Smooth Calibration, Leaky Forecasts, Finite Recall and Nash Dynamics

2 November 2017, 11:15-12:30, bld. 72, room 465

Sergiu Hart (HUJI)

Abstract: We propose to smooth out the calibration score, which measures how good a forecaster, is, by combining nearby forecasts. While regular calibration can be guaranteed only by randomized forecasting procedures, we show that smooth calibration can be guaranteed by deterministic procedures. As a consequence, it does not matter if the forecasts are leaked, i.e., made known in advance: smooth calibration can nevertheless be guaranteed (while regular calibration cannot). Moreover, our procedure has finite recall, is stationary, and all fore-casts lie on a finite grid. To construct it, we deal also with the related setups of online linear regression and weak calibration. Finally, we show that smooth calibration yields uncoupled finite-memory dynamics in n-person games—"smooth calibrated learning"—in which the players play approximate Nash equilibria in almost all periods.

Smooth Calibration, Leaky Forecasts, Finite Recall, and Nash Dynamics*

Dean P. Foster† Sergiu Hart‡

March 30, 2017

Abstract

We propose to smooth out the calibration score, which measures how good a forecaster is, by combining nearby forecasts. While regular calibration can be guaranteed only by randomized forecasting procedures, we show that *smooth calibration* can be guaranteed by *deterministic* procedures. As a consequence, it does not matter if the forecasts are *leaked*, i.e., made known in advance: smooth calibration can nevertheless be guaranteed (while regular calibration cannot). Moreover, our procedure has *finite recall*, is stationary, and all forecasts lie on a finite grid. To construct it, we deal also with the related setups of online linear regression and weak calibration. Finally, we show that smooth calibration yields uncoupled finite-memory *dynamics* in n-person games—"smooth calibrated learning"—in which the players play approximate *Nash equilibria* in almost all periods.

*Previous versions: July 2012, February 2015. Research of the second author was partially supported by a European Research Council (ERC) Advanced Investigator grant. The authors thank Yakov Babichenko for useful comments and suggestions.

†Amazon Inc, New York City, and University of Pennsylvania, Philadelphia. *e-mail*: `dean@foster.net` *web page*: `http://deanfoster.net/`

‡Institute of Mathematics, Department of Economics, and Center for the Study of Rationality, The Hebrew University of Jerusalem. *e-mail*: `hart@huji.ac.il` *web page*: `http://www.ma.huji.ac.il/hart`

Contents

1 Introduction

How good is a forecaster? Assume for concreteness that every day the forecaster issues a forecast of the type "the chance of rain tomorrow is 30%." A simple test one may conduct is to calculate the proportion of rainy days out of those days that the forecast was 30%, and compare it to 30%; and do the same for all other forecasts. A forecaster is said to be *calibrated* if, in the long run, the differences between the actual proportions of rainy days and the forecasts are small—no matter what the weather really was (see Dawid 1982).

What if rain is replaced by an event that is under the control of another agent? If the forecasts are made public before the agent decides on his action—we refer to this setup as *"leaky forecasts"*—then calibration *cannot* be guaranteed; for example, the agent can make the event happen if and only if the forecast is less than 50%, and so the forecasting error (that is, the "calibration score") is always at least 50%. However, if in each period the forecast and the agent's decision are made "simultaneously"—which means that neither one knows the other's decision before making his own—then calibration *can* be guaranteed; see Foster and Vohra (1998). The procedure that yields calibration no matter what the agent's decisions are requires the use of *randomizations* (e.g., with probability 1/2 the forecaster announces 30%, and with probability 1/2 he announces 60%). Indeed, as the discussion at the beginning of this paragraph suggests, one cannot have a deterministic procedure that is calibrated (see Oakes 1985).

Now the standard calibration score is very sensitive: the days when the forecast was, say, 30% are considered separately from the days when the forecast was 29.99% (formally, the calibration score is a highly discontinuous function of the data, i.e., the forecasts and the actions). This suggests that one first combine all days when the forecast was *close to* 30%, and only then compare the 30% with the appropriate average proportion of rainy days. Formally, it amounts to a so-called "smoothing" operation.

Perhaps surprisingly, once we consider smooth calibration, there is no longer a need for randomization when making the forecasts: we will show that there exist deterministic procedures that guarantee smooth calibration, no matter what the agent does. In particular, it follows that it does not matter if the forecasts are made known to the agent before his decision, and so smooth calibration can be guaranteed also when forecasts may be leaked.[1]

Moreover, the forecasting procedure that we construct and which guaran-

[1]Even if the forecast is not leaked, the agent can simulate the deterministic calibration procedure and generate the forecast by himself. Our procedure indeed takes this into account.

tees smooth calibration has finite recall (i.e., only the forecasts and actions of the last R periods are taken into account, for some fixed finite R), and is stationary (i.e., independent of "calendar time": the forecast is the same any time that the "window" of the past R periods is the same).[2] Finally, we can have all the forecasts lie on some finite fixed grid.

The construction starts with the "online linear regression" problem, introduced by Foster (1991), where one wants to generate every period a good linear estimator based only on the data up to that point. We provide a finite-recall stationary algorithm for this problem; see Section 3. We then use this algorithm, together with a fixed-point argument, to obtain "weak calibration," a concept introduced by Kakade and Foster (2004) and Foster and Kakade (2006); see Section 4. Section 5 shows that weak and smooth calibration are essentially equivalent, which yields the existence of smoothly calibrated procedures. Finally, these procedures are used to obtain dynamics ("smoothly calibrated learning") that are uncoupled, have finite memory, and are close to Nash equilibria most of the time.

1.1 Literature

The *calibration problem* has been extensively studied, starting with Dawid (1982), Oakes (1985), and Foster and Vohra (1998); see Olszewski (2015) for a comprehensive survey of the literature. Kakade and Foster (2004) and Foster and Kakade (2006) introduced the notion of *weak calibration*, which shares many properties with smooth calibration. In particular, both can be guaranteed by deterministic procedures, and both are of the "general fixed point" variety: they can find fixed points of arbitrary continuous functions (see for instance the last paragraph in Section 2.3).[3] However, while weak calibration may be at times technically more convenient to work with, smooth calibration is the more natural concept, easy to interpret and understand; it

[2]One way to obtain this is by restarting the procedure once in a while; see, e.g., Lehrer and Solan (2009).

[3]They are thus more "powerful" than the standard calibration procedures, such as those based on Blackwell's approachability, which find *linear* fixed points, such as eigenvectors and invariant probabilities.

is, after all, just a standard smoothing of regular calibration.

The *online regression problem*—see Section 3 for details—was introduced by Foster (1991); for further improvements, see J. Foster (1999), Vovk (2001), Azoury and Warmuth (2001), and the book of Cesa-Bianchi and Lugosi (2006).

2 Model and Result

In this section we present the calibration game in its standard and "leaky" versions, introduce the notion of smooth calibration, and state our main results.

2.1 The Calibration Game

Let[4] $C \subseteq \mathbb{R}^m$ be a compact convex set, and let $A \subseteq C$ (for example, C could be the set of probability distributions $\Delta(A)$ over a finite set A, which is identified with the set of unit vectors in C). The *calibration game* has two players: the "action" player—the "A-player" for short—and the "conjecture" (or "calibrating") player—the "C-player" for short. At each time period $t = 1, 2, ...$, the C-player chooses $c_t \in C$ and the A-player chooses $a_t \in A$. There is full monitoring and perfect recall: at time t both players know the realized history $h_{t-1} = (c_1, a_1, ..., c_{t-1}, a_{t-1}) \in (C \times A)^{t-1}$.

In the *standard* calibration game, c_t and a_t are chosen simultaneously (perhaps in a mixed, i.e., randomized, way). In the *leaky* calibration game, a_t is chosen after c_t has been chosen and revealed; thus, c_t is a function of h_{t-1}, whereas a_t is a function of h_{t-1} *and* c_t. Formally, a pure strategy of the C-player is $\sigma : \cup_{t\geq 1}(C \times A)^{t-1} \to C$, and a pure strategy of the A-player is $\tau : \cup_{t\geq 1}(C\times A)^{t-1} \to A$ in the standard game, and $\tau : \cup_{t\geq 1}(C\times A)^{t-1}\times C \to A$ in the leaky game. A pure strategy of the C-player will also be referred to as *deterministic.*

The *calibration score*—which the C-player wants to minimize—is defined

[4] $\mathbb{R}^m$ denotes the m-dimensional Euclidean space, with the usual ℓ_2-norm $||\cdot||$.

at time $T = 1, 2, ...$ as

$$K_T = \frac{1}{T}\sum_{t=1}^{T} ||\bar{a}_t - c_t||,$$

where

$$\bar{a}_t := \frac{\sum_{s=1}^{T} \mathbf{1}_{c_s = c_t}\, a_s}{\sum_{s=1}^{T} \mathbf{1}_{c_s = c_t}};$$

here $\mathbf{1}_{x=y}$ is the indicator that $x = y$ (i.e., $\mathbf{1}_{x=y} = 1$ when $x = y$ and $\mathbf{1}_{x=y} = 0$ otherwise). Thus K_T is a function of the whole history $h_T \in (C \times A)^T$; it is the mean distance between the forecast c and the average $\bar{a}$ of the actions a chosen in those periods where the forecast was c.

2.2 Smooth Calibration

We introduce the notion of "smooth calibration." A *smoothing function* is a function $\Lambda : C \times C \to [0, 1]$ with $\Lambda(c, c) = 1$ for every c. Its interpretation is that $\Lambda(c', c)$ gives the weight that we assign to c' when we are at c; we will use $\Lambda(c', c)$ instead of the indicator $\mathbf{1}_{c'=c}$ to "smooth" out the forecasts and the average actions. Specifically, put

$$\bar{a}_t^{\Lambda} := \frac{\sum_{s=1}^{T} \Lambda(c_s, c_t)\, a_s}{\sum_{s=1}^{T} \Lambda(c_s, c_t)} \quad \text{and} \quad c_t^{\Lambda} := \frac{\sum_{s=1}^{T} \Lambda(c_s, c_t)\, c_s}{\sum_{s=1}^{T} \Lambda(c_s, c_t)}.$$

The Λ-*smoothed calibration score* at time T is then defined as

$$K_T^{\Lambda} = \frac{1}{T}\sum_{t=1}^{T} ||\bar{a}_t^{\Lambda} - c_t^{\Lambda}||. \tag{1}$$

A standard (and useful) assumption is a Lipschitz condition: there exists $L < \infty$ such that $|\Lambda(c', c) - \Lambda(c'', c)| \leq L||c' - c''||$ for all $c, c', c'' \in C$. Thus, the functions $\Lambda(\cdot, c)$ are uniformly Lipschitz: $\mathcal{L}(\Lambda(\cdot, c)) \leq L$ for every $c \in C$, where $\mathcal{L}(f) := \sup\{\|f(x) - f(y)\| \,/\, \|x - y\| : x, y \in X,\ x \neq y\}$ denotes the *Lipschitz constant* of the function f (if f is not a Lipschitz function then $\mathcal{L}(f) = +\infty$; when $\mathcal{L}(f) \leq L$ we say that f is L-*Lipschitz*).

Two classical examples of Lipschitz smoothing functions are: (i) $\Lambda(c', c) =$

$[1 - ||c' - c||/\delta]_+$ for[5] $\delta > 0$: only points within distance δ of c are considered, and their weight is proportional to the distance from c; and (ii) $\Lambda(c', c) = \exp(-||c' - c||^2/(2\sigma^2))$: the weight is given by a Gaussian (normal) perturbation.

Remarks. *(a)* The original calibration score K_T is obtained when Λ is the indicator function: $\Lambda(c', c) = \mathbf{1}_{c'=c}$ for all $c, c' \in C$.

(b) The normalization $\Lambda(c, c) = 1$ pins down the Lipschitz constant (otherwise one could replace Λ with $\alpha\Lambda$ for small $\alpha > 0$, and so lower the Lipschitz constant without affecting the score).

(c) Smoothing both $\bar{a}_t$ and c_t and then taking the difference is the same as smoothing the difference: $\bar{a}_t^\Lambda - c_t^\Lambda = (\bar{a}_t - c_t)^\Lambda$. Moreover, smoothing a_t is the same as smoothing $\bar{a}_t$, i.e., $\bar{a}_t^\Lambda = a_t^\Lambda$.

(d) An alternative score smoothes only the average action $\bar{a}_t$, but not the forecast c_t:

$$\tilde{K}_T^\Lambda = \frac{1}{T}\sum_{t=1}^{T} ||\bar{a}_t^\Lambda - c_t||.$$

If the smoothing function puts positive weight only in small neighborhoods, i.e., there is $\delta > 0$ such that $\Lambda(c', c) > 0$ only when $||c' - c|| \leq \delta$, then the difference between K_T^Λ and $\tilde{K}_T^\Lambda$ is at most δ (because in this case $||c_t^\Lambda - c_t|| \leq \delta$ for every t). More generally, $|K_T^\Lambda - \tilde{K}_T^\Lambda| \leq \delta$ when $(1/T)\sum_{t=1}^{T} ||c_t^\Lambda - c_t|| \leq \delta$ for any collection of points $c_1, ..., c_T \in C$, which is indeed the case, for instance, for the Gaussian smoothing with small enough σ^2. The reason that we prefer to use K^Λ rather than $\tilde{K}^\Lambda$ is that K^Λ vanishes when there is perfect calibration (i.e., $\bar{a}_t = c_t$ for all t), whereas $\tilde{K}^\Lambda$ is positive; clean statements such as $K_T^\Lambda \leq \varepsilon$ become $\tilde{K}_T^\Lambda \leq \varepsilon + \delta$.

Finally, given $\varepsilon > 0$ and $L < \infty$, we will say that a strategy of the C-player—which is also called a "procedure"—is (ε, L)-*smoothly calibrated* if there is $T_0 \equiv T_0(\varepsilon, L)$ such that

$$K_T^\Lambda = \frac{1}{T}\sum_{t=1}^{T} \left\|\bar{a}_t^\Lambda - c_t^\Lambda\right\| \leq \varepsilon \tag{2}$$

[5] Where $[z]_+ = \max\{z, 0\}$. These Λ functions are sometimes called "tent functions."

holds almost surely, for every strategy of the A-player, every $T > T_0$, and every smoothing function $\Lambda : C \times C \to [0, 1]$ that is L-Lipschitz in the first coordinate. Unlike standard calibration, which can be guaranteed only with high probability, smooth calibration may be obtained by deterministic procedures—as will be shown below—in which case we may well require (2) to always hold (rather than just almost surely).

