Carnegie Mellon University

Dissertation

Competitive Analysis for Machine Learning & Data Science

Michael Spece

Thesis Committee:[1]
Cosma Shalizi
Joseph Kadane
Larry Wasserman
Barnabas Poczos
Pradeep Ravikumar
Aarti Singh

January 30, 2019

[1]Due to its size, all members are Carnegie Mellon University professors.

Dedicated to my parents, Roy and Rose

Contents

Preface

0.1 Explanation of Format

A living version of this document will be made available on arxiv, SSRN, or michaelspece.com.

There is a single running counter for displayed math; but the counters appear with a prefix, as in (Id1.1) meaning the first counted display in Chapter 1, which happens to be an *id*entity. In contrast, definition and result environments of different types (lemmas, theorems, etc.) each have their own counters.

0.1.1 Style

Articles are omitted from proper names. Punctuation is usually placed outside of quotes. I use the abbreviation "iff" for "if and only if". Semi-colons, in addition to their usual function of separating two sentences, may delimit a list of lists, where the inner lists are delimited by commas. The introductory section of a chapter is given a heading title iff it is cross-referenced.

0.2 Acknowledgments

External funding support was provided by (i) Department of Defense through its NDSEG program, (ii) National Science Foundation and PRC's Ministry of Science & Technology through their EAPSI program, and (iii) Institute for New Economic Thinking through its grant INO1400020. Internal funding support was provided by the Machine Learning Department and the Department of Statistics & Data Science.

I am grateful for the many small things my colleagues have done for me over the years, including bestowing me with a codename (Willie Neiswanger, Benjamin Cowley, Junier Oliva, and Kirstin Early), tracking me down in a snowstorm (Yu-Xiang Wang), and helping me train for Muay Thai (Robert Lunde). I appreciate the monumental comprehension, instruction, and encouragement of my advisor, Cosma Shalizi; the tremendous wisdom and tutelage of my mentor, Jay Kadane; and the time and insight of my entire committee. I am indebted to those who have been among my dearest friends for significant portions of my

life, including Xiaoyun Yang, Jingxiong Wang, Runqi Wang, David Hao, and Andrew Kositsky.

Chapter 1

Competitive Analysis

Statistical machines learn from regularity in data and are often designed for stationary or even independent and identically distributed (IID) processes. However, in most real-world applications it is not known how well the data process or non-trivial transformations thereof conform to theoretical assumptions. Moreover, this is impossible to learn when past data may be misleading or otherwise unrepresentative of the future. An adversarial data process, on the other hand, is not subject to probabilistic constraints; instead an adversary can deterministically attempt to mislead or otherwise confuse the machine. Of course designing for an adversary has its own limitation: that of being pessimistic. Despite the disparity between IIDness and adversarialism, for a given application, it may not be known which will better approximate the data. Fortunately, as shown in several supervised settings in Chapter 2, learning from data that is generated by an adaptive adversary is not much harder (statistically) than if it were generated by a static distribution. More precisely, the minimax regret values differ by a constant factor. In that case, a machine optimally designed for an adversary is necessarily competitive with any other, even when the data process is IID.

This idea of being competitive in the worst-case (and therefore all cases) arises in other problems such as regret or excess risk minimization. The next section defines competitive difference and prescribes analysis thereon as a way of characterizing how well (not necessarily quantified) uncertainty can be dealt with. In each of several examples that follow, an algorithm or set of players must cope with incomplete information about the problem or game it is trying to address, such as what is the optimal assumption or decision. This incompleteness impedes an unambiguous objective; the competitive difference provides clarity by furnishing a (fully specified) objective on the basis of which an optimal decision can be formulated.

1.1 Competitive Difference

Given (C1) a loss function $\ell : \mathcal{A} \times \mathcal{D} \to \mathbb{R}$, in which $\mathcal{A}$ can be interpreted as a space of algorithms or actions and $\mathcal{D}$ the set of possible data distributions, (C2) a set of **benchmarks** $\mathcal{B} \subseteq \mathcal{A}$, and (C3) a constant $C \geq 1$, the worst-case **competitive difference** of $a \in \mathcal{A}$ is

$$\mathrm{CD}_a := \sup_{d \in \mathcal{D}} \mathbb{E}\left[\ell(a, d) - C \inf_{a_* \in \mathcal{B}} \ell(a_*, d)\right],$$

assuming the expectation exists. $\mathcal{B}$ can be interpreted as the potential optimal decisions that a decision maker is unsure about.

Similarly, the minimax **competitive difference** is

$$\mathrm{CD} := \inf_{a \in \mathcal{A}} \mathrm{CD}_a. \tag{Id1.1}$$

In addition to this common language, the following general inequality will be useful: For all $a \in \mathcal{A}$,

$$\mathrm{CD}_a \leq \sup_{d \in \mathcal{D}} \mathbb{E}\left[\ell(a, d) - C \inf_{a_* \in \mathcal{B}, d_* \in \mathcal{D}} \ell(a_*, d_*)\right]. \tag{A1.2}$$

So that this bound is non-trivial, it is helpful to assume $\inf_{a_* \in \mathcal{B}, d_* \in \mathcal{D}} \ell(a_*, d_*) > -\infty$.

The central prescription of competitive analysis is the following. CD measures to what extent the problem's uncertainty can be tamed. If the competitive difference of a problem (algorithm, respectively) is sufficiently small, then the problem (algorithm) is said to be **feasible** (**competitive**). What constitutes sufficiently small is made precise in the examples and Section 1.6.1. In case of infeasibility, one would be advised to re-specify the problem (perhaps less ambitiously or supplemented with additional prior information or computational resources) so that it is feasible. The remainder of this dissertation is devoted to feasibility of given problems rather than problem specification (within a given application domain or environment, for example), though the former could provide a menu of feasible options or inspiration for the latter.

1.2 Regret

Regret, defined as a competitive difference by Table 1.1, is the learning objective function that is adopted in this dissertation's study of unknown regularity. It admits an (asymptotic) concept of feasiblity (known as learnability) that is attainable against an adversary yet can provide strong guarantees when the data is regular. For the adversarial case, the parameters[1] defining regret as CD are given in the table.

[1] Parameter here is used in the generic mathematical sense of a variable or term specifying a problem. When parameter is to be understood in the statistical sense of being a random variable capable of estimation, it will be qualified as a **statistical parameter**.

Table 1.1: Regret

Parameter	Special Form
$\mathcal{D}$	$\mathcal{X}^T$
$\mathcal{A}$	$\mathcal{Y}^{\cup_{t=0}^{T-1}\mathcal{X}^t}$
$\mathcal{B}$	$\cup_{y\in\mathcal{Y}}\left[\{y\}^{\cup_{t=0}^{T-1}\mathcal{X}^t}\right]$
C	1
ℓ	$\sum_{t=1}^{T}\ell_0(a_t,d_t)$

$\mathcal{D}$ is sequential data from some space $\mathcal{X}$ of some length $T\in\mathbb{N}_+$, $\mathcal{A}$ is the set of online algorithms thereon with respect to a set of possible $\mathcal{Y}$-valued outputs, $\mathcal{B}$ is the strict subset thereof that produce the same output each period, C is unity, and ℓ is a cumulative loss given the sequences of inputs a_t and outputs d_t. Identifying CD with Reg in the context of regret analysis,

$$\mathrm{Reg}_T(a,d)=\sum_{t=1}^{T}\ell_0(a_t,d_t)-\inf_{a_{**}\in\mathcal{B}}\sum_{t=1}^{T}\ell_0(a_{**,t},d_t),$$

where there is no expectation because the data is adversarial. Learnability is an $o(T)$ bound on the regret.

A connection between regret and "competitive" analysis was apparently recognized in [58, 59]. After presenting additional examples, more will be said about this connection (in Section 1.4.1).

1.3 Unknown Regularity

The previous section merely considered the adversarial case. Data might alternatively be independent and identically distributed. If it is not known which is the case, one can nest these regrets inside a competitive difference that thereby considers both possibilities:

Parameter	Special Form
$\mathcal{D}$	{IID, adversarial}
$\mathcal{A}$	as before
$\mathcal{B}$	$\mathcal{A}$
C	arbitrarily large
ℓ	worst-case regret

Feasibility is an upper bound of 0 on the competitive difference over all possible horizons T. Lower bounds under independent and identically-distributed data that match upper ones under adversarial data imply feasiblity, via Approximation A1.2. As already alluded to, this is often the case: See Sections 2.5 and 2.7 for a discussion of this matching, and note further the lower and upper bounds discussed therein apply under both independent and identically-distributed and adversarial data. In summary, low-regret algorithms are competitive under unknown regularity.

Taking, as the more primitive objective, expected cumulative loss

$$\mathrm{Lum}_T(a,d) := \mathbb{E}\sum_{t=1}^{T}\ell_0(a_t,d_t)$$

(Lum is a portmanteau of *l*oss, c*um*ulative) rather than regret

$$\mathrm{Reg}_T(a,d) = \mathbb{E}\left(\sum_{t=1}^{T}\ell_0(a_t,d_t) - \inf_{a_{**}\in\mathcal{B}}\sum_{t=1}^{T}\ell_0(a_{**,t},d_t)\right),$$

one obtains asymptotic optimality by adopting an algorithm designed for low regret on adversarial data, even when the data is IID, as follows.

Theorem 1 (Asymptotic Minimax Optimality with Error Bound)**.** Let $\mathcal{B}$ be as in Section 1.2 and suppose

(C4) $\mathbb{E}\inf_{a_*\in\mathcal{B}}\sum_{t=1}^{T}\ell_0(a_{*,t},d_t) > -\infty$ and $\mathrm{Lum}_T(a,d) < \infty$ for all $a\in\mathcal{A}$ and d that is IID,

(C5) $a\in\mathcal{A}$ is competitive with respect to the unknown regularity regime and constant C,

(C6) $\sup_{d'\ \mathrm{IID}}\mathrm{Lum}_T(a_*,d') \geq \sup_{d'\ \mathrm{IID}}\mathbb{E}\inf_{a_{**}\in\mathcal{B}}\sum_{t=1}^{T}\ell_0(a_{**,t},d'_t) + C'\sup_{d'\ \mathrm{IID}}\mathrm{Reg}_T(a_*,d')$ for some $C'\geq 0$, and

(C7) $\inf_{a_*\in\mathcal{A}}\sup_{d_*\ \mathrm{IID}}\mathrm{Lum}_T(a_*,d_*) \geq 0$.

