ &\geq C_{\alpha,t}(j) + \frac{1-\alpha}{\alpha\eta} \left(\ln W_{t+1} - \eta \sum_{k=1}^t \alpha^{t-k} \theta_t \right) - \frac{\eta}{8\alpha} \frac{1}{1-\alpha} - \frac{1}{\alpha\eta} \ln n. \end{aligned}$$

Observe that

$$\begin{aligned} \ln W_{t+1} - \eta \sum_{k=1}^t \alpha^{t-k} \ln \theta_t &= \ln \sum_{j=1}^n e^{\eta \sum_{k=1}^t \alpha^{t-k} X_k(j)} - \sum_{k=1}^t \alpha^{t-k} \frac{1}{n} \sum_{j=1}^n X_k(j) \\ &= \ln n + \ln \left(\frac{1}{n} \sum_{j=1}^n e^{y(j)} \right) - \frac{1}{n} \sum_{j=1}^n y(j), \end{aligned}$$

where $y(j) = \eta \sum_{k=1}^t \alpha^{t-k} X_k(j)$. By Jensen's inequality,

$$\frac{1}{n} \sum_{j=1}^n e^{y(j)} \geq e^{\frac{1}{n} \sum_{j=1}^n y(j)},$$

and hence

$$\ln W_{t+1} - \eta \sum_{k=1}^t \alpha^{t-k} \ln \theta_t \geq \ln n + \ln \left(e^{\frac{1}{n} \sum_{j=1}^n y(j)} \right) - \frac{1}{n} \sum_{j=1}^n y(j) = \ln n.$$

Consequently,

$$\begin{aligned} C_{\alpha,t}(0) &\geq C_{\alpha,t}(j) + \frac{1-\alpha}{\alpha\eta} \ln n - \frac{\eta}{8\alpha(1-\alpha)} - \frac{1}{\alpha\eta} \ln n \\ &= C_{\alpha,t}(j) - \frac{\eta}{8\alpha(1-\alpha)} - \frac{\ln n}{\eta}. \end{aligned} \tag{20}$$

Finally, we bound $U_{t_0}(0)$. We evaluate for $j \in \{0, 1, .., n\}$,

$$
\begin{aligned}
(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}C_{\alpha,t}(j) &= (1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\sum_{k=1}^{t}\alpha^{t-k}X_k(j) \\
&= (1-\delta)\left(\begin{array}{c}(\alpha^{t_0-1}+\delta\alpha^{t_0}+\delta^2\alpha^{t_0+1}+...)\,X_1(j)+(\alpha^{t_0-2}+\delta\alpha^{t_0-1}+...)\,X_2(j)\\ +..+(1+\delta\alpha+...)\,X_{t_0}(j)+(\delta+\delta^2\alpha+...)\,X_{t_0+1}(j)...\end{array}\right) \\
&= (1-\delta)\left(\begin{array}{c}\alpha^{t_0-1}(1+\delta\alpha+\delta^2\alpha^2+...)\,X_1(j)+\alpha^{t_0-2}(1+\delta\alpha+...)\,X_2(j)\\ +..+(1+\delta\alpha+...)\,X_{t_0}(j)+\delta(1+\delta\alpha+...)\,X_{t_0+1}(j)+...\end{array}\right) \\
&= (1-\delta)\frac{1}{1-\delta\alpha}\sum_{t=t_0}^{\infty}\delta^{t-t_0}X_t(j)+(1-\delta)\frac{1}{1-\delta\alpha}\sum_{k=1}^{t_0-1}\alpha^{t_0-k}X_k(j) \\
&= \frac{1}{1-\delta\alpha}U_{\delta,t_0}(j)+\frac{1-\delta}{1-\delta\alpha}\sum_{k=1}^{t_0-1}\alpha^{t_0-k}X_k(j).
\end{aligned}
$$

Using (20) we obtain

$$
\begin{aligned}
\frac{U_{t_0}(j)-U_{t_0}(0)}{1-\delta} &= (1-\delta\alpha)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\left[C_{\alpha,t}(j)-C_{\alpha,t}(0)\right]-\sum_{k=1}^{t_0-1}\alpha^{t_0-k}\left[X_k(j)-X_k(0)\right] \\
&\leq \frac{1-\delta\alpha}{1-\delta}\left(\frac{\eta}{8\alpha(1-\alpha)}+\frac{\ln n}{\eta}\right)+\sum_{k=1}^{t_0-1}\alpha^{t_0-k} \\
&= \frac{1-\delta\alpha}{1-\delta}\left(\frac{\eta}{8\alpha(1-\alpha)}+\frac{\ln n}{\eta}\right)+\alpha\frac{1-\delta}{1-\alpha},
\end{aligned}
$$

which completes the proof of (17).

Next, choose α and η that satisfy

$$\frac{1-\alpha}{\alpha}=\frac{2(1-\delta)^{2/3}}{(\ln n)^{1/3}}$$

and $\eta = 2\sqrt{2\alpha(1-\alpha)\ln n}$. Substituting the above η and α into the right-hand side of (17) and using Taylor expansion up to the third term yields

$$U_{t_0}(i)-U_{t_0}(0)\leq\frac{3}{2}\left((1-\delta)\ln n\right)^{\frac{1}{3}}+\frac{1}{2}\left((1-\delta)\ln n\right)^{\frac{2}{3}}.$$

The third term of the expansion is nonpositive and bounded by zero. Since the above

holds for each benchmark i and for each starting period t_0, the statement of the proposition follows immediately. ■

References

Auer, P., N. Cesa-Bianchi, Y. Freund, and R. E. Schapire (1995). Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pp. 322–331.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Journal of the Operational Research Society 20*, 451–468.