2.3 Leaky Forecasts

We willl say that a procedure (i.e., a strategy of the C-player) is (smoothly) *leaky*-calibrated if it is (smoothly) calibrated also in the leaky setup, that is, against an A-player who may choose his action a_t at time t depending on the forecast c_t made by the C-player at time t (i.e., the A-player moves after the C-player). While, as we saw in the Introduction, there are no leaky-calibrated procedures, we will show that there are *smoothly leaky-calibrated* procedures.

Deterministic procedures (i.e., pure strategies of the C-player) are clearly leaky: the A-player can use the procedure at each period t to compute c_t as a function of the history h_{t-1}, and only then determine his action a_t. Thus, in particular, there cannot be deterministic calibrated procedures (because there no leaky such procedures).

In the case of smooth calibration, the procedure that we construct is deterministic, and thus smoothly leaky-calibrated. However, there are also randomized smooothly leaky-calibrated procedures. One example is the simple calibrated procedure of Foster (1999) in the one-dimensional case (where $A = \{$"rain", "no rain"$\}$ and $C = [0, 1]$): the forecast there is "almost deterministic", and so can be shown to be smoothly leaky-calibrated. For another example, see footnote 21 in Section 4 below.

A particular instance of the leaky setup is one where the A-player uses a fixed reaction function $g : C \to A$ that is a continuous mapping of forecasts to actions; thus, $a_t = g(c_t)$ (independently of time t and history h_{t-1}). In this case, smooth leaky-calibration implies that most of the forecasts that are used must be approximate fixed points of g; indeed, in every period in which the forecast is c the action is the same, namely, $g(c)$, and so the average of the actions in all the periods where the forecast is (close to) c is (close to) $g(c)$ (use the continuity of g here); formally, see the arguments in the proof of Theorem 13 in Section 6, in particular, (37). Thus, leaky procedures find (approximate) fixed points for arbitrary continuous functions g, and so must in general be more complex than the procedures that yield calibration (such as those obtained by Blackwell's approachability); cf. the complexity class PPAD (Papadimitriou 1994) in the computer science literature (see also Hazan and Kakade 2012 for the connection to calibration).

2.4 Result

A strategy σ has *finite recall* and is *stationary* if there exists a finite integer $R \geq 1$ and a function $\tilde{\sigma} : (C \times A)^R \to C$ such that

$$\sigma(h_{T-1}) = \tilde{\sigma}(c_{T-R}, a_{T-R}, c_{T-R+1}, a_{T-R+1}, ..., c_{T-1}, a_{T-1})$$

for every $T > R$ and history $h_{T-1} = (c_t, a_t)_{1 \leq t \leq T-1}$. Thus, only the "window" consisting of the last R periods matters; the rest of the history, as well as the calendar time T, do not. Finally, a finite set $D \subseteq C$ is a *δ-grid* for C if for every $c \in C$ there is $d \equiv d(c) \in D$ such that $||d - c|| \leq \delta$.

Our result is:

Theorem 1 *For every $\varepsilon > 0$ and $L < \infty$ there is an (ε, L)-smoothly calibrated procedure. Moreover, the procedure may be taken to be:*

- *deterministic;*
- *leaky;*
- *with finite recall and stationary; and*
- *with all the forecasts lying on a finite grid.*[6]

[6]The sizes R of the recall and δ of the grid depend on ε, L, the dimension m, and the bound on the compact set C.

The proof will proceed as follows. First, we construct deterministic finite-recall algorithms for the *online linear regression* problem (cf. Foster 1991, Azoury and Warmuth 2001); see Theorem 2 in Section 3. Next, we use these algorithms to get deterministic finite-recall *weakly calibrated* procedures (cf. Foster and Kakade 2004, 2006); see Theorem 10 in Section 4. Finally, we obtain smooth calibration from weak calibration; see Section 5.

3 Online Linear Regression

Classical linear regression tries to predict a variable y from a vector x of d variables (and so $y \in \mathbb{R}$ and $x \in \mathbb{R}^d$). There are observations $(x_t, y_t)_t$, and one typically assumes that[7] $y_t = \theta' x_t + \epsilon_t$, where ϵ_t are (zero-mean normally distributed) error terms. The optimal estimator for θ is then given by the least squares method; i.e., θ minimizes $(1/T)\sum_{t=1}^{T} \psi_t(\theta)$ with

$$\psi_t(\theta) := (y_t - \theta' x_t)^2$$

for every t.

In the *online linear regression* problem (Foster 1991; see Section 1.1), the observations arrive sequentially, and at each time period t we want to determine θ_t given the information at that time, namely, $(x_1, y_1), ..., (x_{t-1}, y_{t-1})$ and x_t only. The goal is to bound the difference between the mean square errors in the online case and the offline (i.e., "in hindsight") case; namely,

$$\frac{1}{T}\sum_{t=1}^{T} \psi_t(\theta_t) - \frac{1}{T}\sum_{t=1}^{T} \psi_t(\theta).$$

Thus, an online linear-regression *algorithm* takes as input a sequence $(x_t, y_t)_{t\geq 1}$ in $\mathbb{R}^d \times \mathbb{R}$ and gives as output a sequence $(\theta_t)_{t\geq 1}$ in $\mathbb{R}^d$, such that θ_t is a function only of $x_1, y_1, ..., x_{t-1}, y_{t-1}, x_t$, for each t.

Our result is:

[7]Vectors in $\mathbb{R}^m$ are viewed as column vectors, and θ' denotes the transpose of θ (thus $\theta' x$ is the scalar product $\theta \cdot x$ of θ and x, for $\theta, x \in \mathbb{R}^m$).

Theorem 2 *Let $X, Y > 0$ and $\varepsilon > 0$. Then there exists a positive integer $R_0 \equiv R_0(\varepsilon, X, Y, d)$ such that for every $R > R_0$ there is an R-recall stationary deterministic algorithm that gives $(\theta_t)_{t\geq 1}$, such that*

$$\frac{1}{R}\sum_{t=T-R+1}^{T}[\psi_t(\theta_t)-\psi_t(\theta)] \;\leq\; \varepsilon(1+\|\theta\|^2) \quad \textit{and} \tag{3}$$

$$\frac{1}{T}\sum_{t=1}^{T}[\psi_t(\theta_t)-\psi_t(\theta)] \;\leq\; \varepsilon(1+\|\theta\|^2) \tag{4}$$

hold for every $T \geq R$, every $\theta \in \mathbb{R}^d$, and every sequence $(x_t, y_t)_{t\geq 1}$ in $\mathbb{R}^d \times \mathbb{R}$ with $\|x_t\| \leq X$ and $|y_t| \leq Y$ for all t.

When in addition all the θ_t and θ are bounded by, say,[8] M, the mean square error of our online algorithm is guaranteed not to exceed the optimal offline mean square error by more than $\varepsilon(1 + M)$.

This section is devoted to the proof of Theorem 2. We start from an algorithm of Azoury and Warmuth (2001) (the "forward algorithm"), and construct from it, in a number of steps, another explicit algorithm that satisfies the desired properties (specifically, the "windowed discounted forward algorithm"; see (12) and Proposition 9).

3.1 Forward Algorithm

The starting point is the following algorithm of Azoury and Warmuth (2001, Section 5.4). For each $a > 0$, the *a-forward algorithm* gives[9] $\theta_t = Z_t^{-1}v_t$, where

$$Z_t = aI + \sum_{q=1}^{t} x_q x_q' \quad \text{and} \quad v_t = \sum_{q=1}^{t-1} y_q x_q. \tag{5}$$

Theorem 3 (Azoury and Warmuth 2001) *For every $a > 0$, the a-forward*

[8] For example, require all θ to lie in the unit simplex of $\mathbb{R}^d$.

[9] Z_t^{-1} is the inverse of the $d \times d$ matrix Z_t, and I denotes the identity matrix.

algorithm yields

$$\sum_{t=1}^{T} \psi_t(\theta_t) - \min_{\theta \in \mathbb{R}^d} \left(a \|\theta\|^2 + \sum_{t=1}^{T} \psi_t(\theta) \right) \leq \sum_{t=1}^{T} y_t^2 \left(1 - \frac{\det(Z_{t-1})}{\det(Z_t)} \right) \tag{6}$$

for every $T \geq 1$ *and every sequence* $(x_t, y_t)_{t \geq 1}$ *in* $\mathbb{R}^d \times \mathbb{R}$.

Proof. Theorem 5.6 and Lemma A.1 in Azoury and Warmuth (2001), where Z_t denotes their η_t^{-1} matrix; the second term in their formula (5.17) is non-negative since η_t is a positive definite matrix.[10] □

3.2 Discounted Forward Algorithm

Let $a > 0$ and $0 < \lambda < 1$. The λ*-discounted* a*-forward algorithm* gives $\theta_t = Z_t^{-1} v_t$, where

$$Z_t = aI + \sum_{q=1}^{t} \lambda^{t-q} x_q x_q' \quad \text{and} \quad v_t = \sum_{q=1}^{t-1} \lambda^{t-q} y_q x_q. \tag{7}$$

Proposition 4 *For every* $a > 0$ *and* $0 < \lambda < 1$, *the* λ*-discounted* a*-forward algorithm yields*

$$\sum_{t=1}^{T} \lambda^{T-t} \left[\psi_t(\theta_t) - \psi_t(\theta) \right] \leq a \|\theta\|^2 + \sum_{t=1}^{T} \lambda^{T-t} y_t^2 \left(1 - \lambda^d \frac{\det(Z_{t-1})}{\det(Z_t)} \right) \tag{8}$$

for every $T \geq 1$, *every* $\theta \in \mathbb{R}^d$, *and every sequence* $(x_t, y_t)_{t \geq 1}$ *in* $\mathbb{R}^d \times \mathbb{R}$.

Proof. Let $b := \sqrt{a(1-\lambda)}$. From the sequence $(x_t, y_t)_{t \geq 1}$ we construct a sequence $(\tilde{x}_s, \tilde{y}_s)_{s \geq 1}$ in blocks as follows. For every $t \geq 1$, the t-th block B_t is of size $d+1$ and consists of $(\lambda^{-t/2} b e^{(1)}, 0), ..., (\lambda^{-t/2} b e^{(d)}, 0), (\lambda^{-t/2} x_t, \lambda^{-t/2} y_t)$, where $e^{(i)}$ is the i-th unit vector in $\mathbb{R}^d$. The a-forward algorithm applied to $(\tilde{x}_s, \tilde{y}_s)_{s \geq 1}$ yields the following.

[10] Our statement is different from theirs because ψ_t equals twice L_t, and there is a misprinted sign in the first line of their formula (5.17).

For $s=(d+1)t$, i.e., at the end of the B_t block, we have[11] $\sum_{s\in B_t}\tilde{x}_s\tilde{x}_s' = b^2\lambda^{-t}\sum_{i=1}^d e^{(i)}\left(e^{(i)}\right)' + \lambda^{-t}x_t x_t' = \lambda^{-t}(b^2 I + x_t x_t')$; thus

$$\begin{aligned}\tilde{Z}_{(d+1)t} &= aI + \sum_{q=1}^{t}\sum_{s\in B_t}\tilde{x}_s\tilde{x}_s' = aI + \sum_{q=1}^{t}\lambda^{-q}(b^2 I + x_q x_q') \\ &= \lambda^{-t}\left(aI + \sum_{q=1}^{t}\lambda^{t-q}x_q x_q'\right) = \lambda^{-t}Z_t\end{aligned}$$

(since $\sum_{i=1}^d e^{(i)}\left(e^{(i)}\right)' = I$ and $b^2=(1-\lambda)a$; recall (7)). Together with $\tilde{v}_{(d+1)t} = \sum_{q=1}^t\sum_{s\in B_t}\tilde{y}_s\tilde{x}_s = \sum_{q=1}^t \lambda^{-q}y_q x_q = \lambda^{-t}v_t$ (only the first entry in each block has a nonzero $\tilde{y}$), it follows that $\tilde{\theta}_{(d+1)t}$ indeed equals $\theta_t = Z_t^{-1}v_t$ as given by (7).

Next, for every t we have $\sum_{s\in B_t}\tilde{\psi}_s(\tilde{\theta}_s)\geq\lambda^{-t}\psi_t(\theta_t)$ (all terms in the sum are nonnegative, and we drop all except the last one). Also, for every $\theta\in\mathbb{R}^d$,

$$\sum_{s\in B_t}\tilde{\psi}_s(\theta) = \lambda^{-t}\left(b^2\sum_{i=1}^d(\theta' e^{(i)})^2 + \psi_t(\theta)\right) = \lambda^{-t}\left(b^2\left\|\theta\right\|^2 + \psi_t(\theta)\right).$$

Thus the left-hand side of (6) evaluated at the end of the T-th block B_T satisfies

$$\begin{aligned}LHS &\geq \sum_{t=1}^T\lambda^{-t}\psi_t(\theta_t) - a\left\|\theta\right\|^2 - b^2\left\|\theta\right\|^2\sum_{t=1}^T\lambda^{-t} - \sum_{t=1}^T\lambda^{-t}\psi_t(\theta) \\ &= \sum_{t=1}^T\lambda^{-t}\left[\psi_t(\theta_t)-\psi_t(\theta)\right] - \lambda^{-T}a\left\|\theta\right\|^2.\end{aligned}$$

On the right-hand side we get

$$RHS = \sum_{t=1}^T\sum_{s\in B_t}\tilde{y}_s^2\left(1-\frac{\det(\tilde{Z}_{s-1})}{\det(\tilde{Z}_s)}\right) = \sum_{t=1}^T\lambda^{-t}y_t^2\left(1-\frac{\det(\tilde{Z}_{(d+1)t-1})}{\det(\tilde{Z}_{(d+1)t})}\right)$$

(again, only the last term in each block has nonzero $\tilde{y}_s$). We have seen above that $\tilde{Z}_{(d+1)t}=\lambda^{-t}Z_t$; thus $\tilde{Z}_{(d+1)t-1} = \tilde{Z}_{(d+1)t} - \lambda^{-t}x_t x_t' = \lambda^{-t}(Z_t - x_t x_t') =$

[11] $\tilde{Z}_s$, $\tilde{\psi}_s$, ... pertain to the $(\tilde{x}_s,\tilde{y}_s)_{s\geq 1}$ problem.