Then

$$\sup_{a_*\in\mathcal{A}}\frac{\sup_{d\ \mathrm{IID}}\mathrm{Lum}_T(a,d)}{\sup_{d'\ \mathrm{IID}}\mathrm{Lum}_T(a_*,d)} \leq 1 + \frac{(C-C')\inf_{a_*\in\mathcal{A}}\sup_{d\ \mathrm{IID}}\mathrm{Reg}_T(a_*,d)}{\sup_{d\ \mathrm{IID}}\mathbb{E}\inf_{a_{**}\in\mathcal{B}}\sum_{t=1}^{T}\ell_0(a_{**,t},d_t)}.$$

Proof. For all $a,a_*\in\mathcal{A}$ and IID d,d',

$$\frac{\mathrm{Lum}_T(a,d)}{\mathrm{Lum}_T(a_*,d')} \leq \frac{\mathrm{Lum}_T(a,d)}{\sup_{d'\ \mathrm{IID}}\mathbb{E}\inf_{a_{**}\in\mathcal{B}}\sum_{t=1}^{T}\ell_0(a_{**,t},d'_t) + C'\sup_{d'\ \mathrm{IID}}\mathrm{Reg}_T(a_*,d')}$$

Now adding and subtracting the denominator in the numerator; and ultimately imposing competitiveness on a,

$$= 1 + \frac{\mathrm{Lum}_T(a,d) - \sup_{d' \text{ IID}} \mathbb{E} \inf_{a_{**} \in \mathcal{B}} \sum_{t=1}^T \ell_0(a_{**,t}, d'_t) - C' \sup_{d' \text{ IID}} \mathrm{Reg}_T(a_*, d')}{\sup_{d' \text{ IID}} \mathbb{E} \inf_{a_{**} \in \mathcal{B}} \sum_{t=1}^T \ell_0(a_{**,t}, d'_t) + C' \sup_{d' \text{ IID}} \mathrm{Reg}_T(a_*, d')} \tag{Id1.3}$$

$$\leq 1 + \frac{\mathrm{Lum}_T(a,d) - \mathbb{E} \inf_{a_{**} \in \mathcal{B}} \sum_{t=1}^T \ell_0(a_{**,t}, d_t) - C' \sup_{d' \text{ IID}} \mathrm{Reg}_T(a_*, d')}{\sup_{d' \text{ IID}} \mathbb{E} \inf_{a_{**} \in \mathcal{B}} \sum_{t=1}^T \ell_0(a_{**,t}, d'_t)} \qquad C' \geq 0$$

$$= 1 + \frac{\mathrm{Reg}_T(a,d) - C' \sup_{d' \text{ IID}} \mathrm{Reg}_T(a_*, d')}{\sup_{d' \text{ IID}} \mathbb{E} \inf_{a_{**} \in \mathcal{B}} \sum_{t=1}^T \ell_0(a_{**,t}, d'_t)}$$

$$= 1 + \frac{\mathrm{Reg}_T(a,d) - C \sup_{d' \text{ IID}} \mathrm{Reg}_T(a_*, d') + (C - C') \sup_{d' \text{ IID}} \mathrm{Reg}_T(a_*, d')}{\sup_{d' \text{ IID}} \mathbb{E} \inf_{a_{**} \in \mathcal{B}} \sum_{t=1}^T \ell_0(a_{**,t}, d'_t)}$$

$$\leq 1 + \frac{(C - C') \sup_{d' \text{ IID}} \mathrm{Reg}_T(a_*, d')}{\sup_{d' \text{ IID}} \mathbb{E} \inf_{a_{**} \in \mathcal{B}} \sum_{t=1}^T \ell_0(a_{**,t}, d'_t)} \qquad \text{Competitiveness.}$$

□

Remark: This result highlights the constant C establishing competitiveness and shows its influence not only on regret but cumulative loss. In particular, when $C' = 1$, the exact equality of minimax regrets between the two regularity regimes would ensure exact optimality with respect to cumulative loss.

1.4 Constant Factor Approximations and Competitive Ratios

The examples of this section come from computer science, in which "competitive" in the worst-case sense was first coined ([41, 11]). Both the terms "factor" and "ratio" allude to the following fact.

Proposition 1. Suppose $a \in \mathcal{A}$, (C8) for all $d \in \mathcal{D}$, $\ell(a,d)$ is integrable, and (C9) $\inf_{a_* \in \mathcal{B}, d_* \in \mathcal{D}} \ell(a_*, d_*) > 0$. Then

$$\mathrm{CD}_a \leq 0$$

iff

$$\sup_{d \in \mathcal{D}} \frac{\mathbb{E}\, \ell(a,d)}{\mathbb{E} \inf_{a_* \in \mathcal{B}} \ell(a_*, d)} \leq C.$$

Proof. $\sup_{d \in \mathcal{D}} \frac{\mathbb{E}\, \ell(a,d)}{\mathbb{E} \inf_{a_* \in \mathcal{B}} \ell(a_*, d)} \leq C$ iff

$$\sup_{d \in \mathcal{D}} \left[\frac{\mathbb{E}\, \ell(a,d)}{\mathbb{E} \inf_{a_* \in \mathcal{B}} \ell(a_*, d)} - C \right] = \sup_{d \in \mathcal{D}} \frac{\mathbb{E}\, \ell(a,d) - C\, \mathbb{E} \inf_{a_* \in \mathcal{B}} \ell(a_*, d)}{\mathbb{E} \inf_{a_* \in \mathcal{B}} \ell(a_*, d)} \leq 0$$

iff, for all $d \in \mathcal{D}$, $\mathbb{E}\,\ell(a,d) - C\,\mathbb{E}\inf_{a_* \in \mathcal{B}} \ell(a_*,d) \leq 0$ iff $0 \geq \sup_{d\in\mathcal{D}} \mathbb{E}\,\ell(a,d) - C\,\mathbb{E}\inf_{a_* \in \mathcal{B}} \ell(a_*,d) = \mathrm{CD}_a$. The second iff is by Condition (C9). The last iff is by Condition (C8). □

The problem's data $\mathcal{D}$ may be various aspects of the problem or observations. As for unknown regularity, C is chosen so that the competitive difference is bounded above by 0 and feasiblity corresponds to the existence of such a C.

"Constant factor approximation" connotes a setting where the algorithms of $\mathcal{A}$ are faster or otherwise less complex than those of $\mathcal{B}$. For example, $\mathcal{A}$ might be polynomial time algorithms for the traveling salesmen problem, approximations which have been successfully studied since at least [49]. In addition to studies of computational complexity, this type of comparison arises in statistics. See, for example, Equation 16 of [27].

"Competitive ratio" connotes a narrower setting where $\mathcal{D}$, $\mathcal{A}$, and $\mathcal{B}$ are as for regret; ℓ is amortized (cumulative) computational time complexity (similar to the case of regret, but interpreted more narrowly). One may further allow randomized algorithms and differentiate between data that is generated obliviously or adversarially; in the latter case, the sequence of data can adapt to previous outputs of the algorithm. (Though it is only against randomized algorithms that outputs are not already known in advance to Nature.)

For a recent perspective on the use of this ratio as a performance measure, see [11]. Of the two competitive differences of this section, only the "ratio" one will be treated further, via the following example, whose application of worst-case analysis to list ordering was novel, contrasting with the expected performance analyses beginning nearly two decades earlier.

In [8, 9], the problem is to dynamically order a growing list for faster access (lower computational cost), where accessing the ith item is assumed to cost i and re-ordering the list is free. That cost structure corresponds to a system for low latency response to infrequent (non-burst) queries. Nonetheless, the problem's essence is not limited to data structure design, but more general ordering problems with applications to robotics ([53]).

[8, 9] compare the cumulative access times to that of the best full static list and shows that the ratio is bounded by 2 for certain deterministic algorithms. While a comparison between these two classes appears haphazard at first glance and indeed the authors themselves recognize the "fundamental" difference in a growing versus complete list, the former's cost can be readily bounded in terms of the latter's as follows. Let N be the final number of list items and $c(\texttt{list}, \texttt{query})$ be the cost of querying `query` given ordered list `list`. Require the algorithm to output lists with all N items. The original setting corresponds to removing those items which have yet to be queried. Let $\mathrm{rem}_t(\texttt{list})$ be the list obtained by such removal. The claimed bound can now be precisely stated as, for all sequences of queries and $t \in \{1, \cdots, T\}$, $c(\mathrm{rem}_t(\texttt{list}), \texttt{query}) \leq c(\texttt{list}, \texttt{query})$ (for all lists and queries). The next section shows that it is not only possible to get ratio bounds for this new fixed list size setting, but also regret bounds that imply ratio ones.

1.4.1 Relationship to Regret

Given they both concern online problems, it is not uncommon to see both regret and competitive ratios used in the same work—see for example [53] (which contained constant factor approximations to boot). Of course the framework here links the two concepts explicitly. A more intimate connection between regret and the competitive ratio is the following. Under learnability and an easily checked condition, there is a competitive ratio bound.

Proposition 2. Suppose learnability and

(C10) for all $d \in \mathcal{X}^\infty$, $0 < \inf_{a_* \in \mathcal{B}} \ell(a_*, d_{1:T}) = \Omega(T)$.

Then there exists $C > 1$ such that

$$\inf_{a \in \mathcal{A}} \sup_{d \in \mathcal{X}^T} \left(\ell(a, d) - C \inf_{a_* \in \mathcal{B}} \ell(a_*, d) \right) \leq 0.$$

Proof. Learnability implies there exists a real-valued function f such that

$$\inf_{a \in \mathcal{A}} \sup_{d \in \mathcal{X}^T} \left(\ell(a, d) - C \inf_{a_* \in \mathcal{B}} \ell(a_*, d) \right) \leq f(T) = o(T).$$

One can take

$$C := 1 + \sup_t \frac{f(t)}{\inf_{a_* \in \mathcal{B}, d \in \mathcal{X}^t} \ell(a_*, d)}.$$

□

Because list access costs at least 1 every period (satisfying Condition (C10)) but is bounded by N (satisfying learnability, for expected regret, via a randomized experts algorithm, such as sampling with probabilities given by Hedge (see the proof of Theorem 7), treating the $N!$ possible list orderings as the experts/actions), Proposition 2 applies to the list ordering problem. Using the trivial lower bound of 1 leaves a dependence on the list size, however. A variant of Proposition 2 with weaker conditions admitting a lower bound that is increasing in N could probably give stronger results and whether list size dependence is essential at all could probably be resolved with the techniques of Chapter 2, which is devoted to lower bounds.

1.5 Oracle Relations

An oracle[2] relation, often presented as an inequality, has hitherto been an informal idea[3] based on "ideal risk" ([28, 12]), though a special case of an oracle relation, excess risk, has a precise definition that is a batch analog of regret,

[2] Not to be confused with oracle machine.

[3] Apparently presumed to be easily grasped in context. However, for non-experts, an easy grasp is unlikely to be the case.

including $C := 0$ and the same definition of feasibility.[4] In the Bayesian setting, that is when there is a prior on the potential truth, the ideal risk is the Bayes risk and its excess as a function of the data is also referred to as conditional regret ([30]). Because the data is delivered in a batch, ℓ is the risk $\mathbb{E}\,\ell_0(a_T, d_T)$ corresponding to an instantaneous loss $\ell_0(a_T, d_T)$ and $\mathcal{D}$ could be of the form $\mathcal{X}^T$, as before, or alternatively taken to be unordered rather than sequential data, correspondingly $\mathcal{A}$ estimators that are functions of unordered data. Here $\mathcal{B} := \mathcal{A}$ is a reasonable choice because risk is a less ambitious objective than cumulative loss as used in regret, where the expectation appears outside of the infimum. Incorporating a more general difference operator than minus (that is with a $C > 1$) seems to provide a formalization of oracle relation that is adequate for applications beyond excess risk.

1.6 Extensions

This section generalizes the notion of competitive difference to more directly model ratios, such as those of Sections 1.4 and 1.3–1.5, among other applications.