Blackwell, D. (1956). An analog of the minmax theorem for vector payoffs. *Pacific Journal of Mathematics 6*, 1–8.

Blum, A., E. Even-Dar, and K. Ligett (2006). Routing without regret: on convergence to Nash equilibria of regret-minimizing algorithms in routing games. In *Proceedings of the 25th Annual ACM Symposium on Principles of Distributed Computing*, pp. 45–52.

Camerer, C. and T. H. Ho (1999). Experience-weighted attraction learning in normal form games. *Econometrica 67*, 827–874.

Cesa-Bianchi, N. and G. Lugosi (2003). Potential-based algorithms in on-line prediction and game theory. *Machine Learning 51*, 239–261.

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

Chen, Y. and J. W. Vaughan (2010). A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, pp. 189–198. mimeo.

Clemen, R. T. and R. L. Winkler (2007). Aggregating probability distributions. In W. Edwards, R. Miles, and D. von Winterfeldt (Eds.), *Advances in Decision Analysis*, pp. 154–176. Cambridge University Press.

DeMarzo, P., I. Kremer, and Y. Mansour (2006). Online trading algorithms and robust option pricing. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pp. 477–486.

Erev, I. and A. E. Roth (1998). Prediction how people play games: Reinforcement learning in games with unique strategy equilibrium. *American Economic Review 88*, 848–881.

Foster, D. and R. Vohra (1993). A randomization rule for selecting forecasts. *Operations Research 41*, 704–709.

Foster, D. and R. Vohra (1999). Regret in the online decision problem. *Games and Economic Behavior 29*, 7–35.

Freund, Y. and R. Schapire (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behavior 29*, 79–103.

Fudenberg, D. and D. Levine (1995). Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control 19*, 1065–1089.

Fudenberg, D. and D. Levine (1999). Conditional universal consistency. *Games and Economic Behavior 29*, 104–130.

Guthrie, E. R. (1952). *The Psychology of Learning*. New York: Harper.

Hannan, J. (1957). Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe (Eds.), *Contributions to the Theory of Games, Vol. III*, Annals of Mathematics Studies 39, pp. 97–139. Princeton University Press.

Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68*, 1127–1150.

Hart, S. and A. Mas-Colell (2001). A general class of adaptive strategies. *Journal of Economic Theory 98*, 26–54.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association 58*, 13–30.

Jose, V. R. R., Y. Grushka-Cockayne, and K. C. Lichtendahl, Jr. (2014). Trimmed opinion pools and the crowd's calibration problem. *Management Science 60*, 463–475.

Larrick, R. P. and J. B. Soll (2006). Intuitions about combining opinions: misappreciation of the averaging principle. *Management Science 52*, 111–127.

Lehrer, E. (2003). A wide range no-regret theorem. *Games and Economic Behavior 42*, 101–115.

Lehrer, E. and E. Solan (2009). Approachability with bounded memory. *Games and Economic Behavior 66*, 995–1004.

Littlestone, N. and M. Warmuth (1994). The weighted majority algorithm. *Information and Computation 108*, 212–261.

Mailath, G. J., A. Postlewaite, and L. Samuelson (2005). Contemporaneous perfect epsilon-equilibria. *Games and Economic Behavior 53*, 126–140.

Mallet, V., G. Stoltz, and B. Mauricette (2009). Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research 114*, D05307.

Mansour, Y. (2010). Regret minimization and job scheduling. In *Proceedings of the 36th Conference on Current Trends in Theory and Practice of Computer Science*, pp. 71–76. Springer.

Olszewski, W. and M. Peski (2011). The principal-agent approach to testing experts. *American Economic Journal: Microeconomics 3*, 89–113.

Radner, R. (1980). Collusive behaviour in noncooperative epsilon-equilibria of oligopolies with long but finite lives. *Journal of Economic Theory 22*, 136–154.

Ray, D. and R. Wang (2001). On some implications of backward discounting. New York University, mimeo.

Roth, A. E. and I. Erev (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior 8*, 164–212.

Rubinstein, A. (1998). *Modeling Bounded Rationality*. The MIT Press.

Sánchez, I. (2008). Adaptive combination of forecasts with application to wind energy. *International Journal of Forecasting 24*, 679–693.

Schlag, K. H. and A. Zapechelnyuk (2012). On the impossibility of achieving no regrets in repeated games. *Journal of Economic Behavior and Organization 81*, 153–158.

Strotz, R. H. (1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies 23*, 165–180.

Takimoto, E. and M. Warmuth (2003). Path kernels and multiplicative updates. *Journal of Machine Learning Research 4*, 773–818.

Timmerman, A. (2006). Forecast combinations. In G. Elliott, C. W. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*. Elsevier.

Watson, J. B. (1930). *Behaviorism*. University of Chicago Press.

Winkler, R. L. and S. Makridakis (1983). The combination of forecasts. *Journal of the Royal Statistical Society. Series A 146*, 150–157.

Zapechelnyuk, A. (2008). Better-reply dynamics with bounded recall. *Mathematics of Operations Research 33*, 869–879.