$\lambda^{-t+1}Z_{t-1} + (\lambda^{-t} - \lambda^{-t+1})aI$. Therefore $\det(\tilde{Z}_{(d+1)t-1}) \geq \det(\lambda^{-t+1}Z_{t-1})$ (indeed, if B is a positive definite matrix and $\beta > 0$ then[12] $\det(B + \beta I) > \det(B)$). Therefore we obtain

$$\begin{aligned} RHS &\leq \sum_{t=1}^{T} \lambda^{-t} y_t^2 \left(1 - \frac{\det(\lambda^{-t+1}Z_{t-1})}{\det(\lambda^{-t}Z_t)}\right) \\ &= \sum_{t=1}^{T} \lambda^{-t} y_t^2 \left(1 - \lambda^d \frac{\det(Z_{t-1})}{\det(Z_t)}\right) \end{aligned}$$

(the matrices Z_t are of size $d \times d$, and so $\det(cZ_t) = c^d \det(Z_t)$). Recalling that $LHS \leq RHS$ by (6) and multiplying by λ^T yields the result. □

Remark. From now on it will be convenient to assume that $\|x_t\| \leq 1$ and $|y_t| \leq 1$ (i.e., $X = Y = 1$); for general X and Y, multiply $x_t, y_t, \theta_t, \psi_t, a$ by $X, Y, Y/X, Y^2, X^2$, respectively, in the appropriate formulas.

Proposition 5 *For every $a > 0$ and $1/4 \leq \lambda < 1$ there exists a constant $D_1 \equiv D_1(a, \lambda, d)$ such that the λ-discounted a-forward algorithm yields*

$$\sum_{t=1}^{T} \lambda^{T-t} \left[\psi_t(\theta_t) - \psi_t(\theta)\right] \leq a \|\theta\|^2 + D_1, \tag{9}$$

for every $T \geq 1$, every $\theta \in \mathbb{R}^d$, and every sequence $(x_t, y_t)_{t \geq 1}$ in $\mathbb{R}^d \times \mathbb{R}$ with $\|x_t\| \leq 1$ and $|y_t| \leq 1$ for all t.

Proof. Let $K \geq 1$ be an integer such that $1/4 \leq \lambda^K \leq 1/2$. Given $T \geq 1$, let the integer $m \geq 1$ satisfy $(m-1)K < T \leq mK$. Writing ζ_t for $\det(Z_t)$,

[12] Let $\beta_1, ..., \beta_d > 0$ be the eigenvalues of B; then the eigenvalues of $B + cI$ are $\beta_1 + c, ..., \beta_d + c$, and so $\det(B + cI) = \prod_i(\beta_i + c) > \prod_i \beta_i = \det(B)$.

we have

$$
\begin{aligned}
\sum_{t=1}^{T} \lambda^{T-t}\left(1-\lambda^{d}\frac{\zeta_{t-1}}{\zeta_t}\right) &\le \sum_{t=1}^{mK} \lambda^{T-t}\left(1-\lambda^{d}\frac{\zeta_{t-1}}{\zeta_t}\right) \\
&= \lambda^{T}\sum_{j=1}^{m-1}\sum_{t=jK+1}^{(j+1)K} \lambda^{-t}\ln\left(\lambda^{-d}\frac{\zeta_t}{\zeta_{t-1}}\right) \\
&\le \lambda^{T}\sum_{j=1}^{m-1}\lambda^{-(j+1)K}\sum_{t=jK+1}^{(j+1)K} \ln\left(\lambda^{-d}\frac{\zeta_t}{\zeta_{t-1}}\right) \\
&\le \lambda^{T}\sum_{j=1}^{m-1}\lambda^{-(j+1)K}\ln\left(\lambda^{-dK}\frac{\zeta_{(j+1)K}}{\zeta_{jK}}\right) \qquad (10)
\end{aligned}
$$

(in the second line we have used $1-1/u \le \ln u$ for $0 < u \le 1$, as in (4.21) in Azoury and Warmuth 2001).

Let $B = (b_{ij})$ be a $d \times d$ symmetric positive definite matrix with $|b_{ij}| \le \beta$ for all i, j, and let $a > 0$. Then $a^d \le \det(aI + B) \le d!(a+\beta)^d$. Indeed, the second inequality follows easily since the determinant is the sum of $d!$ products of d elements each). For the first inequality, let $\beta_1, ..., \beta_d > 0$ be the eigenvalues of B; then the eigenvalues of $aI + B$ are $a + \beta_1, ..., a + \beta_d$, and so $\det(aI + B) = \Pi_{i=1}^{d}(a + \beta_i) > a^d$. Applying this to Z_t (using (7), $|x_{t,i}x_{t,j}| \le \|x_t\|^2 \le 1$, and $\sum_{t=1}^{T}\lambda^{T-t} < 1/(1-\lambda)$) yields

$$a^d \le \zeta_t \equiv \det(Z_t) \le d!\left(a + \frac{1}{1-\lambda}\right)^d .$$

Therefore, since $\lambda^{-K} \le 4$, we get

$$\lambda^{-dK}\frac{\zeta_{(j+1)K}}{\zeta_{jK}} \le 4^d d!\left(1 + \frac{1}{a(1-\lambda)}\right)^d =: D,$$

and so (10) is

$$\begin{aligned}
&\le \lambda^T \sum_{j=1}^{m-1} \left(\lambda^{-K}\right)^{j+1} \ln D \le \lambda^T \frac{\lambda^{-K(m+1)} - \lambda^{-2K}}{\lambda^{-K} - 1} \ln D \\
&\le \lambda^T \frac{\lambda^{-T}\lambda^{-K} - 0}{2-1} \ln D = 4 \ln D
\end{aligned}$$

(since $2 \le \lambda^{-K} \le 4$ and $K(m+1) < T + K$). Substituting this in (8) and putting

$$D_1 := 4\left(\ln d! + d \ln 4 + d \ln\left(1 + \frac{1}{a(1-\lambda)}\right)\right) \tag{11}$$

completes the proof. □

3.3 Windowed Discounted Forward Algorithm

From now on it is convenient to put $(x_t, y_t, \theta_t) = (0, 0, 0)$ for all $t \le 0$.

Let $a > 0$, $0 < \lambda < 1$, and integer $R \ge 1$. The *R-windowed λ-discounted a-forward algorithm* gives $\theta_t = Z_t^{-1} v_t$, where[13]

$$Z_t = aI + \sum_{q=t-R+1}^{t} \lambda^{t-q} x_q x_q' \quad \text{and} \quad v_t = \sum_{q=t-R+1}^{t-1} \lambda^{R-q} y_q x_q. \tag{12}$$

Lemma 6 *For every $a > 0$ and $0 < \lambda < 1$ there exists a constant $D_2 \equiv D_2(a, \lambda, d)$ such that if $(\tilde{\theta}_t)_{t\ge1}$ is given by the λ-discounted a-forward algorithm, and $(\theta_t)_{t\ge1}$ is given by the R-windowed λ-discounted a-forward algorithm for some integer $R \ge 1$, then*

$$\left|\psi_t(\tilde{\theta}_t) - \psi_t(\theta_t)\right| \le D_2 \lambda^R \tag{13}$$

for every[14] $t \ge 1$ and every sequence $(x_t, y_t)_{t\ge1}$ in $\mathbb{R}^d \times \mathbb{R}$ with $\|x_t\| \le 1$ and $|y_t| \le 1$ for all t.

To prove this lemma we use the following basic result. The norm of a

[13] The sums below effectively start at $\min\{t - R + 1, 1\}$ (because we put $x_q = 0$ for $q \le 0$).

[14] For $t \le R$ we have $\tilde{\theta}_t = \theta_t$ since they are given by the same formula.

matrix A is $\|A\| := \max_{z\neq 0} \|Az\| / \|z\|$.

Lemma 7 *For $k = 1, 2$, let $c_k = A_k^{-1} b_k$, where A_k is a $d \times d$ symmetric matrix with eigenvalues $\geq \alpha > 0$, and $\|b_k\| \leq M$. Then $\|c_k\| \leq M/\alpha$ and*

$$\|c_1 - c_2\| \leq \frac{1}{\alpha} \|b_1 - b_2\| + \frac{M}{\alpha^2} \|A_1 - A_2\| .$$

Proof. First, $\|c_k\| = \left\|A_k^{-1}\right\| \|b_k\| \leq (1/\alpha)M$ since $\left\|A_k^{-1}\right\|$ is the maximal eigenvalue of A_k^{-1}, which is the reciprocal of the minimal eigenvalue of A_k, and so $\left\|A_k^{-1}\right\| \leq 1/\alpha$.

Second, express $c_1 - c_2$ as $A_1^{-1}(b_1 - b_2) + A_1^{-1}(A_2 - A_1)A_2^{-1}b_2$, to get

$$\|c_1 - c_2\| \leq \left\|A_1^{-1}\right\| \|b_1 - b_2\| + \left\|A_1^{-1}\right\| \|A_2 - A_1\| \left\|A_2^{-1}\right\| \|b_2\|$$

and the proof is complete. □

Proof of Lemma 6. For $t \leq R$ we have $\tilde{\theta}_t \equiv \theta_t$, and so take $t > R$. We have[15] $\|\tilde{v}_t\|, \|v_t\| \leq \sum_{q=1}^{\infty} \lambda^q = 1/(1-\lambda)$. The matrices $\tilde{Z}_t$ and Z_t are the sum of aI and a positive-definite matrix, and so their eigenvalues are $\geq a$. Next,

$$\|\tilde{v}_t - v_t\| = \left\| \sum_{q=1}^{t-R} \lambda^{t-q} y_q x_q \right\| \leq \frac{\lambda^R}{1-\lambda};$$

similarly, for each each element $(\tilde{Z}_t - Z_t)_{ij}$ of $\tilde{Z}_t - Z_t$ we have

$$\left|(\tilde{Z}_t - Z_t)_{ij}\right| = \left| \sum_{q=1}^{t-R} \lambda^{t-q} x_{q,i} x_{q,j} \right| \leq \frac{\lambda^R}{1-\lambda},$$

and so[16] $\left\|\tilde{Z}_t - Z_t\right\| \leq d\lambda^R/(\lambda - 1)$. Using Lemma 7 yields

$$\left\|\tilde{\theta}_t - \theta_t\right\| \leq \frac{1}{a}\frac{\lambda^R}{1-\lambda} + \frac{1}{a^2}\frac{d\lambda^R}{1-\lambda} = \frac{\lambda^R(a+d)}{(1-\lambda)a^2}.$$

[15] $\tilde{v}_t$ and $\tilde{Z}_t$ pertain to the sequence $\tilde{\theta}_t$ given by the λ-discounted a-forward algorithm, whereas v_t and Z_t pertain to the sequence θ_t given by the R-windowed λ-discounted a-forward algorithm.

[16] Because $\|A\| \leq d \max_{i,j} |a_{ij}|$ for any $d \times d$ matrix A.

Hence

$$\begin{aligned}\left|\psi_t(\tilde{\theta}_t)-\psi_t(\theta_t)\right| &= \left|(y_t-\tilde{\theta}_t' x_t)^2-(y_t-\theta_t' x_t)^2\right| \\ &= \left|(\tilde{\theta}_t'-\theta_t')x_t\cdot\left(2y_t-(\tilde{\theta}_t'+\theta_t')x_t\right)\right| \\ &\leq \left\|\tilde{\theta}_t-\theta_t\right\|\left(2+\left\|\tilde{\theta}_t\right\|+\|\theta_t\|\right) \\ &\leq \frac{\lambda^R(a+d)}{(1-\lambda)a^2}\left(2+\frac{2}{(1-\lambda)a}\right)=D_2\lambda^R,\end{aligned}$$

where

$$D_2 := \frac{2\,(a+d)\,(a(1-\lambda)+1)}{a^3(1-\lambda)^2}; \tag{14}$$

this completes the proof. □

Proposition 8 *For every $a>0$ and $1/4\leq\lambda<1$ there exist constants $D_1\equiv D_1(a,\lambda,d)$ and $D_2\equiv D_2(a,\lambda,d)$ such that for every integer $R\geq 1$ the R-windowed λ-discounted a-forward algorithm yields*

$$\begin{aligned}\frac{1}{R}\sum_{t=T-R+1}^{T}\left[\psi_t(\theta_t)-\psi_t(\theta)\right] &\leq \left(a\left\|\theta\right\|^2+D_1\right)\left(1-\lambda+\frac{\lambda}{R}\right) \\ &\quad +\frac{(\|\theta\|+1)^2}{R(1-\lambda)}+D_2\lambda^R\end{aligned} \tag{15}$$

for every $T\geq 1$, every $\theta\in\mathbb{R}^d$, and every sequence $(x_t,y_t)_{t\geq 1}$ in $\mathbb{R}^d\times\mathbb{R}$ with $\|x_t\|\leq 1$ and $|y_t|\leq 1$ for all t.

Proof. Let $\tilde{\theta}_t$ be given by the λ-discounted a-forward algorithm. Put $g_t := \psi_t(\theta_t)-\psi_t(\theta)$ (where θ_t is given by the R-windowed λ-discounted a-forward algorithm) and $\tilde{g}_t := \psi_t(\tilde{\theta}_t)-\psi_t(\theta)$. Apply (9) at T, and also at each one of $T-R+1, T-R+2, ..., T-1$; multiply those by $1-\lambda$ and add them all, to get

$$\begin{aligned}\sum_{t=1}^{T}\lambda^{T-t}\tilde{g}_t+(1-\lambda)\sum_{r=1}^{R-1}\sum_{t=1}^{T-r}\lambda^{T-r-t}\tilde{g}_t &\leq \left(a\left\|\theta\right\|^2+D_1\right)(1+(R-1)(1-\lambda)) \\ &= \left(a\left\|\theta\right\|^2+D_1\right)(R-R\lambda+\lambda).\end{aligned}$$

For $t \leq T - R$, the total coefficient of $\tilde{g}_t$ on the left-hand side above is $\lambda^{T-t} + (1-\lambda)\sum_{r=1}^{R-1} \lambda^{T-r-t} = \lambda^{T-R+1-t}$; for $T - R + 1 \leq t \leq T$, it is $\lambda^{T-t} + (1-\lambda)\sum_{r=1}^{T-t} \lambda^{T-r-t} = 1$. Therefore

$$\sum_{t=1}^{T-R} \lambda^{T-R+1-t}\tilde{g}_t + \sum_{t=T-R+1}^{T} \tilde{g}_t \leq (a\|\theta\|^2 + D_1)(R - R\lambda + \lambda).$$

Now $\tilde{g}_t \geq -\psi_t(\theta) \geq -(\|\theta\|\,\|x_t\| + |y_t|)^2 \geq -(\|\theta\| + 1)^2$, and so

$$\begin{aligned}\sum_{t=T-R+1}^{T} \tilde{g}_t &\leq (a\|\theta\|^2 + D_1)(R - R\lambda + \lambda) + (\|\theta\| + 1)^2 \sum_{t=1}^{T-R} \lambda^{T-R+1-t} \\ &\leq (a\|\theta\|^2 + D_1)(R - R\lambda + \lambda) + \frac{(\|\theta\| + 1)^2}{1-\lambda}.\end{aligned}$$

Divide by R and use $g_t \leq \tilde{g}_t + D_2\lambda^R$ (by Proposition 6). □

Choosing appropriate λ and R allows us to bound the right-hand side of (15).