Given Condition (C1), (C11) sets of benchmark outcomes $\mathcal{B} : \mathcal{A}\times\mathcal{D} \to 2^{\mathcal{A}\times\mathcal{D}}$, and (C12) a difference operator $\Delta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the (generalized) minimax **competitive difference** is

$$\Delta\left(\ell(a_s, d_s), \inf_{(a_*,d_*)\in\mathcal{B}(a_s,d_s)} \ell(a_*, d_*))\right).$$

This definition enjoys more symmetry in the actions and data than in Identity Id1.1. The difference operator is what can directly handle ratios, in addition to the arithmetic difference. Similarly, the (generalized) worst-case **competitive difference** of $a \in \mathcal{A}$ is

$$\mathrm{CD}_a := \mathbb{E}\sup_{d\in\mathcal{D}} \Delta\left(\ell(a, d), \inf_{(a_*,d_*)\in\mathcal{B}(a,d)} \ell(a_*, d_*)\right).$$

One recovers the competitive difference of Section 1.1 by taking $\mathcal{B}$ to be of the form $\mathcal{B}_\mathcal{A} \times \{d\}$ for some $\mathcal{B}_\mathcal{A} \subseteq \mathcal{A}$; and the difference operator of the form

$$\Delta(A, B) \mapsto A - CB,$$

for some $C \geq 1$.

The following inequality generalizes Approximation A1.2. Suppose

1. for every value of its first argument, Δ is non-increasing in its second on $\operatorname{Ran}\ell$
2. for some $\mathcal{B}_* : \mathcal{A} \times \mathcal{D} \to 2^{\mathcal{A}\times\mathcal{D}}$ and all $a \in \mathcal{A}, d \in \mathcal{D}$,

$$\mathcal{B}_*(a, d) \subseteq (\supseteq)\,\mathcal{B}(a, d),$$

respectively,

[4] This notion of excess risk is not to be confused with the epidemiological concept of the same name.

then, for all $a \in \mathcal{A}$,

$$\mathrm{CD}_a \geq (\leq) \sup_{d \in \mathcal{D}} \Delta \left(\ell(a,d), \inf_{(a_*,d_*) \in \mathcal{B}_*(a,d)} \ell(a_*,d_*) \right), \tag{A1.4}$$

respectively.

1.6.1 Parameterized Problems

The following notion more precisely formalizes feasibility from Section 1.1 for problems where $\mathcal{D}$ or ℓ is parametrized by a $T \in \mathbb{N}_+$ (for example, $\mathcal{D}_T := \mathcal{X}^T$ or ℓ_T the worst-case regret for horizon T). Let

$$\mathrm{CD}_T(a) := \sup_{d \in \mathcal{D}_T} \Delta \left(\ell_T(a,d), \inf_{a_* \in \mathcal{B}(a,d)} \ell_T(a_*) \right)$$

and $\mathrm{CD}_T := \inf_{a \in \mathcal{A}} \mathrm{CD}_T(a)$. For a given subset S of "satisfactory performances" ("small" losses) in $\mathbb{R}^\infty$, the problem is **feasible** if and only if $(\mathrm{CD}_T)_{T=1}^\infty \in S$. Those algorithms a for which

$$(\mathrm{CD}_T(a))_{T=1}^\infty \in S$$

are **competitive**.

Feasibility could be cast in terms of a binary-valued difference operator

$$\Delta_{\mathrm{meta}}(a, d^1, d^2, \cdots) := I_{\left[\Delta\left(\ell_T(a,d^T), \inf_{a_* \in \mathcal{B}(a,d)} \ell_T(a_*)\right)\right]_{T=1}^\infty \in S},$$

where $d^T \in \mathcal{D}_T$ for all $T \in \mathbb{N}_+$. If $\mathcal{D}_T := \mathcal{X}^T$ and Nature should be **oblivious** to T, instead

$$\Delta_{\mathrm{meta}}(a, d) := I_{\left[\Delta\left(\ell_T(a,d_{1:T}), \inf_{a_* \in \mathcal{B}(a,d)} \ell_T(a_*)\right)\right]_{T=1}^\infty \in S},$$

in which $d \in \mathcal{X}^\infty$.

For (i) the two examples of Section 1.4 and (ii) unknown regularity (1.3), S is the negative orthant. For regret and excess risk, S are those functions which approach the negative orthant.

1.6.2 Admissibility

For every $a \in \mathcal{A}$, let $\ell(a, \cdot)$ denote the map $d \mapsto \ell(a,d)$; an (arbitrary) $a_* \in \mathcal{A}$ is **admissible** iff $\ell(a_*, \cdot)$ is a maximal element of $\{\ell(a,\cdot) : a \in \mathcal{A}\}$ with respect to the Pareto order.

Admissibility is an attractive property and thereby a useful primitive concept. However, it has limitations when considered in isolation. One, admissibility may be unattainable, analytically difficult, or computationally infeasible. Orthogonally to one, admissibility may lead to non-unique solutions with (a priori) potentially vastly different loss profiles. A reasonable compromise (if not

plagued with the same limitations) then is to be merely competitive but against all admissible actions in the sense of a bound on

$$\inf_{a\in\mathcal{A}}\sup_{d\in\mathcal{D}}\Delta\left(\mathbb{E}\,\ell(a,d),\inf_{a_*\in\mathcal{A}_{\text{admissible}}}\mathbb{E}\,\ell(a_*,d)\right). \tag{Id1.5}$$

Bounding Identity Id1.5 by a competitive difference (under a suitable choice of benchmarks) may provide an escape from the difficulties of a more direct analysis of admissibility. For instance, regret, excess risk, constant factor approximation, and competitive ratio all implicitly provide approximate admissibility by bounding Identity Id1.5. Each of these concepts has been arguably more fruitful than admissibility, especially in the modern literature.

1.6.3 Competitive Difference under a Solution Concept

In a strategic form game, the losses (or payoffs) of the players are specified and a solution of the game is a set of predicted strategies for the players. Certain solution concepts such as Nash equilibrium restricts the possible solutions but does not guarantee uniqueness or that the solution is communally desirable. Re the latter, given a communal loss function ℓ for two players with action sets $\mathcal{A}$ and $\mathcal{D}$, respectively, competitive analysis can be used to characterize the sub-optimality of a given solution as follows.

Given Conditions (C1)–(C12) and (C13) a **solution** $a_s, d_s \in \mathcal{A}\times\mathcal{D}$, the (variant of) **competitive difference** is

$$\Delta\left(\ell(a_s,d_s),\inf_{(a_*,d_*)\in\mathcal{B}(a_s,d_s)}\ell(a_*,d_*))\right).$$

Re which solution to select, one may consider two extremes among the possibilities: **anarchy**, the worst case with respect to the communal loss function, and **stability**, the best case.

Now to recover the prices of anarchy and stability take (i) Δ as division; (ii) the solutions as anarchy and stability, respectively, under the concept of Nash equilibrium and given player loss functions $\ell_1 : \mathcal{A}\times\mathcal{D}$ and $\ell_2 : \mathcal{A}\times\mathcal{D}$; and (iii) $\mathcal{B}$ as identically $\mathcal{A}\times\mathcal{D}$.

A designer or law maker would seek a system that aligns self interest with that of the community, by ensuring low prices of stability and anarchy. Feasibility thus corresponds to a price "close to 1" ([47]). In the context of given game solutions, feasibility purely depends on the game rather than the strategies, as there is no longer a concept of worst-case competitive difference.

A generalization of competitive difference to n players for all $n > 2$ is straight-forward and omitted.

1.7 Chapter Review

Worst-case analysis is a form of preparation, in the wake or anticipation of uncertainty that is difficult to quantify. Competitive differences quickly and

succinctly unifies several major concepts, from various fields, namely (i) constant factor approximations and competitive ratios from computer science, (ii) regret from machine learning, and (iii) oracle relations, including excess risk, from statistics. Generalizing further—from the minimax perspective to solution concepts—additionally recovers the prices of stability and anarchy from game theory. The general prescription in Section 1.1 provides a systematic methodology for helping to deal with uncertainty, particularly incomplete information, since minimizing the competitive difference furnishes an objective. It also provides a concise language for describing the effects of incomplete information on (not purely Bayesian)[5] learning, as seen in Section 1.3. Under the perspective of competitive analysis, the theoretical insights of Chapter 2 can be seen as the application of a coherent methodology.

In more general terms, specification of an unknown reality is one of the fundamental challenges of science. Given multiple possible realities or worlds, one can attempt to be competitive in all of them.

[5]The purely Bayesian approach is to presume, or enforce through subjective information, complete specification of a problem. In other words, incompletely specified problems do not exist in a Bayesian world. That is not to say one cannot apply Bayesian principles to certain aspects of an incompletely specified problem. Indeed, Bayesian principles can co-exist with those in a world of competitive differences.

Chapter 2

Regret Bounds

2.1 Introduction

This chapter develops machinery for non-asymptotic, high-dimensional lower regret bounds with application to learning via expert advice under metric loss, that is a distance function between estimates and truth ([21]).

Consider, for example, a problem with T periods in which every period each of N experts[1] proffers advice and Nature assigns a loss function that is convex and bounded over the expert advice. [13, 15] shows the worst-case regret is, to within a constant factor, $\sqrt{T \log N}$ as $T, N \to \infty$ (in that order, under absolute loss). That rate is grossly inaccurate in high enough dimensions $N \gg T$. Indeed, by boundedness one immediately obtains the (asymptotically equivalent) upper bound $O\left(\min\left\{T, \sqrt{T \log N}\right\}\right)$. Though a priori this still might be inaccurate (in particular, for $\log N \approx T$), it turns out that the non-asymptotic minimax rate is the same, even when the expert advice and loss functions are constrained to be IID, as shown by Theorem 4 of the section (2.4) on applications.

Re the underlying machinery, a single set of sufficient conditions (described by Corollary 5) applies in an abstract setting that encompasses not only experts with absolute loss, but metric loss; moreover, these conditions seem flexible enough to precisely account for the problem's complexity, obtaining optimal bounds in the convex case (Sections 2.5 and 2.7). As for the relevance of metric loss, it allows for data to be elements of an arbitrary real vector space (the predictor function is assumed to be linear so does not quite generalize to arbitrary metric spaces), including a space of functional data.

Beyond experts, because of their elegant properties (high generality achieved concisely), a particular class of abstract vector spaces—Banach ones—have received recent attention, with applications to online linear or convex optimization (for example, [56, 55, 6, 23]). A setup considered herein (Section 2.4.2) once again (i) ventures beyond Banach spaces, to give a brief but still elegant

[1] Experts can be human decision-makers, computer emulations or imitations thereof, models, or actions.

treatment of online linear optimization (which is sufficient to generate lower bounds for convex optimization), and (ii) derives a non-asymptotic $\omega\left(\sqrt{T}\right)$ lower bound.

The machinery applied to experts can be re-used for linear optimization, helping to show how closely related these problems are. The lower bounds share the form $O\left(\sqrt{T}\min\left\{\log\kappa, (\log\kappa)^{\alpha}\sqrt{T}\right\}\right)$, where κ and T are the problem's complexity and horizon, respectively, and $\alpha \geq 0$ is also problem dependent. In particular, each bound is $\Omega\left(\sqrt{T}\right)$ and non-asymptotic in the problem's dimension.