Proposition 9 *For every $\varepsilon > 0$ and $a > 0$ there is $\lambda_0 \equiv \lambda_0(\varepsilon, a, d) < 1$ such that for every $\lambda_0 < \lambda < 1$ there is $R_0 \equiv R_0(\varepsilon, a, d, \lambda) \geq 1$ such that for every $R \geq R_0$ the R-windowed λ-discounted a-forward algorithm yields*

$$\frac{1}{R}\sum_{t=T-R+1}^{T} [\psi_t(\theta_t) - \psi_t(\theta)] \leq \varepsilon(1 + \|\theta\|^2), \tag{16}$$

$$\frac{1}{T}\sum_{t=1}^{T} [\psi_t(\theta_t) - \psi_t(\theta)] \leq \varepsilon(1 + \|\theta\|^2), \tag{17}$$

for every $T \geq R$, every $\theta \in \mathbb{R}^d$, and every sequence $(x_t, y_t)_{t\geq 1}$ in $\mathbb{R}^d \times \mathbb{R}$ with $\|x_t\| \leq 1$ and $|y_t| \leq 1$ for all t.

Proof. The right-hand side of (15) is

$$\leq \left(D_1(1-\lambda) + \frac{D_1}{R} + \frac{2}{R(1-\lambda)} + D_2\lambda^R\right) + \|\theta\|^2 \left(a(1-\lambda) + \frac{a}{R} + \frac{2}{R(1-\lambda)}\right)$$

(use $\lambda/R \leq 1/R$ and $(\|\theta\| + 1)^2 \leq 2\|\theta\|^2 + 2$). First, take $1/4 \leq \lambda_0 < 1$ close enough to 1 so that $a(1 - \lambda_0) \leq \varepsilon/4$ and $D_1(a, \lambda_0, d) \cdot (1 - \lambda_0) \leq \varepsilon/4$ (recall formula (11) for D_1 and use $\lim_{x \to 0^+} x \ln x = 0$). Then, given $\lambda \in [\lambda_0, 1)$, take $R_0 \geq 1$ large enough so that $a/R_0 \leq \varepsilon/4$, $D_1(a, \lambda, d)/R_0 \leq \varepsilon/4$, $2/(R_0(1 - \lambda)) \leq \varepsilon/4$, and $D_2(a, \lambda, d)\lambda^{R_0} \leq \varepsilon/4$. This shows (16) for every $T \geq 1$.

In particular, for $T' < R$ we get $(1/R) \sum_{t=1}^{T'} [\psi_t(\theta_t) - \psi_t(\theta)] \leq \varepsilon(1 + \|\theta\|^2)$ (because $(x_t, y_t, \theta_t) = (0, 0, 0)$ for all $t \leq 0$). For $T \geq R$, add inequality (16) for the disjoint blocks of size R that end at $t = T$, together with the above inequality for the initial smaller block of size $T' < R$ when T is not a multiple of R, to get $(1/R) \sum_{t=1}^{T} [\psi_t(\theta_t) - \psi_t(\theta)] \leq \lceil T/R \rceil \varepsilon(1 + \|\theta\|^2) \leq 2(T/R)\varepsilon(1 + \|\theta\|^2)$. Replacing ε with $\varepsilon/2$ yields (17). □

Remark. Similar arguments show that, for $\lambda_0 \leq \lambda < 1$, the discounted average is also small:

$$\frac{1 - \lambda}{1 - \lambda^T} \sum_{t=1}^{T} \lambda^{T-t} [\psi_t(\theta_t) - \psi_t(\theta)] \leq \varepsilon(1 + \|\theta\|^2).$$

Proposition 9 yields the main result of this section, Theorem 2.

Proof of Theorem 2. Use Proposition 9 (with, say, $a = 1$), and rescale everything by X and Y appropriately (see the Remark before Proposition 5). □

4 Weak Calibration

The notion of "weak calibration" was introduced by Kakade and Foster (2004) and Foster and Kakade (2006). The idea is as follows. Given a "test" function $w : C \to \{0, 1\}$ that indicates which forecasts c to consider, let the corresponding score be[17] $S_T^w := ||(1/T) \sum_{t=1}^{T} w(c_t)(a_t - c_t)||$. It can be shown

[17] The S_T scores are norms of averages, rather than averages of norms like the K_T scores. "Windowed" versions of the scores may also be considered (with the average taken over the last R periods only; cf. (3)).

that if S_T^w is small for every such w, then the calibration score K_T is also small.[18]

Now instead of the discontinuous indicator functions, *weak calibration* requires that S_T^w be small for continuous "weight" functions $w : C \to [0, 1]$. Specifically, let $\varepsilon > 0$ and $L < \infty$. A procedure (i.e., a strategy of the C-player in the calibration game) is (ε, L)-*weakly calibrated* if there is $T_0 \equiv T_0(\varepsilon, L)$ such that

$$S_T^w = \left\| \frac{1}{T} \sum_{t=1}^{T} w(c_t)(a_t - c_t) \right\| \leq \varepsilon \tag{18}$$

holds for every strategy of the A-player, every $T > T_0$, and every weight function $w : C \to [0, 1]$ that is L-Lipschitz (i.e., $\mathcal{L}(w) \leq L$).

The importance of weak calibration is that, unlike regular calibration, it can be guaranteed by deterministic procedures (which are thus leaky): Kakade and Foster (2004) and Foster and Kakade (2006) have proven the existence of deterministic (ε, L)-weakly calibrated procedures. Moreover, as we will show in the next section, weak calibration is essentially equivalent to smooth calibration.

We now provide a deterministic (ε, L)-weakly calibrated procedure that in addition has finite recall and is stationary.

Theorem 10 *For every $\varepsilon > 0$ and $L < \infty$ there exists an (ε, L)-weakly calibrated deterministic procedure that has finite recall and is stationary; moreover, all its forecasts may be taken to lie on a finite grid.*

Proof. Without loss of generality assume that $A \subseteq C \subseteq [0, 1]^m$ (one can always translate the sets A and C—which does not affect (18)—and rescale them—which just rescales the Lipschitz constant); assume also that $L \geq 1$ (as L increases there are more Lipschitz functions) and $\varepsilon \leq 1$.

[18] Specifically, if $S_T^w \leq \varepsilon$ for all $w : C \to \{0, 1\}$ then $K_T \leq 2m\varepsilon$. Indeed, for each coordinate $i = 1, ..., m$, let C_+^i be the set of all c_t such that $\bar{a}_{t,i} > c_{t,i}$, and C_-^i the set of all c_t such that $\bar{a}_{t,i} < c_{t,i}$. Taking w to be the indicator of C_+^i yields $S_T^w = (1/T) \sum_t [\bar{a}_{t,i} - c_{t,i}]_+ \leq \varepsilon$ (where $[z]_+ := \max\{z, 0\}$); similarly, the indicator of C_-^i yields $(1/T) \sum_t [\bar{a}_{t,i} - c_{t,i}]_- \leq \varepsilon$. Adding the two inequalities gives $(1/T) \sum_t |\bar{a}_{t,i} - c_{t,i}| \leq 2\varepsilon$. Since this holds for each one of the m coordinates, it follows that $K_T \leq 2m\varepsilon$.

For every $b \in \mathbb{R}^m$ let $\gamma(b) := \arg\min_{c\in C} \|c - b\|$ be the closest point to b in C (it is well defined and unique since C is a convex compact set); then

$$\|c - b\| \geq \|c - \gamma(b)\| \tag{19}$$

for every $c \in C$ (because $\|c-b\|^2 = \|c-\gamma(b)\|^2 + \|b-\gamma(b)\|^2 - 2(b-\gamma(b))\cdot(c-\gamma(b))$ and the third term is ≤ 0).

Let $\varepsilon_1 := \varepsilon/(2\sqrt{m})$. Denote by W_L the set of weight functions $w : C \to [0,1]$ with $\mathcal{L}(w) \leq L$. By Lemma 16 in the Appendix, for every $w \in W_L$ there is a vector $\varpi \equiv \varpi_w \in [0,1]^d$ such that[19]

$$\max_{c\in C}\left| w(c) - \sum_{i=1}^{d} \varpi_i f_i(c) \right| \leq \varepsilon_1. \tag{20}$$

Denote $F(c) := (f_1(c), ..., f_d(c)) \in [0,1]^d$; thus $\|F(c)\| \leq \sqrt{d}$. Without loss of generality we assume that the coordinate functions are included in the set $\{f_1, ..., f_d\}$, say, $f_j(c) = c_j$ for $j = 1, ..., m$ (thus $d > m$, in fact d is much larger than m).

Let $\varepsilon_2 := \varepsilon/(m + m(1+d)^2 + d^2)$ (where d is given above, and depends on ε, m, and L) and $\varepsilon_3 := (\varepsilon_2)^2$.

Let λ and R be given by Theorem 2 and Proposition 9 for $a = 1$, $X = \sqrt{d}$, $Y = 1$, and $\varepsilon = \varepsilon_3$. For each $j = 1, ..., m$ consider the sequence $(x_t, y_t^{(j)})_{t\geq 1} = (F(c_t), a_{t,j})_{t\geq 1}$ in $\mathbb{R}^d \times \mathbb{R}$, where $a_t \in A$ is determined by the A-player, and $c_t \in C$ is constructed inductively as follows.

[19]Since W_L is compact in the sup norm, there are $f_1, ..., f_d \in W_L$ such that for every $w \in W_L$ there is $1 \leq i \leq d$ with $\max_{c\in C} |w(c) - f_i(c)| \leq \varepsilon_1$. Lemma 16 improves this, in getting a much smaller d by using linear combinations with bounded coefficients.

Let the history be $h_{t-1} = (c_1, a_1, ..., c_{t-1}, a_{t-1})$. For each $b \in \mathbb{R}^m$, let[20]

$$\begin{aligned}
Z_t(b) &= I + \sum_{q=1}^{R-1} \lambda^{R-q} x_q x_q' + F(\gamma(b))F(\gamma(b))' \in \mathbb{R}^{d \times d}, \\
v_t^{(j)} &= \sum_{q=1}^{R-1} \lambda^{R-q} a_{q,j} x_q \in \mathbb{R}^d, \\
H_{t,j}(b) &= \left(Z_t(b)^{-1} v_t^{(j)} \right)' F(\gamma(b)) \in \mathbb{R}, \\
H_t(b) &= (H_{t,1}(b), ..., H_{t,m}(b)) \in \mathbb{R}^m
\end{aligned}$$

(where $x_q = F(c_q)$ for $q < t$); thus, $H_t(b) = H_t(\gamma(b))$ for every $b \in \mathbb{R}^m$.

We have $\left\| v_t^{(j)} \right\| \leq \sqrt{d}\lambda/(1-\lambda)$ (since $|a_{q,j}| \leq 1$ and $\|x_q\| = \|F(c_q)\| \leq \sqrt{d}$), and so $\left\| Z_t(b)^{-1} v_t^{(j)} \right\| \leq \sqrt{d}\lambda/(1-\lambda)$ by Lemma 7 ($Z_t(b)$ is positive definite and its eigenvalues are ≥ 1), which finally implies that $\left| H_{t,j}(b) \right| \leq \sqrt{d}\lambda/(1-\lambda) \cdot \sqrt{d} = d\lambda/(1-\lambda) =: K$. Therefore the restriction of H_t to the compact and convex set $[-K, K]^m$, which is clearly a continuous function (since, again, $Z_t(b)$ is positive definite and its eigenvalues are ≥ 1), has a fixed point (by Brouwer's Fixed-Point Theorem), which we denote b_t (any fixed point will do);[21] put $c_t := \gamma(b_t) \in C$. Thus

$$c_t = \gamma(b_t) \;\text{ and }\; b_t = H_t(b_t) = H_t(c_t).$$

Define $x_t := F(\gamma(b_t)) = F(c_t)$ and $\theta_t^{(j)} := Z_t(b_t)^{-1} v_t^{(j)} = Z_t(c_t)^{-1} v_t^{(j)} \in \mathbb{R}^d$. Then $Z_t(c_t) = I + \sum_{q=1}^{R} \lambda^{R-q} x_q x_q'$, and thus it corresponds to the R-windowed λ-discounted 1-forward algorithm (see (12)). Therefore, for every $j = 1, ..., m$ and every $\theta^{(j)} \in \mathbb{R}^d$ we have by (4)

$$\frac{1}{T} \sum_{t=1}^{T} \left[\psi_t^{(j)}(\theta_t^{(j)}) - \psi_t^{(j)}(\theta^{(j)}) \right] \leq \varepsilon_3 \left(1 + \left\| \theta^{(j)} \right\|^2 \right) \tag{21}$$

[20] A subscript j stands for the j-th coordinate (e.g., $a_{t,j}$ is the j-th coordinate of a_t), whereas a superscript j refers to the j-th procedure (e.g., $v_t^{(j)}$).

[21] There may be more than one fixed point here, in which case we may choose the fixed point at random, and obtain a *randomized* procedure that satisfies everything the deterministic procedure does. Using it yields in Theorem 1 a randomized procedure that is smoothly *leaky*-calibrated (cf. Section 2.3).

for all $T \geq T_0 \equiv R$, where $\psi_t^{(j)}(\theta) = (a_{t,j} - \theta' x_t)^2$, and thus $\psi_t^{(j)}(\theta_t^{(j)}) = (a_{t,j} - b_{t,j})^2$ (recall that $b_{t,j} = H_{t,j}(b_t) = \left(\theta_t^{(j)}\right)' F(\gamma(b_t)) = \left(\theta_t^{(j)}\right)' x_t$). Summing over j yields

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\left[\psi_t^{(j)}(\theta_t^{(j)}) - \psi_t^{(j)}(\theta^{(j)})\right] \leq \varepsilon_3 \left(m + \sum_{j=1}^{m}\left\|\theta^{(j)}\right\|^2\right).$$

Now $\sum_{j=1}^{m} \psi_t^{(j)}(\theta_t^{(j)}) = \sum_{j=1}^{m}(a_{t,j} - b_{t,j})^2 = \|a_t - b_t\|^2 \geq \|a_t - \gamma(b_t)\|^2 = \|a_t - c_t\|^2 = \sum_{j=1}^{m} (a_{t,j} - c_{t,j})^2$ (by the definition of $\gamma(b_t)$ and (19), since $a_t \in A \subseteq C$), and therefore

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\left[(a_{t,j} - c_{t,j})^2 - \psi_t^{(j)}(\theta^{(j)})\right] \leq \varepsilon_3 \left(m + \sum_{j=1}^{m}\left\|\theta^{(j)}\right\|^2\right). \tag{22}$$

Given a weight function $w \in W_L$, let the vector $\varpi \equiv \varpi_w \in [0,1]^d$ satisfy (20), i.e., $|w(c) - \varpi' F(c)| \leq \varepsilon_1$ for all $c \in C$. Take $u = (u_j)_{j=1,...,m} \in \mathbb{R}^m$ with $\|u\| = 1$. For every $j = 1, ..., m$, take $\theta^{(j)} = e^{(j)} + \varepsilon_2 u_j \varpi \in \mathbb{R}^d$, where $e^{(j)} \in \mathbb{R}^d$ is the j-th unit vector; thus $\left\|\theta^{(j)}\right\| \leq 1 + \varepsilon_2 d \leq 1 + d$ (since $\varepsilon_2 \leq \varepsilon \leq 1$). We have

$$(\theta^{(j)})' x_t = (\theta^{(j)})' F(c_t) = c_{t,j} + \varepsilon_2 \left(\varpi' F(c_t)\right) u_j$$

(since $f_j(c) = c_j$ for $j \leq m$), and hence

$$\begin{aligned}(a_{t,j} - c_{t,j})^2 - \psi_t^{(j)}(\theta^{(j)}) &= (a_{t,j} - c_{t,j})^2 - (a_{t,j} - c_{t,j} - \varepsilon_2 \left(\varpi' F(c_t)\right) u_j)^2 \\ &= 2\varepsilon_2 \left(\varpi' F(c_t)\right) u_j (a_{t,j} - c_{t,j}) - (\varepsilon_2 \varpi' F(c_t)^2 u_j^2.\end{aligned}$$

Summing over $j = 1, ..., m$ yields

$$\begin{aligned}\sum_{j=1}^{m}\left[(a_{t,j} - c_{t,j})^2 - \psi_t^{(j)}(\theta^{(j)})\right] &= 2\varepsilon_2 \left(\varpi' F(c_t)\right) u'(a_t - c_t) - (\varepsilon_2 \varpi' F(c_t)^2 \|u\|^2 \\ &\geq 2\varepsilon_2 \left(\varpi' F(c_t)\right) u'(a_t - c_t) - (\varepsilon_2)^2 d^2 \\ &\geq 2\varepsilon_2 w(c_t) u'(a_t - c_t) - \varepsilon_1 \cdot 2\varepsilon_2 \|u\| \|a_t - c_t\| - (\varepsilon_2)^2 d^2 \\ &\geq 2\varepsilon_2 w(c_t) u'(a_t - c_t) - 2\varepsilon_1 \varepsilon_2 \sqrt{m} - (\varepsilon_2)^2 d^2\end{aligned}$$

(since: $\|u\| = 1$, $|\varpi' F(c)| \leq d$ [the coordinates of ϖ are between -1 and 1 and those of $F(c)$ between 0 and 1], $\|a_t - c_t\| \leq \sqrt{m}$ (since $a_t, c_t \in [0,1]^m$, and recall (20)).

Together with (22) we get (recall that $\varepsilon_3 = (\varepsilon_2)^2$ and $\varepsilon_1 = \varepsilon/(2\sqrt{m})$):

$$2\varepsilon_2 \cdot \frac{1}{T}\sum_{t=1}^{T} w(c_t)u'(a_t - c_t) \leq (\varepsilon_2)^2(m + m(1+d)^2 + d^2) + \varepsilon\varepsilon_2;$$

hence, dividing by $2\varepsilon_2$ and recalling that $\varepsilon_2 = \varepsilon/(m + m(1+d)^2 + d^2))$:

$$u \cdot \frac{1}{T}\sum_{t=1}^{T} w(c_t)(a_t - c_t) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since $u \in \mathbb{R}^m$ with $\|u\| = 1$ was arbitrary, the proof is complete.

For the "moreover" statement, let $\varepsilon_4 := \varepsilon/(L\sqrt{m} + 1)$, and take $D \subseteq C$ to be a finite ε_4-grid in C; i.e., for every $c \in C$ there is $d(c) \in D$ with $||d(c) - c|| \leq \varepsilon_4$. Replace the forecast c_T obtained above with $\tilde{c}_T := d(c_T)$; then, for every $a_T \in A$, we have

$$||w(c_T)(a_T - c_T) - w(\tilde{c}_T)(a_T - \tilde{c}_T)|| \leq L\varepsilon_4\sqrt{m} + \varepsilon_4 = \varepsilon.$$

Therefore the score S_T^w changes by at most ε, and so it is at most 2ε. □

5 Smooth Calibration

In this section we show (Propositions 11 and 12) that weak calibration and smooth calibration are essentially equivalent (albeit with different constants ε, L). The existence of weakly calibrated procedures (Theorem 10, proved in the previous section) then implies the existence of smoothly calibrated procedures, which proves Theorem 1.

We first show how to go from weak to smooth calibration.

Proposition 11 *An* (ε, L)*-weakly calibrated procedure is* (ε', L)*-smoothly calibrated, where*[22] $\varepsilon' = \Omega\left(\sqrt{\varepsilon L^m}\right)$.

Proof. Take $\varepsilon' := \sqrt{\varepsilon L^m}(2\sqrt{m})^{m/2}(2+\sqrt{m})$.

Let $(D_j)_{j=1,...,M}$ be a partition of $[0,1]^m \supseteq C$ into disjoint cubes with side $1/(2L\sqrt{m})$; the diameter of each cube is thus $1/(2L)$, and the number of cubes is $M = (2L\sqrt{m})^m$.

Fix $a_t, c_t \in C \subseteq [0,1]^m$ for $t = 1, ..., T$, and a smoothing function Λ that is L-Lipschitz in its first coordinate. Assume that the weak calibaration score S_T^w (given by (18)) satisfies $S_T^w \leq \varepsilon$ for every weight function w in W_L; we will show that the smooth calibration score K_T^Λ (given by (1)) satisfies $K_T^\Lambda \leq \varepsilon'$.

For every $t = 1, ..., T$, take $w_t(c) := \Lambda(c, c_t)$; then w_t is a weight function with $\mathcal{L}(w_t) \leq L$, and so, by our assumption,

$$\left\| \sum_{s=1}^{T} w_t(c_s)(a_s - c_s) \right\| \leq T\varepsilon. \tag{23}$$

Now $\bar{a}_t^\Lambda - c_t^\Lambda = \sum_{s=1}^T w_t(c_s)(a_s - c_s)/W_t$, where $W_t := \sum_{s=1}^T w_t(c_s)$, and so (23) yields

$$\left\|\bar{a}_t^\Lambda - c_t^\Lambda\right\| \leq T\varepsilon \frac{1}{W_t}. \tag{24}$$

When W_t is large, (24) provides a good bound; we will show that, for a large proportion of indices t, this is indeed the case.

Let $V \subseteq \{1, ..., T\}$ be the set of indices t such that the cube D_j that contains c_t includes at least a fraction $\sqrt{\varepsilon_1}$ of $c_1, ..., c_T$, i.e., $|\{s \leq T : c_s \in D_j\}| \geq T\sqrt{\varepsilon_1}$, where $\varepsilon_1 := \varepsilon/M$. Then $w_t(c_s) \geq 1/2$ for every such $c_s \in D_j$ (because $||c_t - c_s|| \leq 1/(2L)$ and so, by the Lipschitz condition, $1 - w_t(c_s) = w_t(c_t) - w_t(c_s) \leq L||c_t - c_s|| \leq 1/2$). Therefore $W_t \geq T\sqrt{\varepsilon_1}/2$ (there are at least $T\sqrt{\varepsilon}$ such c_s, and each one contributes at least $1/2$ to the sum

[22] We have not tried to optimize the estimate of ε'. The notations $f(x) = O(g(x))$, $f(x) = \Omega(g(x))$, and $f(x) = \Theta(g(x))$ mean, as usual, that there are constants $c, c' > 0$ such that for all x we have, respectively, $f(x) \leq cg(x)$, $f(x) \geq c'g(x)$, and $c'g(x) \leq f(x) \leq cg(x)$. In our case x stands for (ε, L); the dimension m is assumed fixed.

$\sum_s w_t(c_s) = W_t$), and so (24) yields

$$\left\|\bar{a}_t^{\Lambda} - c_t^{\Lambda}\right\| \leq T\varepsilon \frac{2}{T\sqrt{\varepsilon_1}} = 2\sqrt{M\varepsilon} \tag{25}$$

for each $t \in V$. If $t \notin V$ then c_t belongs to one of the cubes D_j that contains less than $T\sqrt{\varepsilon_1}$ of the $c_1, ..., c_T$, and so in total there are no more than $M \cdot T\sqrt{\varepsilon_1} = \sqrt{M\varepsilon}T$ indices $t \notin V$. For each one the bound $\left\|\bar{a}_t^{\Lambda} - c_t^{\Lambda}\right\| \leq \sqrt{m}$ (because both points belong to $C \subseteq [0,1]^m$) yields

$$\sum_{t \notin V} \left\|\bar{a}_t^{\Lambda} - c_t^{\Lambda}\right\| \leq \sqrt{m}\sqrt{M\varepsilon}T. \tag{26}$$

Adding (25) for all $t \in V$ together with (26) yields

$$\begin{aligned}\sum_{t=1}^{T} \left\|\bar{a}_t^{\Lambda} - c_t^{\Lambda}\right\| &\leq T \cdot 2\sqrt{M\varepsilon} + \sqrt{m}\sqrt{M\varepsilon}T = \sqrt{M\varepsilon}(2+\sqrt{m})T \\ &= \sqrt{\varepsilon L^m}(2\sqrt{m})^{m/2}(2+\sqrt{m})T \leq \varepsilon' T\end{aligned}$$

(because $M = (2L\sqrt{m})^m$), and so $K_T^{\Lambda} \leq \varepsilon'$, as claimed. □

The existence of smoothly calibrated procedures with the desired properties follows.

Proof of Theorems 1. Apply Theorem 10 and Proposition 11, and recall (Section 2.3) that for deterministic procedures leaks do not matter. □

Next, we show how to go from smooth to weak calibration.

Proposition 12 *An (ε, L)-smoothly calibrated procedure is (ε', L')-weakly calibrated, where*[23,24] $\varepsilon' = \Omega\left(\sqrt{\varepsilon L^m}\right)$ *and* $L' = O\left(\sqrt{\varepsilon L^{m+2}}\right)$.

Proof. Let $(D_j)_{j=1,...,M}$ be a partition of $[0,1]^m \supseteq C$ into disjoint cubes with side $\delta := 1/(L\sqrt{m})$; the diameter of each cube is thus $\delta\sqrt{m} = 1/L$, and the number of cubes is $M = \delta^{-m} = L^m m^{m/2}$. Let $\varepsilon_1 := \sqrt{\varepsilon L^m}$ and

[23] Again, we have not tried to optimize the estimates for ε' and L'.

[24] Thus, given ε' and L', one may take $L = \Theta(L'/\varepsilon')$ and $\varepsilon = \Theta((\varepsilon')^{m+2}/(L')^m)$.

$\varepsilon_2 := \varepsilon_1 m^{m/2}/M = \sqrt{\varepsilon/L^m}$. Take $L' := L\varepsilon_1/2 = \sqrt{\varepsilon L^{m+2}}/2$ and $\varepsilon' = \varepsilon_1(1+\sqrt{m}+\sqrt{m^{m+1}}) = \sqrt{\varepsilon L^m}(1+\sqrt{m}+\sqrt{m^{m+1}})$.

Fix $a_t, c_t \in C \subseteq [0,1]^m$ for $t = 1, ..., T$, and a weight function w in $W_{L'}$. Assume that $K_T^\Lambda \leq \varepsilon$ holds for every smoothing function Λ that is L-Lipschitz in the first coordinate; we will show that $S_T^w \leq \varepsilon'$ (where K_T^Λ and S_T^w are given by (1) and (18), respectively).

Let $V \subseteq \{1, ..., T\}$ be the set of indices t such that the cube D_j that contains c_t includes at least a fraction ε_2 of $c_1, ..., c_T$, i.e., $|\{s \leq T : c_s \in D_j\}| \geq \varepsilon_2 T$. Then

$$T - |V| = |\{t \leq T : t \notin V\}| < M \cdot \varepsilon_2 T = \varepsilon_2 MT, \tag{27}$$

because there are at most M cubes containing less than $\varepsilon_2 T$ points each.

We distinguish two cases.

Case 1: $\max_{t\in V} w(c_t) < \varepsilon_1$. Since $||a_t - c_t|| \leq \sqrt{m}$, we have $||\sum_{t\in V} w(c_t)(a_t - c_t)|| \leq |V| \cdot \varepsilon_1 \cdot \sqrt{m} \leq \varepsilon_1\sqrt{m}T$ (use $|V| \leq T$), and $||\sum_{t\notin V} w(c_t)(a_t - c_t)|| \leq (T - |V|) \cdot 1 \cdot \sqrt{m} = \varepsilon_2 M\sqrt{m}T$ (use (27)). Adding and dividing by T yields

$$S_T^w \leq (\varepsilon_1 + \varepsilon_2 M)\sqrt{m} = \varepsilon_1\sqrt{m}(1 + m^{m/2}) < K\varepsilon_1 = \varepsilon'.$$

Case 2: $\max_{t\in V} w(c_t) \geq \varepsilon_1$. Let $s \in V$ be such that $w(c_s) = \max_{t\in V} w(c_t) \geq \varepsilon_1$, and let $R \subseteq V$ be the set of indices r such that c_r lies in the same cube D_j as c_s.