The remainder of this paper is organized as follows. The next section specifies the general format of the learning problems to be considered. Section 2.3 abstracts the max regret as an extreme value and derives general minimax regret bounds (under the previously mentioned sufficient conditions), which are applied in Section 2.4 to examples. Section 2.5 considers matching upper bounds, while Section 2.6 summarizes a key computational issue.

2.2 Problem Setup: A Sequential Game

The sequential nature of the decision problem is tracked through co-ordinates, using the following notation.

Definition 1 (Co-Ordinates)**.** Superscripts on set elements denote time indices.

Learning is an alternating-turn game between **Learner** and **Nature** appearing in Game 1. All sets are assumed non-empty.

Given: $T \in \mathbb{Z}_+$; sets $\mathcal{X}, Y, Z$; $F \subseteq Z^{\mathcal{X}}$; $\mathrm{d} : Z \times Y \to \mathbb{R}$
for $t = 1, \cdots, T$ **do**

1. Nature reveals $x^t \in \mathcal{X}$
2. The learner reveals $f^t \in F$
3. Nature privately decides $v^t \in Y$
4. $y^t := Uv^t$ is publicly revealed

end

Game 1: Prediction

Though this is set up as a prediction problem, it subsumes online linear optimization, as shown in Section 2.4.2.

2.2.1 Player Strategies

Each decision f^t is (implicitly) a map from the history of observations

$$(x^1, y^1), \cdots, (x^{t-1}, y^{t-1}), x^t$$

(to F).

The following constraint on Nature is assumed:

(C2) $((x^t, y^t))_{t=1}^T$ are IID and there exists a countable set that almost surely contains (x^1, y^1).

Whenever f^t, x, or y appears, it is to be implicitly assumed it is a strategy as just described, unless explicitly defined otherwise.

2.2.2 Objective

Let

$$\mathrm{R}_T(f, x, y) := \sum_{t=1}^{T} \mathrm{d}\big(f^t(x^t), y^t\big) - \inf_{f_* \in F} \sum_{t=1}^{T} \mathrm{d}\big(f_*^t(x^t), y^t\big).$$

Learner's objective is its expected regret or cumulative loss in excess of the best decision's performance,

$$\mathrm{Reg}_T(f, x, y) := \mathbb{E}\,\mathrm{R}_T(f, x, y) \tag{Id2.1}$$

(assuming the expectation exists), in the worst case:

$$\mathrm{RWC}_T(f) := \sup_{x,y} \mathrm{Reg}_T(f, x, y). \tag{Id2.2}$$

The **value** of Game 1 is

$$\mathrm{Val}_T := \inf_f \mathrm{RWC}_T(f).$$

Extending Reg to all maps z from the history of observations to Z, let z^t be its image in the tth period and

$$\mathrm{Reg}_T(z, x, y) := \mathbb{E}\,\mathrm{R}_T(f, x, y). \tag{Id2.3}$$

RWC and Val can be extended analogously. Denote the new value by $\underline{\mathrm{Val}}_T$, where the underline signifies it is potentially smaller:

$$\mathrm{Val}_T \geq \underline{\mathrm{Val}}_T.$$

2.3 Lower Bounds

2.3.1 Approximation of Sums

As the cumulative loss is a sum, it is useful to have general methods for approximating sums. In the following, let (L_t) denote a sequence of random variables, that is:

Definition 2 (Random Variable)**.** Consistent with Condition (C2), a **random variable** is a real-valued function for which there exists a countable set to which the function maps almost surely.

This countability restriction simplifies the analysis by helping skirt issues of measurability and allowing one to define expectation in terms of series rather than a suprema over them. It is more a choice of presentation than a means to achieving stronger results.

Definition 3 (Centeredness)**.** A random variable is **centered** iff its expectation is 0.

Theorem 2 (Marcinkiewicz-Zygmund)**.** Suppose $(L_t)_{t\in\mathbb{Z}_+}$ are independent and each L_t is centered. Then

$$\mathbb{E}\left|\sum_{t=1}^{T} L_t\right| \geq \frac{1}{\sqrt{2}}\,\mathbb{E}\sqrt{\sum_{t=1}^{T} L_t^2}.$$

Proof. See [31]. □

Definition 4 (Essential Supremum)**.** For every random variable L, let $\|L\|_\infty$ be its essential supremum, that is $\inf\{C : |L| \leq C \text{ almost surely}\}$.

Proposition 3 (Expectation of Square Root)**.** Suppose $\mathbb{P}(L \geq 0) = 1$ and $\mathbb{P}(L = 0) < 1$. Then $\|L\|_\infty > 0$ and

$$\mathbb{E}\sqrt{L} \geq \frac{\mathbb{E}\,L}{\sqrt{\|L\|_\infty}}$$

Proof. If L is not almost surely zero, by countable additivity, there exists a finite upper bound on $|L|$ that holds with positive probability, giving $\|L\|_\infty > 0$.

$$\begin{aligned}
\|L\|_1 &= \mathbb{E}\left(\sqrt{L}\right)^2 && \text{Non-negativity} \\
&\leq \left\|\sqrt{L}\right\|_\infty \left\|\sqrt{L}\right\|_1 && \frac{1}{\infty} + \frac{1}{1} = 1, \text{Hölder} \\
&= \sqrt{\|L\|_\infty}\,\mathbb{E}\sqrt{L}.
\end{aligned}$$

□

Corollary 1. Suppose the conditions of Theorem 2 and that L_1 is not almost surely 0. Then $\|L_1\|_\infty > 0$ and

$$\mathbb{E}\left|\sum_{t=1}^{T} L_t\right| \geq \frac{\mathrm{Var}(L_1)}{\sqrt{2}\|L_1\|_\infty}\sqrt{T}.$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left|\sum_{t=1}^{T} L_t\right| \\
&\geq \frac{1}{\sqrt{2}}\,\mathbb{E}\sqrt{\sum_{t=1}^{T} L_t^2} && \text{Theorem 2} \\
&\geq \frac{T\,\mathbb{E}\,L_1^2}{\sqrt{2T}\|L_1\|_\infty} && \text{Proposition 3} \\
&= \frac{\mathrm{Var}(L_1)}{\sqrt{2}\|L_1\|_\infty}\sqrt{T} && L_1 \text{ centered.}
\end{aligned}$$

□

2.3.2 Max in Two Dimensions

In order to account for the maximum's dependence on N, it is useful to establish initial conditions at $N = 2$.

Lemma 1 (Symmetrization)**.** Let $(L_k)_{k=1}^{\infty}$ be random variables. Then, for $N > 1$,

$$\mathbb{E}\max\{L_1, \cdots, L_N\} \geq \frac{1}{2}\,\mathbb{E}|L_1 - L_2|.$$

Proof. For $N > 1$,

$$\mathbb{E}\max\{L_1, \cdots, L_N\} \geq \mathbb{E}\max\{L_1, L_2\}.$$

L_1 and L_2 are implicitly integrable (having a well defined expectation that is not $\pm\infty$).

$$\begin{aligned}
&\mathbb{E}\max\{L_1, L_2\} \\
&= \frac{1}{2}\,\mathbb{E}\left(L_1 + L_2 + \max\{L_1 - L_2, L_2 - L_1\}\right). \qquad \text{(Id2.4)}
\end{aligned}$$

When L_1, L_2 are centered, Identity Id2.4 simplifies to

$$\frac{1}{2}\,\mathbb{E}\max\{L_1 - L_2, L_2 - L_1\} = \frac{1}{2}\,\mathbb{E}|L_1 - L_2|.$$

□

Remark: $L_1 - L_2$ is symmetric when L_1 and L_2 are IID.

Theorem 3 (Max's Growth for Bounded Random Variables)**.** Suppose each random variable of $(L_{n,t})_{n,t\in\mathbb{Z}_+}$ is centered, and that

(C3) $L_{1,1}$ does not almost surely equal $L_{2,1}$, and

(C4) $((L_{n,t})_{n\in\mathbb{Z}_+})_{t\in\mathbb{Z}}$ are IID (over t).

Then, for all $N > 1$,

$$\mathbb{E}\max_{n\in\{1,\cdots,N\}}\sum_{t=1}^{T} L_{n,t} \geq \frac{\mathrm{Var}(L_{1,1} - L_{2,1})}{2\sqrt{2}\|L_{1,1} - L_{2,1}\|_\infty}\sqrt{T}.$$

Proof. Because $N > 1$, by Lemma 1,

$$\begin{aligned}\mathbb{E}\max_{n\in\{1,\cdots,N\}}\sum_{t=1}^{T} L_{n,t} &\geq \frac{1}{2}\,\mathbb{E}\left|\sum_{t=1}^{T} L_{1,t} - \sum_{t=1}^{T} L_{2,t}\right| \\ &= \frac{1}{2}\,\mathbb{E}\left|\sum_{t=1}^{T} d_t\right|,\end{aligned} \tag{A2.5}$$

in which $d_t := L_{1,t} - L_{2,t}$.

It remains to lower bound $\mathbb{E}\left|\sum_{t=1}^{T} d_t\right|$. (d_t) obeys the conditions of Theorem 2 and d_1 is not almost surely 0. By Corollary 1,

$$\mathbb{E}\left|\sum_{t=1}^{T} d_t\right| \geq \frac{\mathrm{Var}(d_1)}{\sqrt{2}\|d_1\|_\infty}\sqrt{T}.$$

□

Remark: $L_{1,1} - L_{2,1}$, as appears in Theorem 3, even under dependence, may be a known distribution from which to derive its variance (and maximum magnitude).

Proposition 4. Suppose L_1, L_2 are independent and $\mathbb{E}\,L_2 = 0$. Then

$$\mathbb{E}|L_1 + L_2| \geq \mathbb{E}|L_1|. \tag{A2.6}$$

Proof.

$$\begin{aligned}\mathbb{E}\left(|L_1 + L_2|\,|\,L_1 \geq 0\right) &\geq \mathbb{E}\left(L_1 + L_2|\,L_1 \geq 0\right) \\ &= \mathbb{E}\left(|L_1|\,|\,L_1 \geq 0\right) + \mathbb{E}\,L_2.\end{aligned}$$

Similarly, albeit more complicatedly,

$$\begin{aligned}
&\mathbb{E}\left(|L_1+L_2||\, L_1<0\right)\\
=&\mathbb{P}(L_2<0)\mathbb{E}\left(-L_1-L_2|\, L_1<0, L_2<0\right)\\
+&\mathbb{P}(L_2\geq 0)\mathbb{E}\left(||L_1|-|L_2|||\, L_1<0, L_2\geq 0\right)\\
\geq&\mathbb{P}(L_2<0)\mathbb{E}\left(-L_1-L_2|\, L_1<0, L_2<0\right)\\
+&\mathbb{P}(L_2\geq 0)\mathbb{E}\left(|L_1|-|L_2||\, L_1<0, L_2\geq 0\right)\\
=&\mathbb{E}\left(|L_1||\, L_1<0\right)+\mathbb{P}(L_2<0)\mathbb{E}\left(-L_2|\, L_2<0\right)\\
+&\mathbb{P}(L_2\geq 0)\mathbb{E}\left(-L_2|\, L_2\geq 0\right)\\
=&\mathbb{E}\left(|L_1||\, L_1<0\right)-\mathbb{E}\, L_2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}|L_1+L_2|\\
\geq&\mathbb{E}|L_1|+\left(\mathbb{P}(L_1\geq 0)-\mathbb{P}(L_1<0)\right)\mathbb{E}\, L_2.
\end{aligned}$$

□

Lemma 2 (Monotonicity)**.** Suppose $(L_{k,t})_{1,1}^{2,\infty}$ are centered and obey Condition (C4). Then

$$\mathbb{E}\max\left\{\sum_{t=1}^{T}L_{1,t},\sum_{t=1}^{T}L_{2,t}\right\}$$

is non-decreasing as a function of T.