For each r in R, proceed as follows. First, we have $|w(c_s) - w(c_r)| \leq L'||c_s - c_r|| \leq L' \cdot \delta\sqrt{m} = L\varepsilon_1/2 \cdot (1/L) = \varepsilon_1/2$, and so

$$w(c_r) \geq w(c_s) - \frac{\varepsilon_1}{2} \geq \varepsilon_1 - \frac{\varepsilon_1}{2} = \frac{\varepsilon_1}{2}. \tag{28}$$

Next, put $w^r(c) := \min\{w(c), w(c_r)\}$ and $\Lambda(c, c_r) := w^r(c)/w(c_r)$ for $r \in R$ (and, for $t \notin R$, put, say, $\Lambda(c, c_t) = 1$ for all c); then $\mathcal{L}(\Lambda(\cdot, c_r)) \leq \mathcal{L}(w)/w(c_r) \leq L'/(\varepsilon_1/2) = L$, and so, by our assumption

$$\frac{1}{T}\sum_{r\in R} ||\bar{a}_r^\Lambda - c_r^\Lambda|| \leq \frac{1}{T}\sum_{t\leq T} ||\bar{a}_t^\Lambda - c_t^\Lambda|| \leq K_T^\Lambda \leq \varepsilon. \tag{29}$$

We will now show that S_T^w is close to an appropriate multiple of $||\bar{a}_r^\Lambda - c_r^\Lambda||$, for each r in R. For t in V we have $w(c_t) \leq w(c_s)$, and so $0 \leq w(c_t) - w^r(c_t) \leq w(c_s) - w(c_r) \leq \varepsilon_1/2$ (recall (28)), which gives

$$\left\| \sum_{t \in V} \left(w(c_t) - w^r(c_t)\right)(a_t - c_t) \right\| \leq |V| \cdot \frac{\varepsilon_1}{2} \cdot \sqrt{m} \leq \frac{1}{2}\varepsilon_1 \sqrt{m} T.$$

For $t \notin V$ we have $0 \leq w(c_t) - w^r(c_t) \leq 1$, and so (recall (27))

$$\left\| \sum_{t \notin V} \left(w(c_t) - w^r(c_t)\right)(a_t - c_t) \right\| \leq (T - |V|) \cdot 1 \cdot \sqrt{m} \leq \varepsilon_2 M \sqrt{m} T.$$

Adding the two inequalities and dividing by T yields

$$\begin{aligned} S_T^w &= \left\| \frac{1}{T} \sum_{t=1}^{T} w(c_t)(a_t - c_t) \right\| \leq \left\| \frac{1}{T} \sum_{t=1}^{T} w^r(c_t)(a_t - c_t) \right\| + \sqrt{m}\left(\frac{\varepsilon_1}{2} + \varepsilon_2 M\right) \\ &\leq ||\bar{a}_r^\Lambda - c_r^\Lambda|| + \sqrt{m}\left(\frac{\varepsilon_1}{2} + \varepsilon_2 M\right), \end{aligned}$$

because $\sum_{t \leq T} w^r(c_t)(a_t - c_t) = \left(\sum_{t \leq T} w^r(c_t)\right)(\bar{a}_r^\Lambda - c_r^\Lambda)$ and $\sum_{t \leq T} w^r(c_t) \leq T$. Averaging over all r in the set R, whose size is $|R| \geq \varepsilon_2 T$ (all points in the same cube as c_s), and then recalling (29), finally gives

$$\begin{aligned} S_T^w &\leq \frac{1}{|R|} \sum_{r \in R} ||\bar{a}_r^\Lambda - c_r^\Lambda|| + \sqrt{m}\left(\frac{\varepsilon_1}{2} + \varepsilon_2 M\right) \\ &\leq \frac{1}{\varepsilon_2} \frac{1}{T} \sum_{r \in R} ||\bar{a}_r^\Lambda - c_r^\Lambda|| + \sqrt{m}\left(\frac{\varepsilon_1}{2} + \varepsilon_2 M\right) \\ &\leq \frac{1}{\varepsilon_2}\varepsilon + \sqrt{m}\left(\frac{\varepsilon_1}{2} + \varepsilon_2 M\right) = \frac{\varepsilon \sqrt{L^m}}{\sqrt{\varepsilon}} + \varepsilon_1\left(\frac{\sqrt{m}}{2} + m^{(m+1)/2}\right) \\ &= \varepsilon_1\left(1 + \frac{\sqrt{m}}{2} + m^{(m+1)/2}\right) < K\varepsilon_1 = \varepsilon', \end{aligned}$$

completing the proof. □

6 Nash Equilibrium Dynamics

In this section we use our results on smooth calibration to obtain dynamics in n-person games that are in the long run close to Nash equilibria most of the time.

A (finite) *game* is given by a finite set of players N, and, for each player $i \in N$, a finite set of actions[25] A^i and a payoff function $u^i : A \to \mathbb{R}$, where $A := \prod_{i\in N} A^i$ denotes the set of action combinations of all players. The set of mixed actions of player i is $X^i := \Delta(A^i)$, which is the unit simplex (i.e., the set of probability distributions) on A^i; we identify the pure actions in A^i with the unit vectors of X^i, and so $A^i \subseteq X^i$. Put $X := \prod_{i\in N} X^i$ for the set of mixed-action combinations. Let $C := \Delta(A)$ be the set of joint distributions on action combinations; thus $A \subseteq X \subseteq C \subseteq [0,1]^m$, where $m = \prod_{i\in N} |A^i|$. The payoff functions u^i are linearly extended to C, and thus $u^i : C \to \mathbb{R}$.

For each player i, a combination of mixed actions of the other players $x^{-i} = (x^j)_{j\neq i} \in \prod_{j\neq i} X^j =: X^{-i}$, and $\varepsilon \geq 0$, let $\mathrm{BR}^i_\varepsilon(x^{-i}) := \{x^i \in X^i : u^i(x^i, x^{-i}) \geq \max_{y^i\in X^i} u^i(y^i, x^{-i}) - \varepsilon\}$ denote the set of *ε-best replies* of i to x^{-i}. A (mixed) action combination $x \in X$ is a *Nash ε-equilibrium* if $x^i \in \mathrm{BR}^i_\varepsilon(x^{-i})$ for every $i \in N$; let $\mathrm{NE}(\varepsilon) \subseteq X$ denote the set of Nash ε-equilibria of the game.

A (discrete-time) *dynamic* consists of each player $i \in N$ playing a pure action $a^i_t \in A^i$ at each time period $t = 1, 2, ...$; put $a_t = (a^i_t)_{i\in N} \in A$. There is perfect monitoring: at the end of period t all players observe a_t. The dynamic is *uncoupled* (Hart and Mas-Colell 2003, 2006, 2013) if the play of every player i may depend only on player i's payoff function u^i (and not on the other players' payoff functions). Formally, such a dynamic is given by a mapping for each player i from the history $h_{t-1} = (a_1, ..., a_{t-1})$ and his own payoff function u^i into $X^i = \Delta(A^i)$ (player i's choice may be random); we will call such mappings *uncoupled.* Let $x^i_t \in X^i$ denote the mixed action that player i plays at time t, and put $x_t = (x^i_t)_{i\in N} \in X$.

The dynamics we consider are continuous, smooth variants of the "cali-

[25]We refer to one-shot choices as "actions" rather than "strategies," the latter term being used for repeated interactions.

brated learning" introduced by Foster and Vohra (1997). *Calibrated learning* consists of each player best-replying to calibrated forecasts on the other players' actions; it results in the joint distribution of play converging in the long run to the set of correlated equilibria of the game. Kakade and Foster (2004) defined *publicly calibrated learning*, where each player approximately best-replies to a public weakly calibrated forecast on all players' actions, and proved that most of the time the play is an approximate Nash equilibrium. We consider instead *smooth calibrated learning*, where weak calibration is replaced with the more natural smooth calibration; it amounts to taking calibrated learning and smoothing out both the forecasts and the best replies.

Formally, a *smooth calibrated learning* dynamic is given by:

(D1) An (ε_c, L_c)-smoothly calibrated deterministic procedure, which yields at time t a "forecast" $c_t \in C$.

(D2) An L_g-Lipschitz ε_g-approximate best-reply mapping $g : C \to \prod_{i\in N} \Delta(A^i)$; i.e., $g^i(c) \in \mathrm{BR}^i_{\varepsilon_g}(c^{-i})$ for every $i \in N$ and[26] $c^{-i} \in \Delta(A^{-i})$, $g(c) = (g^i(c))_{i\in N}$, and $\mathcal{L}(g) \le L_g$.

(D3) Each player runs the procedure in (D1), generating at time t a forecast $c_t \in C$; then each player i plays at period t the mixed action[27] $x^i_t := g^i(c_t) \in \Delta(A^i)$, where g^i is given by (D2). All players observe the action combination $a_t = (a^i_t)_{i\in N} \in A$ that has actually been played, and remember it together with the forecast[28] $c_t \in C$.

Smooth calibrated learning is a stochastic uncoupled dynamic: stochastic because the players use mixed actions, and uncoupled because the payoff

[26]For a joint distribution $c \in C = \Delta(A)$ on action combinations and a player i, we denote by $c^{-i} \in \Delta(A^{-i})$ the marginal of c on A^{-i} (when $x = (x^i)_{i\in N} \in \prod_{i\in N} \Delta(A^i)$ is a product distribution then $x^{-i} = (x^j)_{j\neq i} \in \prod_{j\neq i} \Delta(A^j)$). The ε-best-reply correspondence $\mathrm{BR}^i_\varepsilon$ is defined also for $c^{-i} \in \Delta(A^{-i})$, i.e., $\mathrm{BR}^i_\varepsilon(c^{-i}) := \{x^i \in \Delta(A^i) : u^i(x^i, c^{-i}) \ge \max_{y^i\in\Delta(A^i)} u^i(y^i, c^{-i}) - \varepsilon\}$, where (y^i, c^{-i}) denotes the product of the two distributions, y^i (on A^i) and c^{-i} (on A^{-i}).

[27]Thus $\mathbb{P}[a_t = a \mid h_{t-1}] = \prod_{i\in N} x^i_t(a^i)$ for every $a = (a^i)_{i\in N} \in A$, where h_{t-1} is the history and $x^i_t(a^i)$ is the probability that $x^i_t \in \Delta(A^i)$ assigns to the pure action $a^i \in A^i$.

[28]As we will see below, it suffices to remember the actions and forecasts of the last R periods only (for an appropriate finite R).

function of player i is used by player i only, in constructing his approximate best-reply mapping g^i.

Our result is:

Theorem 13 *Fix the finite set of players N, the finite action spaces A^i for all $i \in N$, and the payoff bound $U < \infty$. For every $\varepsilon > 0$ there exist stochastic uncoupled dynamics—e.g., smooth calibrated learning—that have finite memory and are stationary, such that*

$$\liminf_{T \to \infty} \frac{1}{T} |\{t \leq T : x_t \in \mathrm{NE}(\varepsilon)\}| \geq 1 - \varepsilon \quad \text{(a.s.)}$$

for every finite game with payoff functions $(u^i)_{i \in N}$ that are bounded by U (i.e., $|u^i(a)| \leq U$ for all $i \in N$ and $a \in A$).

The idea of the proof is as follows. First, assume that the forecasts c_t are in fact calibrated (rather than just smoothly calibrated) and, moreover, that they are calibrated with respect to the mixed plays x_t (rather than with respect to the actual plays a_t). Because x_t is given by a fixed function of c_t, namely, $x_t = g(c_t)$, the sequence of mixed plays in those periods when the forecast was a certain c is the constant sequence $g(c), ..., g(c)$, whose average is $g(c)$, and calibration then implies that $g(c)$ must be close to c (most of the time, i.e., for forecasts that appear with positive frequency). But we have only smooth calibration; however, because g is a Lipschitz function, if c and $g(c)$ are far from one another then so are c' and $g(c')$ for any c' close to c, and so the average of such $g(c')$ is also far from c, contradicting smooth calibration. Thus, most of the time $g(c_t)$ is close to c_t, and hence $g(g(c_t))$ is close to $g(c_t)$ (because g is Lipschitz)—which says that $g(c_t)$ is close to an approximate best reply to itself, i.e., $g(c_t)$ is an approximate Nash equilibrium. Finally, an appropriate use of a strong law of large numbers shows that if the actual plays a_t are (smoothly) calibrated then so are their expectations, i.e., the mixed plays x_t.

Proof. This proof goes along similar lines to the proof of Kakade and Foster

(2004) for publicly calibrated dynamics (which is the only other calibration-based Nash dynamic to date[29]).

The existence of a deterministic smoothly calibrated procedure is given by Theorem 1, and that of a Lipschitz approximate best-reply mapping is given by Lemma 17 in the Appendix (the function u^i is linear in c^i, and $|u^i(a)| \leq U$ for all $a \in A$ implies that $\mathcal{L}(u^i) \leq 2\sqrt{m}U =: L_u$). For each period t, let $c_t \in C$ be the forecast, $x_t = g(c_t) \in \prod_{i \in N} \Delta(A^i)$ the mixed actions, and $a_t \in A$ the realized pure actions (c_t, x_t, and a_t all depend on the history).

Let W be the set of weight functions $w : C \to [0,1]$ with $\mathcal{L}(w) \leq L_c$, and let $w_1, ..., w_K \in W$ be an[30] ε_1-net of W. From[31] $\mathbb{E}[a_t \mid h_{t-1}] = x_t = g(c_t)$ it follows that $\mathbb{E}[w_k(c_t)a_t \mid h_{t-1}] = w_k(c_t)g(c_t)$ for every $k = 1, ..., K$, and hence, by the Strong Law of Large Numbers for Dependent Random Variables (see Loève 1978, Theorem 32.1.E: $(1/T)\sum_{t=1}^{T}(X_t - \mathbb{E}[X_t|h_{t-1}]) \to 0$ as $T \to \infty$ a.s., for random variables X_t that are, in particular, uniformly bounded; note that there are finitely many k):

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} w_k(c_t)(a_t - g(c_t)) = 0 \quad \text{for all } 1 \leq k \leq K \quad \text{(a.s.)}. \tag{30}$$

Thus, for each one of the (almost all) infinite histories h_∞ where (30) holds, there is a finite T_1 (depending on h_∞) such that $\left\| \sum_{t=1}^{T} w_k(c_t)(a_t - g(c_t)) \right\| \leq \varepsilon_1 T$ for all $T > T_1$ and all $1 \leq k \leq K$; since the w_k constitute an ε_1-net of W and $||c|| \leq \sqrt{m}$ for every $c \in C$, it follows that

$$\left\| \sum_{t=1}^{T} w(c_t)(a_t - g(c_t)) \right\| \leq (1 + 2\sqrt{m})\varepsilon_1 T \quad \text{for all } T > T_1 \text{ and all } w \in W.$$

[29] Recall footnote 3.

[30] The relations between the various ε and L appearing in the proof will be given at the end of the proof.

[31] Recall that h_{t-1} denotes the history, and we identify the pure actions $a^i \in A^i$ with the unit vectors in C^i.

Take $\Lambda(x,c) = [1 - L_c||x-c||]_+$, then $\Lambda(\cdot, c) \in W$, and so, for every $t \leq T$,

$$\left\|\sum_{s=1}^{T} \lambda_s(a_s - g(c_s))\right\| \leq (1+2\sqrt{m})\varepsilon_1 T, \tag{31}$$

where $\lambda_s := \Lambda(c_s, c_t)$. Because $\lambda_s > 0$ only when $||c_s - c_t|| < 1/L_c$, and thus $||g(c_s) - g(c_t)|| < L_g/L_c$, it follows that

$$\left\|\sum_{s=1}^{T} \lambda_s(a_s - g(c_t))\right\| \leq (1+2\sqrt{m})\varepsilon_1 T + \sum_{s=1}^{T} \lambda_s \frac{L_g}{L_c}.$$

Recalling that $\sum_s \lambda_s a_s = \left(\sum_s \lambda_s\right) \bar{a}_t^{\Lambda}$ then yields

$$||\bar{a}_t^{\Lambda} - g(c_t)|| \leq \frac{T}{\sum_{s=1}^{T} \lambda_s}(1+2\sqrt{m})\varepsilon_1 + \frac{L_g}{L_c}. \tag{32}$$

As in the Proof of Proposition 11, let $(D_j)_{j=1,\ldots,M}$ be a partition of $[0,1]^m \supseteq C$ into disjoint cubes with side $1/(2L_c\sqrt{m})$; the diameter of each cube is thus $1/(2L_c)$, and the number of cubes is $M = (2L_c\sqrt{m})^m$. For each infinite history h_∞ where (30) holds and every $T > T_1$, let $V \subseteq \{1, 2, ..., T\}$ be the set of indices t such that the cube D_j to which c_t belongs contains at least $\varepsilon_2 T$ of $c_1, ..., c_T$. If c_s and c_t belong to the same cube D_j then $||c_s - c_t|| \leq \operatorname{diam}(D_j) = 1/(2L_c)$, and so $\lambda_s = [1 - L_c||c_s - c_t||]_+ \geq 1/2$. Therefore $t \in V$ implies that $\sum_{s=1}^{T} \lambda_s \geq \varepsilon_2 T/2$, which, using (32), yields

$$||\bar{a}_t^{\Lambda} - g(c_t)|| \leq \frac{2}{\varepsilon_2}(1+2\sqrt{m})\varepsilon_1 + \frac{L_g}{L_c}. \tag{33}$$

If $t \notin V$ then c_t belongs to a cube D_j that contains less than $\varepsilon_2 T$ of $c_1, ..., c_T$, and so there are at most $M \cdot \varepsilon_2 T$ such t (the number of cubes is M). Using $||\bar{a}_t^{\Lambda} - g(c_t)|| \leq 2\sqrt{m}$ for $t \notin V$ and (33) for $t \in V$, we finally get

$$\frac{1}{T}\sum_{t=1}^{T} ||\bar{a}_t^{\Lambda} - g(c_t)|| \leq \frac{2}{\varepsilon_2}(1+2\sqrt{m})\varepsilon_1 + \frac{L_g}{L_c} + 2\sqrt{m}M\varepsilon_2. \tag{34}$$

Let T_0 be such that the smooth calibration score $K_T^{\Lambda} = (1/T)\sum_{t \leq T} ||\bar{a}_t^{\Lambda} -$

$c_t^\Lambda|| \leq \varepsilon_c$ for any $T > T_0$. Because, again, $\lambda_s > 0$ only when $||c_s - c_t|| < 1/L_c$, and c_t^Λ is a weighted average of such c_s, it follows that $||c_t^\Lambda - c_t|| < 1/L_c$, and so

$$\frac{1}{T}\sum_{t=1}^{T} ||\bar{a}_t^\Lambda - c_t|| \leq \varepsilon_c + \frac{1}{L_c}$$

for every $T > T_0$. Adding this to (34) yields

$$\frac{1}{T}\sum_{t=1}^{T} \|g(c_t) - c_t\| \leq \varepsilon_3 \tag{35}$$

for almost every infinite history and $T > \max\{T_0, T_1\}$, where

$$\varepsilon_3 := \varepsilon_c + 2(1 + 2\sqrt{m})\varepsilon_1/\varepsilon_2 + (L_g + 1)/L_c + 2\sqrt{m}M\varepsilon_2. \tag{36}$$

From (35) it immediately follows that, for every $\varepsilon_4 > 0$,

$$\frac{1}{T}\left|\{t \leq T : \|g(c_t) - c_t\| > \varepsilon_4\}\right| \leq \frac{1}{\varepsilon_4}\frac{1}{T}\sum_{t=1}^{T} \|(g(c_t) - c_t)\| \leq \frac{\varepsilon_3}{\varepsilon_4}. \tag{37}$$

If $\|g(c_t) - c_t\| \leq \varepsilon_4$ then $x_t = g(c_t)$ satisfies

$$\|g(x_t) - x_t\| = \|g(g(c_t)) - g(c_t)\| \leq L_g \|g(c_t) - c_t\| \leq L_g\varepsilon_4,$$

and so

$$\begin{aligned} u^i(x_t) &\geq u^i(g^i(x_t), x_t^{-i}) - L_u \left\|g^i(x_t) - x_t^i\right\| \\ &\geq \max_{y^i \in \Delta(A^i)} u^i(y^i, x_t^{-i}) - \varepsilon_g - L_u L_g \varepsilon_4 \end{aligned}$$

(for the second inequality we have used $g^i(x) \in \mathrm{BR}^i_{\varepsilon_g}(x^{-i})$). Therefore $\|g(c_t) - c_t\| \leq \varepsilon_4$ implies that $x_t \in \mathrm{NE}(\varepsilon_5)$, where

$$\varepsilon_5 := \varepsilon_g + L_u L_g \varepsilon_4, \tag{38}$$

and so, from (35) and (37) we get

$$\frac{1}{T}\left|\{t \leq T : x_t \notin \mathrm{NE}(\varepsilon_5)\}\right| \leq \frac{\varepsilon_3}{\varepsilon_4}$$

for all large enough T, for almost every infinite history.

To make both $\varepsilon_3/\varepsilon_4$ and ε_5 equal to ε one may take, for instance (see (36) and (38)),

$$\begin{aligned}
\varepsilon_g &= \frac{\varepsilon}{2}, \; L_g = O\left(\left(\frac{2L_u}{\varepsilon}\right)^{m+1}\right) \\
\varepsilon_4 &= \frac{\varepsilon}{2L_uL_g}, \; \varepsilon_3 = \varepsilon_4\varepsilon = \frac{\varepsilon^2}{2L_uL_g}, \\
\varepsilon_c &= \frac{\varepsilon_3}{4} = \frac{\varepsilon^2}{8L_uL_g}, \quad L_c = \frac{4(L_g+1)}{\varepsilon_3} = \frac{32L_uL_g(L_g+1)}{\varepsilon^2}, \\
\varepsilon_2 &= \frac{\varepsilon_3}{8\sqrt{m}M}, \; \varepsilon_1 = \frac{(\varepsilon_3)^2}{64\sqrt{m}(1+2\sqrt{m})M}
\end{aligned}$$

(recall that $L_u = 2\sqrt{m}U$ and $M = (2\sqrt{m}L_c)^m$).

Finally, the smoothly calibrated procedure that all the players use has finite recall and is stationary. However, while in the calibration game of Sections 4 and 5 both c_t and a_t are monitored and thus become part of the recall window, in the n-person game only a_t is monitored (while the forecast c_t is computed by each player separately, but is *not* played). Therefore, in order to run the calibrated procedure, in the n-person game each player needs to remember at time T, in addition to the last R action combinations $a_{T-R}, ..., a_{T-1}$, also the last R forecasts $c_{T-R}, ..., c_{T-1}$. "Finite recall" of size R in the calibration procedure therefore becomes "finite memory" of size $2R$ in the game dynamic (the memory contains $2R$ elements of[32] C). □

Remarks. *(a) Nash dynamics.* Uncoupled dynamics where Nash ε-equilibria are played $1-\varepsilon$ of the time were first proposed by Foster and Young (2003), followed by Kakade and Foster (2004), Foster and Young (2006), Hart and Mas-Colell (2006), Germano and Lugosi (2007), Young (2009),

[32]For a similar transition from finite recall to finite memory, see Theorem 7 in Hart and Mas-Colell (2006).

Babichenko (2012), and others (see also Remark (f) below).

(b) Coordination. All players need to coordinate before playing the game on the smoothly (or weakly) calibrated procedure that they will run; thus, at every period t they all generate the same forecast t. By contrast, in the original calibrated learning of Foster and Vohra (1997)—which leads to correlated equilibria—every player may use his own calibrated procedure.

This fits the so-called *Conservation Coordination Law* for game dynamics, which says that some form of "coordination" must be present, either in the limit static equilibrium concept (such as correlated equilibrium), or else in the dynamic leading to it (as in the case of Nash equilibrium). See Hart and Mas-Colell (2003, footnote 19) and Hart (2005, footnote 19).

(c) Deterministic calibration. In order for all the players to generate the same forecasts, it is not enough that they all use the same procedure; in addition, the forecasts must be deterministic (otherwise the randomizations, which are carried out independently by the players, may lead to different actual forecasts). This is the reason that we use smoothly calibrated procedures rather than fully calibrated ones (cf. Oakes 1985 and Foster and Vohra 1998).

(d) Leaky calibration. One may use a common *randomized* smoothly calibrated procedure, provided that the randomizations are carried out publicly (i.e., they must be leaked!). Alternatively, a "central bureau of statistics" may provide each period the forecast for all players.

(e) Forecasting almost independent play. The proof of Theorem 13 above shows that most of the time the forecasts c_t are close to being independent across players, i.e., close to X; indeed, $g(c_t) \in X$ and c_t is close to $g(c_t)$.

(f) Exhaustive search. Dynamics that perform exhaustive search can also be used to get the result of Theorem[33] 13. Take for instance a finite grid on C, say, $D = \{d_1, ..., d_M\} \subseteq C$, that is fine enough so that there always is a pure Nash ε-equilibrium on the grid. Let the dynamic go over the points $d_1, d_2, ...$ in sequence until the first time that $d_T^i \in \mathrm{BR}_\varepsilon^i(d_T^{-i})$ for all i, following which d_T is played forever. This is implemented by having for every player i a distinct action $a_0^i \in A^i$ that is played at time t only when $d_t^i \in \mathrm{BR}_\varepsilon^i(d_t^{-i})$ (otherwise

[33] We thank Yakov Babichenko for suggesting this.

a different action is played); once the action combination $a_0 = (a_0^i)_{i\in N} \in A$ is played, say, at time T, each player i plays d_T^i at all $t > T$. This dynamic is uncoupled (each player only considers $\mathrm{BR}_\varepsilon^i$) and has memory of size 2 (i.e., 2 elements of C): for $t \leq T$, it consists of d_{t-1} and a_{t-1} (the last checked point and the last played action combination); for $t > T$, it consists of d_T and a_0. Of course, all players need to coordinate before playing the game on the sequence $d_1, d_2, ..., d_M$ and the action combination a_0.

(g) Continuous action spaces. The result of Theorem 13 easily extends to continuous action spaces and approximate *pure* Nash equilibria. Assume that for each player $i \in N$ the set of actions $A^i = C^i$ is a convex compact subset of some Euclidean space (such games arise, for instance, from exchange economies where the actions are net trades; see, e.g., Hart and Mas-Colell 2015). Thus $A = C = \prod_{i\in N} C^i$ is a compact convex set in some Euclidean space, say, $\mathbb{R}^m$.

For every $\varepsilon \geq 0$, the set of *pure ε-best replies* of player i to $c^{-i} \in C^{-i}$ is $\mathrm{PBR}_\varepsilon^i(c^{-i}) := \{c^i \in C^i : u^i(c^i, c^{-i}) \geq \max_{b^i \in C^i} u^i(b^i, c^{-i}) - \varepsilon\}$. An action combination $a \in C$ is a *pure Nash ε-equilibrium* if $a^i \in \mathrm{PBR}_\varepsilon^i(a^{-i})$ for every $i \in N$; let $\mathrm{PNE}(\varepsilon) \subseteq C$ denote the set of pure Nash ε-equilibria.

Smooth calibrated learning is defined as above, except that now the approximate best replies are pure actions (the play is $a_t = g(c_t)$, and it is monitored by all players). Our result here is:

Theorem 14 *Fix the finite set of players N, the convex compact action spaces A^i for all $i \in N$, and the Lipschitz bound $L < \infty$. For every $\varepsilon > 0$ there exist stochastic uncoupled dynamics—e.g., smooth calibrated learning—that have finite memory and are stationary, and a $T_0 \equiv T_0(\varepsilon, L)$ such that for every $T \geq T_0$,*

$$\frac{1}{T}\left|\{t \leq T : a_t \in \mathrm{PNE}(\varepsilon)\}\right| \geq 1 - \varepsilon$$

for every game with payoff functions $(u^i)_{i\in N}$ that are L-Lipschitz (i.e., $\mathcal{L}(u^i) \leq L$) and quasi-concave in one's own action (i.e., $u^i(c^i, c^{-i})$ is quasi-concave in $c^i \in C^i$ for every $c^{-i} \in C^{-i}$), for all $i \in N$.

Proof. We now have $A^i = C^i$ and $a_t = g(c_t)$, and everything is deterministic. Proceed as in the proof of Theorem 13, skipping the use of the Law of Large Numbers and taking $\varepsilon_1 = 0$ in (31) (and $T_1 = 0$). □

(h) Additional player. To get finite recall rather than finite memory one may add an artificial player, say, player 0, with action set $A^0 := C$ and constant payoff function $u^0 \equiv 0$, who plays at each period t the forecast c_t, i.e., $a_t^0 = c_t$; this way the forecasts become part of the recall of all players.

(i) Reaction function and fixed points. The Proof of Theorem 13 shows that in the leaky calibration game, if the A-player uses a stationary strategy given by a Lipschitz "reaction" function g (i.e., he plays $g(c_t)$ at time t), then smooth calibration implies that the forecasts c_t are close to fixed points of g most of the time.

References

Azoury, K. S. and M. K. Warmuth (2001), "Relative Loss Bounds for On-Line Density Estimation with the Exponential Family of Distributions," *Machine Learning* 43, 211–246.

Babichenko, Y. (2012), "Completely Uncoupled Dynamics and Nash Equilibria," *Games and Economic Behavior* 76, 1–14.

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

Dawid, A. (1982), "The Well-Calibrated Bayesian," *Journal of the American Statistical Association* 77, 605–613.