Proof. For all $k\in\{1,2\}$, let $S_{k,T}:=\sum_{t=1}^{T+1}L_{1,t}$.

By Lemma 1,

$$\begin{aligned}
&\mathbb{E}\max\left\{\sum_{t=1}^{T+1}L_{1,t},\sum_{t=1}^{T+1}L_{2,t}\right\}\\
&=\mathbb{E}\max\left\{L_{1,T+1}+S_{1,T},L_{2,T+1}+S_{2,T}\right\}\\
&=\frac{1}{2}\,\mathbb{E}|S_{1,T}-S_{2,T}+L_{1,T+1}-L_{2,T+1}|\\
&=\frac{1}{2}\,\mathbb{E}|\Delta S+\Delta L|, \qquad \text{(Id2.7)}
\end{aligned}$$

in which $\Delta S:=S_{1,T}-S_{2,T}$ and $\Delta L:=L_{1,T+1}-L_{2,T+1}$ are independent and centered; whereas,

$$\mathbb{E}\max\left\{\sum_{t=1}^{T}L_{1,t},\sum_{t=1}^{T}L_{2,t}\right\}=\frac{1}{2}\,\mathbb{E}|\Delta S|.$$

Invoking Proposition 4 concludes.

□

2.3.3 Max in High Dimensions

It remains to account for the maximum's dependence on κ, which either has a (i) bounded or (ii) unbounded growth for a fixed horizon T. In case (i), there is hope for a lower bound of the separable form $f(\kappa)g(T)$. Bounded growth make this impossible (except for situations where f is bounded). The remainder of this section is devoted to determining potentially non-separable decompositions.

An idea for simplifying the max of a sum is the following.

Lemma 3 (Tempered Super-Additivity for Two Terms)**.** Suppose D is a set;

$$B \subseteq \{\beta : \exists \alpha_0 : (\alpha_0, \beta) \in D\}$$

and, for all $\beta \in B$, $A_\beta \subseteq \{\alpha : (\alpha, \beta) \in D\}$ are non-empty;

(C5) $a : D \to \mathbb{R}$; $b : D \to \mathbb{R}$; and

(C6) there exist $\alpha_* : \beta \in B \mapsto \alpha_*(\beta) \in A_\beta$ and $\beta_* \in B$ such that, for all $\beta \in B$,

$$\begin{aligned}(\alpha_*(\beta), \beta) &\in D \\ a(\alpha_*(\beta), \beta) &\geq \sup_{\alpha \in A_\beta} a(\alpha, \beta) \\ b(\alpha_*(\beta_*), \beta_*) &\geq \sup_{\beta \in B} b(\alpha_*(\beta), \beta).\end{aligned}$$

Then

$$\sup_{(\alpha,\beta)\in D} [a(\alpha, \beta) + b(\alpha, \beta)] \geq \sup_{\alpha \in A_{\beta_*}} a(\alpha, \beta_*) + \sup_{\beta \in B} b(\alpha_*(\beta), \beta).$$

Proof. Because a supremum is at least as great as any particular point and $(\alpha_*(\beta_*), \beta_*) \in D$,

$$\begin{aligned}\sup_{(\alpha,\beta)\in D} a(\alpha, \beta) + b(\alpha, \beta) &\geq a(\alpha_*(\beta_*), \beta_*) + b(\alpha_*(\beta_*), \beta_*) \\ &\geq \sup_{\alpha \in A_{\beta_*}} a(\alpha, \beta_*) + \sup_{\beta \in B} b(\alpha_*(\beta), \beta),\end{aligned}$$

by Condition (C6). □

The following corollary shows one way by which the foregoing idea applies to product spaces.

Corollary 2 (Tempered Super-Additivity Specialized to a Product Space)**.** Suppose

(C7) $\{A, B\}$ are finite, non-empty sets; and

Condition (C5) with $D := A \times B$. Then, for all α_* mapping $\beta \in B$ to a branch of $\operatorname{argmax}_{\alpha \in A} a(\alpha, \beta)$ and $\beta_* \in \operatorname{argmax}_{\beta \in B} b(\alpha_*(\beta), \beta)$,

$$\begin{aligned}\max_{(\alpha,\beta)\in D} [a(\alpha,\beta) + b(\alpha,\beta)] &\geq \max_{\alpha \in A} a(\alpha, \beta_*) + \max_{\beta \in B} b(\alpha_*(\beta), \beta) \\ &\geq \min_{\beta \in B} \max_{\alpha \in A} a(\alpha, \beta) + \min_{\alpha \in A} \max_{\beta \in B} b(\alpha, \beta).\end{aligned}$$

Proof. In Lemma 3, one can take $A_\beta, B, \alpha_*, \beta_*$ to be, respectively, the A, B, α_*, β_* of this corollary. □

Corollary 3 (Tempered Super-Additivity)**.** Suppose

(C8) $\{A_i\}_{i=1}^K$ are finite, non-empty sets;

and, for all $i \in \{1, \cdots, K\}$, $a_i : \times_{i=1}^K A_i \to \mathbb{R}$. Then

$$\max_{\alpha \in \times_{i=1}^K A_i} \sum_{i=1}^K a_i(\alpha) \geq \sum_{i=1}^K \min_{\alpha_{-k} \in \times_{i \neq k} A_i} \max_{\alpha_k \in A_K} a(\alpha). \tag{A2.8}$$

As to obtaining product spaces, it suffices to extend a function, and use the following observation.

Proposition 5. Suppose $\{A_i\}_{i \in \{1, \cdots, K\}}$ are nonempty sets, $D \subseteq \times_{i=1}^K A_i$, $\iota : \times_{i=1}^K A_i \to D$ is an injection, and $a \times_{i=1}^K A_i :\to \mathbb{R}$. Then

$$\sup_{\alpha \in D} a(\alpha) \geq \sup_{\alpha' \in \times_{i=1}^K A_i} a\left(\iota(\alpha')\right).$$

More work needs to be done when considering expected maxima, because one would like the minima in sum decomposition Approximation A2.8 to appear outside the expectation.

Definition 5 (Distributions)**.** $\sim$ means distributed as, $\perp$ (statistically) independent.

The following two propositions follow directly from definitions.

Proposition 6 (Replacement)**.** If $a \sim b$ and f is a deterministic function, then $\mathbb{E} f(a) = \mathbb{E} f(b)$.

Proposition 7 (Identical Distributions)**.** If $a \perp b$, $a \perp c$, and $b \sim c$, then $(a, b) \sim (a, c)$.

Lemma 4 (Mean Version)**.** Suppose

(C9) D is a set;

(C10) $A := \{\alpha : \exists \beta_0 : (\alpha, \beta_0) \in D\}$ and $B := \{\beta : \exists \alpha_0 : (\alpha_0, \beta) \in D\}$ are non-empty;

(C11) Ω is a sample space;

(C12) a, b are independent random functions such that for all $\omega \in \Omega$, $a^\omega, b^\omega : D \to \mathbb{R}$;

(C13) $\sup_{(\alpha,\beta)\in D} a(\alpha,\beta) + b(\alpha,\beta)$ is integrable; and

(C14) there exist measurable $\alpha_* : \mathbb{R}^D \to A$ and $\beta_* : \mathbb{R}^D \times \mathbb{R}^D \to B$ such that, for all $\beta \in B$,

$$\begin{aligned}(\alpha_*^a, \beta) &\in D\\ a(\alpha_*^a, \beta) &= \sup_{\alpha:(\alpha,\beta)\in D} a(\alpha,\beta)\\ b(\alpha_*^a, \beta_*^{a,b}) &= \sup_{\beta\in B} b(\alpha_*^a, \beta)\\ a &\perp \beta_*^{a,b}.\end{aligned}$$

Then

$$\mathbb{E}\sup_{(\alpha,\beta)\in D} a(\alpha,\beta) + b(\alpha,\beta) \geq \inf_{\beta\in B}\mathbb{E}\sup_{\alpha:(\alpha,\beta)\in D} a(\alpha,\beta) + \inf_{\alpha\in A}\mathbb{E}\sup_{\beta:(\alpha,\beta)\in D} b(\alpha,\beta).$$

Proof. By Lemma 3 (integrability ensures the expectation exists and distributes with inequality in the designated direction),

$$\mathbb{E}\sup_{(\alpha,\beta)\in D} a(\alpha,\beta) + b(\alpha,\beta) \geq \mathbb{E}\sup_{\alpha:(\alpha,\beta)\in D} a(\alpha,\beta_*^{a,b}) + \mathbb{E}\sup_{\beta\in B} b(\alpha_*^a,\beta).$$

By Condition (C14), there exists $\beta^b \sim \beta_*^{a,b}$. By Proposition 7, $(a,\beta_*^{a,b}) \sim (a,\beta^b)$. Then

$$\begin{aligned}\mathbb{E}\sup_{\alpha:\left(\alpha,\beta_*^{a,b}\right)\in D} a(\alpha,\beta_*^{a,b}) &= \mathbb{E}\sup_{\alpha:(\alpha,\beta^b)\in D} a(\alpha,\beta^b) && \text{Proposition 6}\\ &= \mathop{\mathbb{E}}_b\mathop{\mathbb{E}}_a\sup_{\alpha:(\alpha,\beta)\in D} a(\alpha,\beta) && a\perp b\\ &\geq \inf_{\beta\in B}\mathop{\mathbb{E}}_a\sup_{\alpha:(\alpha,\beta)\in D} a(\alpha,\beta)\\ &= \inf_{\beta\in B}\mathbb{E}\sup_{\alpha:(\alpha,\beta)\in D} a(\alpha,\beta).\end{aligned}$$

Similarly (but more simply because α_* depends only on a),

$$\begin{aligned}\mathbb{E}\sup_{\beta\in B} b(\alpha_*^a,\beta) &= \mathop{\mathbb{E}}_a\mathop{\mathbb{E}}_b\sup_{\beta\in B} b(\alpha_*^a,\beta) && a\perp b\\ &\geq \inf_{\alpha\in A}\mathop{\mathbb{E}}_b\sup_{\beta:(\alpha,\beta)\in D} b(\alpha,\beta)\\ &= \inf_{\alpha\in A}\mathbb{E}\sup_{\beta:(\alpha,\beta)\in D} b(\alpha,\beta).\end{aligned}$$

□

Corollary 4 (Mean Version for Product Spaces)**.** Suppose Condition (C7),

(C15) Condition (C12) with $D := A \times B$;

(C16) for all $\alpha, \alpha' \in A$ and $\beta \in B$,

$$b(\alpha, \beta) \sim b(\alpha', \beta),$$

and Condition (C13). Then

$$\mathbb{E} \sup_{(\alpha,\beta)\in D} a(\alpha,\beta) + b(\alpha,\beta) \geq \inf_{\beta\in B} \mathbb{E} \sup_{\alpha:(\alpha,\beta)\in D} a(\alpha,\beta) + \inf_{\alpha\in A} \mathbb{E} \sup_{\beta:(\alpha,\beta)\in D} b(\alpha,\beta).$$

Proof. In Lemma 4, one can take A_β, B to be the A, B of this corollary, $\alpha^a_*(\beta)$ a branch of $\operatorname{argmax}_{\alpha\in A} a(\alpha,\beta)$, and $\beta^{a,b}_* \in \operatorname{argmax}_{\beta\in B} b(\alpha^a_*(\beta),\beta)$. Then, by Condition (C16), $a \perp \beta^{a,b}_*$. □

Lemma 5 (Recurrence Lower Bound)**.** For all $a, b \in \mathbb{Z}$, let $[a, b]$ ($(a, b]$) denote the discrete interval $\{a, a+1, \cdots, b\}$ ($\{a+1, \cdots, b\}$, respectively). Suppose Condition (C4) and

(C17) for all $i, j \in [1, N]$, $k \in \{i, j\}$, and $T_0 \in [1, T-1]$,

$$\mathbb{P}\left(\sum_{t=T_0+1}^{T} L_{k,t} = \min_{k_*\in\{i,j\}} \sum_{t=T_0+1}^{T} L_{k_*,t} \right) \geq 1/2.$$

For all $\mathcal{N} \subseteq [1, N]$, let

$$M_{\mathcal{N},T} := \mathbb{E} \max_{n\in\mathcal{N}} \sum_{t=1}^{T} L_{n,t}. \tag{Id2.9}$$

Let $\iota : [1, \lceil N/2 \rceil] \times \{0, 1\} \to [1, N]$. Then, for $N > 1$ and $T_0 \in [0, T]$,

$$M_{[1,N],T} \geq \min_{j\in\{0,1\}} M_{\iota([1,\lceil N/2\rceil]\times\{j\}),T_0} + \min_{i\in[1,\lceil N/2\rceil]^{\{0,1\}}} M_{\iota((i(0),0)\cup(i(1),1)),T-T_0}. \tag{A2.10}$$

Suppose also $N = 2^k$ for some $k \in \mathbb{Z}_+$. Then there exists a re-ordering, say σ, such that for

$$M^\sigma_{\mathcal{N},T} := \mathbb{E} \max_{n\in\mathcal{N}} \sum_{t=1}^{T} L_{\sigma(n),t},$$

$$M_{[1,N],kT} \geq \sum_{t=1}^{k} \min_{\substack{i\in[1,N/2^t],\\ j\in(N/2^t,N/2^{t-1}]}} M^\sigma_{\{i,j\},T}. \tag{A2.11}$$

Proof.

$$\max_{k\in[1,N]} L^{k}_{1:T} \geq \max_{k\in[1,\lceil N/2\rceil]\times\{0,1\}} L^{\iota(k)}_{1:T}$$
$$= \max_{k\in[1,\lceil N/2\rceil]\times\{0,1\}} \left(L^{\iota(k)}_{1:T_0} + L^{\iota(k)}_{T_0.T}\right). \tag{A2.12}$$

By Corollary 2 with $b(k) := L^{\iota(k)}_{T_0.T}$ and $B := \{0,1\}$, Approximation A2.12 is lower bounded by

$$\max_{\alpha\in A} a(\alpha,\beta_*) + \max_{\beta\in B} b(\alpha_*(\beta),\beta) = \max_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,j_*)}_{1:T_0} + \max_{j\in\{0,1\}} L^{\iota(i_*(j),j)}_{T_0.T},$$

where i_* is a branch of $\operatorname{argmax}_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,\cdot)}_{1:T_0}$ as a function of j and

$$j_* \in \operatorname*{argmax}_{j\in\{0,1\}} L^{\iota(i_*(j),j)}_{T_0.T}$$

can otherwise be random.

Therefore,

$$M_{[1,N],T} \geq \mathbb{E} \max_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,j_*)}_{1:T_0} + \mathbb{E} \max_{j\in\{0,1\}} L^{\iota(i_*(j),j)}_{T_0.T}.$$

By Condition (C17), conditioned on i_*, $0 \in \operatorname{argmax}_{j\in\{0,1\}} L^{\iota(i_*(j),j)}_{T_0.T}$ with at least probability $1/2$, and similarly for 1. Thus, j_* can be given a distribution that endows 0 with probability $1/2$. Furthermore, by Condition (C4), conditioned on i_*, $\operatorname{argmax}_{j\in\{0,1\}} L^{\iota(i_*(j),j)}_{T_0.T}$ is independent of $L_{1:T_0}$. Then

$$\begin{aligned}
&\mathbb{E} \max_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,j_*)}_{1:T_0} \\
=&\,\mathbb{E}\,\mathbb{E}\left(\max_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,j_*)}_{1:T_0} \,\middle|\, i_*, L_{1:T_0}\right) \\
=&\,\mathbb{E}\left(\frac{1}{2}\max_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,0)}_{1:T_0} + \frac{1}{2}\max_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,1)}_{1:T_0}\right) \\
=&\,\frac{1}{2}\left(\mathbb{E}\max_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,0)}_{1:T_0} + \mathbb{E}\max_{i\in[1,\lceil N/2\rceil]} L^{\iota(i,1)}_{1:T_0}\right) \\
\geq&\,\min_{j\in\{0,1\}} M_{\iota([1,\lceil N/2\rceil]\times\{j\}),T_0}.
\end{aligned}$$

As for the second term of Approximation A2.10,

$$\begin{aligned}
\mathbb{E}\max_{j\in\{0,1\}} L^{\iota(i_*(j),j)}_{T_0.T} &= \mathbb{E}\,\mathbb{E}\left(\max_{j\in\{0,1\}} L^{\iota(i_*(j),j)}_{T_0.T} \,\middle|\, i_*\right) \\
&\geq \min_{i\in[1,\lceil N/2\rceil]^{\{0,1\}}} \mathbb{E}\left(\max_{j\in\{0,1\}} L^{\iota(i_*(j),j)}_{T_0.T} \,\middle|\, i_* = i\right) \\
&= \min_{i\in[1,\lceil N/2\rceil]^{\{0,1\}}} \mathbb{E}\left(\max_{j\in\{0,1\}} L^{\iota(i(j),j)}_{T_0.T} \,\middle|\, i_* = i\right) \\
&= \min_{i\in[1,\lceil N/2\rceil]^{\{0,1\}}} \mathbb{E}\max_{j\in\{0,1\}} L^{\iota(i(j),j)}_{T_0.T},
\end{aligned} \tag{A2.13}$$

because of independence (via Condition (C4)) of i_* and $L_{T_0.T}$. By Approximation A2.13,

$$\begin{aligned}\mathbb{E}\max_{j\in\{0,1\}} L_{T_0.T}^{\iota(i_*(j),j)} &\geq \min_{i\in[1,\lceil N/2\rceil]^{\{0,1\}}} \mathbb{E}\max_{j\in\{0,1\}} L_{T_0.T}^{\iota(i(j),j)} \\ &= \min_{i\in[1,\lceil N/2\rceil]^{\{0,1\}}} \mathbb{E}\max_{j\in\{0,1\}} L_{1:T-T_0}^{\iota(i(j),j)},\end{aligned}$$

because L_t are distributed identically over time. Therefore, Approximation A2.10 holds.

Now let $k \in \mathbb{Z}_+$, $N = 2^k$, and the horizon be kT in place of T, with the latter replacing T_0.

Under an appropriate ordering σ_{k-1},

$$M_{[1,2^{k-1}],(k-1)T}^{\sigma_1} = \min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{k-1}],\\(2^{k-1},2^k]\end{smallmatrix}\right\}} M_{\mathcal{N},(k-1)T}^{\sigma_1}.$$

If σ_j obeys

$$\begin{aligned}M_{[1,2^{k-1}],(k-1)T}^{\sigma_j} &= \min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{k-1}],\\(2^{k-1},2^k]\end{smallmatrix}\right\}} M_{\mathcal{N},(k-1)T}^{\sigma_j} \\ &\vdots \\ M_{[1,2^j],jT}^{\sigma_j} &= \min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{j}],\\(2^{j},2^{j+1}]\end{smallmatrix}\right\}} M_{\mathcal{N},jT}^{\sigma_j},\end{aligned}$$

then there exists σ_{j-1} obeying $\sigma_{j-1}(i) = \sigma_j(i)$ for $i > 2^j$ and

$$\begin{aligned}M_{[1,2^{k-1}],(k-1)T}^{\sigma_{j-1}} &= \min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{k-1}],\\(2^{k-1},2^k]\end{smallmatrix}\right\}} M_{\mathcal{N},(k-1)T}^{\sigma_{j-1}} \\ &\vdots \\ M_{[1,2^{j-1}],(j-1)T}^{\sigma_{j-1}} &= \min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{j-1}],\\(2^{j-1},2^j]\end{smallmatrix}\right\}} M_{\mathcal{N},(j-1)T}^{\sigma_{j-1}}.\end{aligned}$$

Therefore, by induction, there exists σ_1 such that

$$\begin{aligned}M_{[1,2^{k-1}],(k-1)T}^{\sigma_1} &= \min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{k-1}],\\(2^{k-1},2^k]\end{smallmatrix}\right\}} M_{\mathcal{N},(k-1)T}^{\sigma_1} \\ &\vdots \\ M_{\{1,2\},T}^{\sigma_1} &= \min_{\mathcal{N}\in\{\{1,2\},\{3,4\}\}} M_{\mathcal{N},T}^{\sigma_1}.\end{aligned}$$

By the assumption of identical distributions (Condition (C4)), Condition (C17) applies equally well to each epoch $[1,T]$, $(T,2T]$, $\cdots$, $((k-1)T,kT]$, not only to the last one. Therefore, the reasoning establishing Approximation A2.10 applies to each epoch. Thus, for every $j \in \{2,\cdots,k\}$,

$$\begin{aligned}
&M^{\sigma_1}_{[1,N/2^{k-j}],jT}\\
=&M^{\sigma_1}_{[1,2^j],jT}\\
\geq&\min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{j-1}],\\(2^{j-1},2^j]\end{smallmatrix}\right\}} M^{\sigma_1}_{\mathcal{N},T_0}+\min_{\substack{i\in[1,2^{j-1}],\\ i'\in(2^{j-1},2^j]}} M^{\sigma_1}_{\{i,i'\},jT-T_0}\\
=&\min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{j-1}],\\(2^{j-1},2^j]\end{smallmatrix}\right\}} M^{\sigma_1}_{\mathcal{N},(j-1)T}+\min_{\substack{i\in[1,2^{j-1}],\\ i'\in(2^{j-1},2^j]}} M^{\sigma_1}_{\{i,i'\},T},
\end{aligned}$$

having taken $T_0 := (j-1)T$. Therefore,

$$\begin{aligned}
&M^{\sigma_1}_{[1,2^j],jT}\\
\geq&\min_{\mathcal{N}\in\left\{\begin{smallmatrix}[1,2^{j-1}],\\(2^{j-1},2^j]\end{smallmatrix}\right\}} M^{\sigma_1}_{\mathcal{N},(j-1)T}+\min_{\substack{i\in[1,2^{j-1}],\\ i'\in(2^{j-1},2^j]}} M^{\sigma_1}_{\{i,i'\},T}\\
=&M^{\sigma_1}_{[1,2^{j-1}],(j-1)T}+\min_{\substack{i\in[1,2^{j-1}],\\ i'\in(2^{j-1},2^j]}} M^{\sigma_1}_{\{i,i'\},T},
\end{aligned}$$

by construction of σ_1. By induction, Approximation A2.11 holds. □

In the following, separability of dimension and time in the lower bound occurs at certain scales.