Foster, J. (1999), "On Relative Loss Bounds in Generalized Linear Regression," in *12th International Symposium on Fundamentals of Computation Theory (FCT '99)*, 269–280.

Foster, D. P. (1991), "Prediction in the Worst Case," *The Annals of Statistics* 19, 1084–1090.

Foster, D. P. (1999), "A Proof of Calibration via Blackwell's Approachability Theorem," *Games and Economic Behavior* 29, 73–78.

Foster, D. P. and S. M. Kakade (2006), "Calibration via Regression," *IEEE Information Theory Workshop 2006.*

Foster, D. P. and R. V. Vohra (1997), "Calibrated Learning and Correlated Equilibrium," *Games and Economic Behavior* 21, 40–55.

Foster, D. P. and R. V. Vohra (1998), "Asymptotic Calibration," *Biometrika* 85, 379–390.

Foster, D. P. and H. P. Young (2003), "Learning, Hypothesis Testing, and Nash Equilibrium," *Games and Economic Behavior* 45, 73–96.

Foster, D. P. and H. P. Young (2006), "Regret Testing: Learning to Play Nash Equilibrium Without Knowing You Have an Opponent," *Theoretical Economics* 1, 341–367.

Germano, F. and G. Lugosi (2007), "Global Nash Convergence of Foster and Young's Regret Testing," *Games and Economic Behavior* 60, 135–154.

Hart, S. (2005), "Adaptive Heuristics," *Econometrica* 73, 1401–1430; Chapter 11 of Hart and Mas-Colell (2013).

Hart, S. and A. Mas-Colell (2000), "A Simple Adaptive Procedure Leading to Correlated Equilibrium," *Econometrica* 68, 1127–1150; Chapter 2 of Hart and Mas-Colell (2013).

Hart, S. and A. Mas-Colell (2001), "A General Class of Adaptive Strategies," *Journal of Economic Theory* 98, 26–54; Chapter 3 of Hart and Mas-Colell (2013).

Hart, S. and A. Mas-Colell (2003), "Uncoupled Dynamics Do Not Lead to Nash Equilibrium," *American Economic Review* 93, 1830–1836; Chapter 7 of Hart and Mas-Colell (2013).

Hart, S. and A. Mas-Colell (2006), "Stochastic Uncoupled Dynamics and Nash Equilibrium," *Games and Economic Behavior* 57, 286–303; Chapter 8 of Hart and Mas-Colell (2013).

Hart, S. and A. Mas-Colell (2013), *Simple Adaptive Strategies: From Regret-Matching to Uncoupled Dynamics*, World Scientific, 2013.

Hart, S. and A. Mas-Colell (2015), "Markets, Correlation, and Regret-Matching," *Games and Economic Behavior* 93, 42–54.

Hazan, E. and S. Kakade (2012), "(Weak) Calibration is Computationally Hard," *25th Annual Conference on Learning Theory (COLT '12); JMLR: Workshop and Conference Proceedings* 23 (2012), 3.1–3.10.

Kakade, S. M. and D. P. Foster (2004), "Deterministic Calibration and Nash Equilibrium," in *17th Annual Conference on Learning Theory (COLT '04)*; *Journal of Computer and System Sciences* 74 (2008), 115–130.

Lehrer, E. and E. Solan (2009), "Approachability with Bounded Memory," *Games and Economic Behavior* 66, 995–1004.

Loève, M. (1978), *Probability Theory, Vol. II*, 4th edition, Springer.

Oakes, D. (1985), "Self-calibrating Priors Do Not Exist," *Journal of the American Statistical Association* 80, 339.

Olszewski, W. (2015), "Calibration and Expert Testing," in *Handbook of Game Theory, Vol. 4*, H. P. Young and S. Zamir (editors), Springer, 949–984.

Papadimitriou, C. (1994). "On the Complexity of the Parity Argument and Other Inefficient Proofs of Existence," *Journal of Computer and System Sciences* 48, 498–532.

Vovk, V. (2001), "Competitive On-Line Statistics," *International Statistical Review* 69, 213–248.

Young, H. P. (2009), "Learning by Trial and Error," *Games and Economic Behavior* 65, 626–643.

A Appendix

Let C be a compact subset of $\mathbb{R}^m$ and let $\varepsilon > 0$. A *maximal* 2ε*-net* in C is a maximal collection of points $z_1, ..., z_K \in C$ such that $\|z_k - z_j\| \geq 2\varepsilon$ for all $k \neq j$; maximality implies $\cup_{k=1}^K B(z_k, 2\varepsilon) \supseteq C$. Let $\alpha_k(x) := [3\varepsilon - \|x - z_k\|]_+$, and put $\bar{\alpha}(x) := \sum_{k=1}^K \alpha_k(x)$. For every $x \in C$ we have $0 \leq \alpha_k(x) \leq 3\varepsilon$ and $\bar{\alpha}(x) \geq \varepsilon$ (since $\alpha_k(x) \geq \varepsilon$ when $x \in B(z_k, 2\varepsilon)$, and the union of these balls covers C). Finally, define $\beta_k(x) := \alpha_k(x)/\bar{\alpha}(x)$.

Lemma 15 *The functions* $(\beta_k)_{1 \leq k \leq K}$ *satisfy the following properties:*

(i) $\beta_k(x) \geq 0$ *for all* $x \in C$ *and all* k.

(ii) $\sum_{k=1}^K \beta_k(x) = 1$ *for all* $x \in C$.

(iii) $\beta_k(x) = 0$ *for all* $x \notin B(z_k, 3\varepsilon)$.

(iv) *For each* $x \in C$ *there are at most*[34] 4^m *indices* k *such that* $\beta_k(x) > 0$.

(v) $\mathcal{L}(\beta_k) \leq 4^{m+2}/\varepsilon$ *for every* k.

Proof. (i) and (ii) are immediate. For (iii), we have $\beta_k(x) > 0$ iff $\alpha_k(x) > 0$ iff $\|x - z_k\| < 3\varepsilon$. This implies that $B(z_k, \varepsilon) \subseteq B(x, 4\varepsilon)$. The open balls of radius ε with centers at z_k are disjoint (because $\|z_k - z_j\| \geq 2\varepsilon$ for $k \neq j$), and so there can be at most 4^m such balls included in $B(x, 4\varepsilon)$ whose volume is 4^m times larger; this proves (iv). For every $x, y \in C$:

$$\begin{aligned}
|\beta_k(x) - \beta_k(y)| &\leq \left|\frac{\alpha_k(x)}{\bar{\alpha}(x)} - \frac{\alpha_k(y)}{\bar{\alpha}(x)}\right| + \left|\frac{\alpha_k(y)}{\bar{\alpha}(x)} - \frac{\alpha_k(y)}{\bar{\alpha}(y)}\right| \\
&\leq \frac{1}{\bar{\alpha}(x)} |\alpha_k(x) - \alpha_k(y)| + \frac{\alpha_k(y)}{\bar{\alpha}(x)\bar{\alpha}(y)} \sum_{j=1}^{K} |\alpha_j(x) - \alpha_j(y)| \\
&\leq \frac{1}{\varepsilon} \|x - y\| + \frac{3\varepsilon}{\varepsilon \cdot \varepsilon} 2 \cdot 4^m \|x - y\| \leq \frac{4^{m+2}}{\varepsilon} \|x - y\|
\end{aligned}$$

(since $\bar{\alpha}(x) \geq \varepsilon$, $\alpha_k(x) \leq 3\varepsilon$, and there are at most $2 \cdot 4^m$ indices j where neither $\alpha_j(x)$ nor $\alpha_j(y)$ vanish); this proves (v). □

Thus, the functions $(\beta_k)_{1 \leq k \leq K}$ constitute a *Lipschitz partition of unity* that is subordinate to the maximal 2ε-net $z_1, ..., z_K$. Next, we obtain a basis for the Lipschitz functions on C.

Lemma 16 *Let* W_L *be the set of functions* $w : C \to [0, 1]$ *with* $\mathcal{L}(w) \leq L$. *Then for every* $\varepsilon > 0$ *there exist* d *functions* $f_1, ..., f_d \in W_L$ *such that for every* $w \in W_L$ *there is a vector* $\varpi \equiv \varpi_w \in [0, 1]^d$ *satisfying*

$$\max_{x \in C} \left| w(x) - \sum_{i=1}^{d} \varpi_i f_i(x) \right| < \varepsilon.$$

Moreover, one can take $d = O(L^m / \varepsilon^{m+1})$.

[34] We have not tried to get the best bounds in (iv) and (v); indeed, they may be easily reduced.

Proof. Put $\varepsilon_1 := \varepsilon/(3L)$. Let $z_1, ..., z_K$ be a maximal $2\varepsilon_1$-net on C, and let $\beta_1, ..., \beta_K$ be the corresponding Lipschitz partition of unity given by Lemma 15 (for ε_1).

Given $w \in W_L$, let $v(x) := \sum_{k=1}^{N} w(z_k)\beta_k(x)$; then $w(z_k) \in [0,1]$ and we have

$$\begin{aligned} |w(x) - v(x)| &= \left|\sum_{k=1}^{N} (w(x) - w(z_k))\,\beta_k(x)\right| \leq \sum_{k:\beta_k(x)>0} \beta_k(x)\,|w(x) - w(z_k)| \\ &\leq \sum_{k:\beta_k(x)>0} \beta_k(x) 3\varepsilon_1 L = 3\varepsilon_1 L, \end{aligned}$$

since $\beta_k(x) > 0$ implies $\|x - z_k\| < 3\varepsilon_1$ and thus $|w(x) - w(z_k)| \leq L\,\|x - z_k\| \leq L \cdot 3\varepsilon_1$ (because $\mathcal{L}(w) \leq L$).

Now $\mathcal{L}(\beta_k) \leq 4^{m+2}/\varepsilon_1$ by (v) of Lemma 15; we thus replace each β_k by the sum of $Q = \lceil 4^{m+2}/(\varepsilon_1 L)\rceil$ identical copies of $(1/Q)\beta_k$—denote them $f_{k,1}, ..., f_{k,Q}$—which thus satisfy $\mathcal{L}(f_{k,q}) = (1/Q)\mathcal{L}(\beta_k) \leq L$, and so

$$\left| w(x) - \sum_{k=1}^{K}\sum_{q=1}^{Q} w(z_k) f_{k,q}(x) \right| = |w(x) - v(x)| \leq 3\varepsilon_1 L = \varepsilon.$$

The $d = KQ$ functions $(f_{k,q})_{1\leq k\leq K, 1\leq q\leq Q}$ yield our result.

Finally, $K = O(\varepsilon_1^{-m})$ (because C contains the K disjoint open balls of radius ε_1 centered at the z_k) and $Q \leq 4^{m+2}/(\varepsilon_1 L) + 1$, and so $d = KQ = O(\varepsilon_1^{-m-1} L^{-1}) = O(\varepsilon^{-m-1} L^m)$. □

In the game setup we construct ε-best reply functions that are Lipschitz. The following lemma applies when the action spaces are finite (as in Theorem 13), and also when they are continuous (as in Theorem 14). In the latter case there is no need to use mixed actions, and so $C^i = A^i$ and $C = A = \prod_{i\in N} A^i$, and the sets $\Delta(A^i)$, $\Delta(A^{-i})$, and $\Delta(A)$ are identified with A^i, A^{-i}, and A (or, C^i, C^{-i}, and C), respectively; also, $\mathrm{BR}^i_\varepsilon$ stands for $\mathrm{PBR}^i_\varepsilon$, the set of *pure* ε-best replies.

Lemma 17 *Assume that for each player $i \in N$ the function $u^i : C \to \mathbb{R}$ is a Lipschitz function with $\mathcal{L}(u^i) \leq L$, and $u^i(\cdot, c^{-i})$ is quasi-concave on*

$\Delta(A^i)$ *for every fixed* $c^{-i} \in \Delta(A^{-i})$*. Then for every* $\varepsilon > 0$ *there is a Lipschitz function* $g : C \to C$ *such that* $g^i(c) \in \mathrm{BR}^i_\varepsilon(c^{-i})$ *for all* $c \in C$ *and all* $i \in N$*, and* $\mathcal{L}(g) = O((L/\varepsilon)^{m+1})$.

Proof. Put $\varepsilon_1 := \varepsilon/(6L)$. Let $z_1, ..., z_K \in C$ be a maximal $2\varepsilon_1$-net on C, and let $\beta_1, ..., \beta_K$ be the subordinated Lipschitz partition of unity given by Lemma 15. For each $i \in N$ and $1 \leq k \leq K$ take $x^i_k \in \mathrm{BR}^i_0(z^{-i}_k)$, and define $g^i(c) := \sum_{k=1}^K \beta_k(c) x^i_k$. Because $\beta_k(c) > 0$ if and only if $\|c - z_k\| < 3\varepsilon_1$, it follows that $x^i_k \in \mathrm{BR}^i_\varepsilon(c^{-i})$ (indeed, for every $y^i \in \Delta(A^i)$ we have $u^i(x^i_k, c^{-i}) > u^i(x^i_k, z^{-i}_k) - 3L\varepsilon_1 \geq u^i(y^i, z^{-i}_k) - 3L\varepsilon_1 > u^i(y^i, c^{-i}) - 6L\varepsilon_1 = \varepsilon$, where we have used $\mathcal{L}(u^i) \leq L$ twice, and $x^i_k \in \mathrm{BR}^i_0(z^{-i}_k)$). The set $\mathrm{BR}^i_\varepsilon(c^{-i})$ is convex by the quasi-concavity assumption, and so $g^i(c)$, as an average of such x^i_k, belongs to $\mathrm{BR}^i_\varepsilon(c^{-i})$.

Now $\max_{c \in C} ||c|| \leq \sqrt{m}$ (because $C \subseteq [0,1]^m$), and so $\|x_k\| \leq \sqrt{m}$ (where $x_k = (x^i_k)_{i \in N}$) for all k, and $K \leq (\sqrt{m}/\varepsilon_1)^m$ (because $C \subseteq B(0, \sqrt{m})$ contains the K disjoint open balls of radius ε_1 centered at the points z_k). Therefore the Lipschitz constant of $g(c) = \sum_{k=1}^K \beta_k(c) x_k$ satisfies $\mathcal{L}(g) \leq \sum_{k=1}^K \|x_k\| \, \mathcal{L}(\beta_k) \leq (\sqrt{m}/\varepsilon_1)^m \sqrt{m}\, 4^{m+2}/\varepsilon_1 = O(\varepsilon^{-m-1} L^{m+1})$ (see Lemma 15 (v)). □