Lemma 6 (Quantized Dimension-Dependent Bound)**.** Suppose the conditions of Theorem 3 and (C17). Then, for all integers $K \geq 1$,

$$M_{\{1,\cdots,2^K\},KT} \geq K\frac{\mathrm{Var}(L_{1,1}-L_{2,1})}{4\sqrt{2}b}\sqrt{T}, \tag{A2.14}$$

in which $M_{\{1,\cdots,2^K\},KT}$ is defined by Identity Id2.9.

Proof. By Lemma 5, to show Approximation A2.14, it suffices to show

$$\min_{\mathrm{card}(\mathcal{N})=2} M_{\mathcal{N},T} \geq \frac{\mathrm{Var}(L_{1,1}-L_{2,1})}{4\sqrt{2}b}\sqrt{T},$$

which is the content of Theorem 3.

□

2.3.4 Game Value Bounds

Once again, consider Game 1.

Proposition 8 (Commutation Lower Bound)**.** Suppose there exists x, y such that $\inf_{f_* \in F} \mathbb{E}\, \mathrm{d}\big(f_*(x^1), y^1\big) < \infty$.

$$\mathrm{Val}_T \geq \sup_{x,y} \left[T \inf_{f_* \in F} \mathbb{E}\, \mathrm{d}\big(f_*(x^1), y^1\big) - \mathbb{E} \inf_{f_* \in F} \sum_{t=1}^{T} \mathrm{d}\big(f_*(x^t), y^t\big) \right].$$

Proof. The finiteness assumption ensures the expectation is additive.

For every $x, y, f_0^t \in F, t \in \{1, \cdots, T\}$;

$$\begin{aligned} \mathbb{E}\, \mathrm{d}\big(f_0^t(x^t), y^t\big) &\geq \inf_{f_* \in F} \mathbb{E}\, \mathrm{d}\big(f_*(x^t), y^t\big) \\ &= \inf_{f_* \in F} \mathbb{E}\, \mathrm{d}\big(f_*(x^1), y^1\big). \end{aligned}$$

Hence,

$$\mathbb{E} \sum_{t=1}^{T} \mathrm{d}\big(f_0^t(x^t), y^t\big) \geq T \inf_{f_* \in F} \mathbb{E}\, \mathrm{d}\big(f_*(x^1), y^1\big).$$

□

Definition 6 (Set of Minimizers)**.** Given Nature's strategy x, y, let $F_{\mathrm{opt}} \subseteq F$ denote the set of minimizers attaining

$$\inf_{f_* \in F} \mathbb{E}\, \mathrm{d}\big(f_*(x^1), y^1\big).$$

Regret can be written as a transformation of a single sum of differences. It turns out this sum can be bounded by one where the terms are centered, for which the following notation is useful.

Definition 7 (Centered Difference)**.** For all x, y and $f_* \in F_{\mathrm{opt}}$, $d_t(f_*, x, y) := \mathbb{E}[\mathrm{d}\big(f_*(x^1), y^1\big)] - \mathrm{d}(f_*(x^t), y^t)$ and

$$D_T(f_*) := \sum_{t=1}^{T} d_t(f_*, x, y).$$

Lemma 7 (Sum of Centered Differences Lower Bound)**.** Suppose x and y are such that

(C18) there exists $f_* \in F_{\mathrm{opt}}$ such that $\mathrm{d}\big(f_*(x^t), y^1\big)$ is integrable.

Then, for all non-empty $F_* \subseteq F_{\mathrm{opt}}$ and $T \in \mathbb{Z}_+$,

$$\mathrm{Val}_T \geq \inf_{f^t} \mathbb{E}\, \mathrm{R}_T\,(f, x, y) \geq \mathbb{E} \sup_{f_* \in F_*} D_T(f_*). \tag{A2.15}$$

If, further,

(C19) $\inf_{f_* \in F} \mathbb{E}\, \mathrm{d}\big(f_*(x^1), y^1\big) \leq \inf_{f_* \in Z^X} \mathbb{E}\, \mathrm{d}\big(f_*(x^1), y^1\big)$,

then

$$\underline{\mathrm{Val}}_T \geq \mathbb{E} \sup_{f_* \in F_*} D_T(f_*).$$

Proof. Suppose the lemma's conditions and let $C := \inf_{f_* \in F_*} \mathbb{E}\, \mathrm{d}\big(f_*(x^t)^1, y^1\big)$. By Condition (C18) and Proposition 8 (regardless of x and y),

$$\begin{aligned}
\mathrm{Val}_T &\geq \mathbb{E}\left[TC - \inf_{f_* \in F} \sum_{t=1}^{T} \mathrm{d}\big(f_*(x^t), y^t\big)\right] \\
&\geq \mathbb{E}\left[TC - \inf_{f_* \in F_*} \sum_{t=1}^{T} \mathrm{d}\big(f_*(x^t), y^t\big)\right] \quad \text{Condition (C20), Approximation A1.4.}
\end{aligned} \tag{A2.16}$$

By Definition 6, for all $f^t \in F_*$,

$$C = \mathbb{E}\, \mathrm{d}\big(f^1(x^1), y^1\big). \tag{Id2.17}$$

Plugging Identity Id2.17 into Approximation A2.16,

$$\mathrm{Val}_T \geq \mathbb{E} \sup_{f^t \in F_*} \sum_{t=1}^{T} [\mathbb{E}\, \mathrm{d}\big(f^1(x^1), y^1\big) - \mathrm{d}\big(f^t(x^t), y^t\big)].$$

□

Low Dimensionality

Remark 1. Under the conditions of Lemma 7, given any two distinct points $f_1^t, f_2^t \in F_{\mathrm{opt}}$, Theorem 3 can be applied with

$$\begin{aligned}
F_* &:= \{f_1^t, f_2^t\} \\
L_{1,t} &:= \mathbb{E}[\mathrm{d}\big(f_1^t(x^1), y^1\big)] - \mathrm{d}\big(f_1^t(x^1), y^1\big) \\
L_{2,1} &:= \mathbb{E}[\mathrm{d}\big(f_2^t(x^1), y^1\big)] - \mathrm{d}\big(f_2^t(x_1^1), y^1\big),
\end{aligned}$$

provided Condition (C3) is satisfied. Indeed, then $L_{1,t}$ and $L_{2,t}$ are centered (integrable by Condition (C18)) and obey Condition (C4). Note two such points may exist even when $X \subseteq \mathbb{R}$, as with the upcoming online linear optimization problem in Section 2.4.2.

High Dimensionality

Given the choice of a space denoted F_*, let κ be its cardinality. To consider dimensionality effects, it turns out that it suffices to derive lower bounds that depend on κ for a space subject to the conditions of Corollary 5. In applications, these are intended to be proved with the aid of Lemma 5 (whose lower bounds are of a form similar to Approximation A2.18). This proof technique is illustrated in the next corollary.

Definition 8 (Support)**.** For a X-valued random variate x, let $\operatorname{supp}(x) := \bigcap_{A \subseteq X : \mathbb{P}(x \in A) = 1} A$.

Corollary 5 (Value Lower Bound)**.** Suppose the conditions of Lemma 7;

(C20) $F_* \subseteq F_{\text{opt}}$;

(C21) $\kappa < \infty$; and

(C22) $g : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{Z}_+$, $h : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{R}$, and w_*, w'_* are such that for all $T \in \mathbb{Z}_+$,

$$\mathbb{E} \sup_{f^t \in F_*} D_T(f_*, x, y) \geq h(\kappa, T) \, \mathbb{E} \max_{w \in \{f_*, f'_*\}} D_{g(\kappa, T)}(f^t). \tag{A2.18}$$

Then, for all $T \in \mathbb{Z}_+$,

$$\operatorname{Val}_T \geq C h(\kappa, T) \sqrt{g(\kappa, T)}, \tag{A2.19}$$

in which

$$\begin{aligned} C &:= \frac{\operatorname{Var}\big(\mathrm{d}\big(f_*(x^1), y^1\big) - \mathrm{d}\big(f'_*(x^1), y^1\big)\big)}{4\sqrt{2} b} \\ b &:= \sup_{f^t \in \{f_*, f'_*\}, x_* \in \operatorname{supp}(x), y_* \in \operatorname{supp}(y)} |\mathrm{d}(f(x_*), y_*)|, \end{aligned} \tag{Id2.20}$$

in which supp denotes the smallest set containing its argument almost surely.[2]

Proof. Given the stated conditions, by Lemma 7, it suffices to show

$$\mathbb{E} \max_{f_* \in \{f_*, f'_*\}} \sum_{t=1}^{T} d_t(f_*, x, y) \geq C\sqrt{T},$$

which follows from Theorem 3, where the necessary bounds do not depend on t due to Condition (C2). □

Corollary 6 (Value Lower Bound for Bounded Growth)**.** Suppose Conditions (C18), (C20), (C21), and

(C23) for all $t \in \mathbb{Z}_+$ and $f, f', f'' \in F_*$ such that $f'' \in \{f, f'\}$,

$$\mathbb{P}\Big(D_T(f'') - D_t(f'') = \min_{k \in \{f, f'\}} D_T(k) - D_t(k)\Big) \geq 1/2.$$

Then, for all $T \in \mathbb{Z}_+$,

$$\operatorname{Val}_T \geq \tilde{C} \min\left\{T, \sqrt{T \log_2(\kappa)}\right\},$$

[2] Recall Condition (C2).

in which

$$\tilde{C} := \frac{\min \mathrm{Var}\big(\mathrm{d}\big(f_*(x^1), y^1\big) - \mathrm{d}\big(f'_*(x^1), y^1\big)\big)}{8 \log_2(3) \sup_{\substack{f \in F_*, \\ x \in X, y \in Y}} |\mathrm{d}(f(x^t), y)|}, \tag{Id2.21}$$

where the minimum is over distinct $f_*, f'_* \in F_*$. If, further, Condition (C19) holds, then, for all $T \in \mathbb{Z}_+$,

$$\underline{\mathrm{Val}}_T \geq \tilde{C} \min\left\{T, \sqrt{T \log_2(\kappa)}\right\},$$

Proof. Indexing the elements of F_* by $\{1, \cdots, \kappa\}$, and applying Lemma 7, bounds the value by $M_{\{1,\cdots,\kappa\},T}$ as defined by Identity Id2.9, with the definitions given by Remark 1.

Condition (C23) satisfies (C17).

Let all logarithms that appear be base 2 and $K := \min\{\log \kappa, T\}$. By Lemma 5 and monotonicity (Lemma 2), for all $\kappa, T \in \mathbb{Z}_+$,

$$\begin{aligned} M_{\{1,\cdots,\kappa\},T} &\geq \lfloor K \rfloor \min_{\mathrm{card}(\mathcal{N})=2} M_{\mathcal{N}, \lfloor \frac{T}{K} \rfloor} \\ &\geq C \lfloor K \rfloor \sqrt{\left\lfloor \frac{T}{K} \right\rfloor}, \end{aligned} \tag{A2.22}$$

by Corollary 5, with

$$\begin{aligned} g(\kappa, T) &= \left\lfloor \frac{T}{K} \right\rfloor \\ h(\kappa, T) &= \lfloor K \rfloor \\ \{f_*, f'_*\} &= \operatorname*{argmin}_{\mathrm{card}(\mathcal{N})=2} M_{\mathcal{N}, \lfloor \frac{T}{K} \rfloor}. \end{aligned}$$

The two factors C and $\lfloor K \rfloor \sqrt{\lfloor \frac{T}{K} \rfloor}$ of Approximation A2.22 will be bounded in turn.

$$\begin{aligned} C &\geq \min_{f_*, f'_* \in F_*} \frac{\mathrm{Var}\big(\mathrm{d}\big(f_*(x^1), y^1\big) - \mathrm{d}\big(f'_*(x^1), y^1\big)\big)}{4\sqrt{2}b} \\ b &\leq \sup_{f_* \in F_*, x \in X, y \in Y} |\mathrm{d}\big(f_*(x^t), y\big)|. \end{aligned}$$

where a max would make sense for $F_{*,N}$ by Condition (C21).

$$\begin{aligned}
&\lfloor K \rfloor \sqrt{\left\lfloor \frac{T}{K} \right\rfloor} \\
&\geq \min\left\{\frac{\log\kappa}{\log 3}, T\right\} \sqrt{\max\left\{\left\lfloor \frac{T}{\log\kappa} \right\rfloor, 1\right\}} \\
&= \min\left\{\frac{\log\kappa}{\log 3}, T\right\} \max\left\{1, \sqrt{\left\lfloor \frac{T}{\log\kappa} \right\rfloor}\right\} \\
&\geq \min\left\{\frac{\log\kappa}{\log 3}, T\right\} \max\left\{1, \sqrt{\frac{T}{2\log\kappa}}\right\} \\
&\geq \min\left\{\frac{\log\kappa}{\log 3}\sqrt{\frac{T}{2\log\kappa}}, T\right\} \\
&= \min\left\{\frac{\sqrt{\log\kappa}}{\sqrt{2}\log 3}\sqrt{T}, T\right\} \\
&\geq \frac{1}{\sqrt{2}\log 3} \min\left\{\sqrt{T\log\kappa}, T\right\}.
\end{aligned}$$

$4\sqrt{2}\sqrt{2} = 8$.

□

2.4 Choice of F_*

For applications to experts and online linear optimization, for a given dimensionality, it suffices to take F_* as the vertices of F (when F can be regarded as a linear space).

2.4.1 Experts

(C24) For a given dimensionality N, Game 1 recovers the experts setting via $X := Z^N$ and F becoming the set of N coordinate projections on X.

This is essentially the setting of [15] (see its p. 7), with special cases in [13][3] (absolute loss) and [47] (linear loss). [32] generalizes certain aspects of the latter (allowing F to depend on time)—see its p. 4. One recovers a special case of the experts setting here by fixing its $N_t, C(i,t)$ by N, i, respectively, for all $i \in \{1, \cdots, N\}$. The case where randomized decisions are allowed is considered in [3].

The following results are meant to help check the conditions of Corollary 6.

[3] Which interestingly uses "ℓ" to denote the horizon.

Proposition 9 (Symmetric Extreme)**.** Suppose $L_1 - L_2$ is symmetrically distributed. Then

$$\mathbb{P}(L_1 = \min\{L_1, L_2\}) \geq \frac{1}{2}.$$

Proof.

$$\begin{aligned}\mathbb{P}(L_1 = \min\{L_1, L_2\}) &= \mathbb{P}(L_1 \leq L_2) \\ &= 1 - \mathbb{P}(L_1 > L_2).\end{aligned}$$

$$\begin{aligned}1 &= \mathbb{P}(L_1 < L_2) + \mathbb{P}(L_1 = L_2) + \mathbb{P}(L_1 > L_2) \\ &= 2\,\mathbb{P}(L_1 > L_2) + \mathbb{P}(L_1 = L_2),\end{aligned}$$

by symmetry. Hence,

$$\mathbb{P}(L_1 > L_2) \leq \frac{1}{2}.$$

□

Proposition 10 (Exchangeability of Identical Distributions)**.** Suppose L_1 and L_2 are identically distributed and that each has a support of no more than two real numbers. Then they are exchangeable.

Proof. If the support is a singleton, then the result is trivial. Suppose the support is two points x_1, x_2.

Because $\mathbb{P}(L_1 = x_1) = \mathbb{P}(L_2 = x_1)$,

$$\begin{aligned}&\mathbb{P}(L_1 = x_1, L_2 = x_1) + \mathbb{P}(L_1 = x_1, L_2 = x_2) \\ &= \mathbb{P}(L_1 = x_1, L_2 = x_1) + \mathbb{P}(L_1 = x_2, L_2 = x_1),\end{aligned}$$

which is equivalent to

$$\mathbb{P}(L_1 = x_1, L_2 = x_2) = \mathbb{P}(L_1 = x_2, L_2 = x_1).$$

□

The following is relevant to the satisfaction of Condition (C23).

Corollary 7. If, for every $t \in \{1, \cdots, T\}$, $L_{1,t}$ and $L_{2,t}$ are identically distributed on two distinct real numbers, then

$$\mathbb{P}\left(\sum_{t=1}^{T} L_{1,t} = \min\left\{\sum_{t=1}^{T} L_{1,t}, \sum_{t=1}^{T} L_{2,t}\right\}\right) \geq \frac{1}{2}.$$

Proof. Proposition 10 provides exchangeability of each pair of tth terms in the sums, so exchanging all of them shows the sums themselves are exchangeable. Exchangeability implies symmetry. Proposition 9 concludes. □

Lemma 8. Suppose Condition (C24) and

(C25) there exists $z, z' \in Z$, $y_0 \in Y$, and strategy y such that (i) y_0 has mass, (ii)

$$\mathbb{E}\,\mathrm{d}(z, y) = \mathbb{E}\,\mathrm{d}(z', y) \leq \inf_{z_* \in Z} \mathbb{E}\,\mathrm{d}(z, y),$$

and (iii) for all $z'' \in \{z, z'\}$, $\mathbb{E}|\mathrm{d}(z'', y)| < \infty$.

Then, letting the components of x be uniformly distributed on $\{z, z'\}$ and $F_* := F$, Conditions (C18)–(C21) and (C23) are satisfied.

Proof. Condition (C24) guarantees (C20) and (C21).

Condition (C25) and the restriction of x to $\{z, z'\}^N$ guarantees Conditions (C18)–(C20).

By Corollary 7, Condition (C24) and the construction of x to be uniform together satisfy (C23). □

Lemma 9. Suppose (C26) there exist distinct $y_0, y_0' \in Z \cap Y$, and (C27) d is a metric on $Z \cup Y$. Then Condition (C25) holds with y uniformly distributed over $\{y_0, y_0'\}$ independently of x, $z := y_0$, and $z' := y_0'$.

Proof. There exists y uniformly distributed over $\{y_0, y_0'\}$. Then

$$\begin{aligned}
\inf_{z \in Z} \mathbb{E}\,\mathrm{d}(z, y) &= \inf_{z \in Z} \frac{1}{2} \left[\mathrm{d}(z, y_0) + \mathrm{d}(z, y_0')\right] \\
&= \inf_{z \in Z} \frac{1}{2} \left[\mathrm{d}(y_0, z) + \mathrm{d}(z, y_0')\right] && \text{Symmetry} \\
&\geq \inf_{z \in Z} \frac{1}{2} \mathrm{d}(y_0, y_0') && \text{Triangle} \\
&= \frac{1}{2} \mathrm{d}(y_0, y_0') \\
&= \frac{1}{2} \left(\mathrm{d}(y_0, y_0') + \mathrm{d}(y_0, y_0)\right) && \text{(Id2.23)} \\
&= \frac{1}{2} \left(\mathrm{d}(y_0, y_0') + \mathrm{d}(y_0', y_0')\right), && \text{(Id2.24)}
\end{aligned}$$

where Identities Id2.23 and Id2.24 are equal to, respectively, $\mathbb{E}\,\mathrm{d}(y_0, y)$ and $\mathbb{E}\,\mathrm{d}(y_0', y)$. □

Proposition 11 (Variance Concerning Losses)**.** Let x, y, z be uniformly and independently distributed on $\{y_0, y_0'\}$ and d a metric. Then

$$\mathrm{Var}(\mathrm{d}(x, z) - \mathrm{d}(y, z)) = \frac{\mathrm{d}(y_0, y_0')}{2}.$$

Proof. Let $B \sim \text{Bernoulli}(1/2)$.

$$\begin{aligned}
&\operatorname{Var}(\mathrm{d}(x,z) - \mathrm{d}(y,z)) \\
&= \mathbb{E} \operatorname{Var}(\mathrm{d}(x,z) - \mathrm{d}(y,z) | z) + \operatorname{Var}(\mathbb{E}(\mathrm{d}(x,z) - \mathrm{d}(y,z) | z)) \\
&= 2\,\mathbb{E} \operatorname{Var}(\mathrm{d}(x,z) | z) \\
&= 2\,\mathrm{d}(y_0, y_0')^2 \operatorname{Var}(B) \\
&= \frac{\mathrm{d}(y_0, y_0')^2}{2}.
\end{aligned}$$

□

Consequently, one has the following.

Theorem 4 (Value Lower Bound)**.** Suppose Conditions (C24) and (C26)–(C27). Then, for all $T \in \mathbb{Z}_+$,

$$\underline{\mathrm{Val}}_T \geq \frac{\mathrm{d}(y_0, y_0')}{16 \log_2 3} \min\left\{T, \sqrt{T \log_2 N}\right\}.$$

Proof. By Lemma 9, Condition (C25) holds with $z := y_0$ and $z' := y_0'$.

Given the satisfaction of Conditions (C24) and (C25), by Lemma 8, the conditions of Corollary 6 ((C18), (C20), (C21), and (C23)) are satisfied.

Re what the latter result says,

$$\tilde{C}_N = \frac{\min_{i \neq j} \operatorname{Var}\big(\mathrm{d}\big(x_i^1, y^1\big) - \mathrm{d}\big(x_j^1, y^1\big)\big)}{8 \log 3 \sup_{\substack{i_* \in \{1, \cdots, N\}, \\ x \in X, y \in Y}} |\mathrm{d}(x_{i_*}, y)|}.$$

Let the components of x be independent so that

$$\min_{i \neq j} \operatorname{Var}\big(\mathrm{d}\big(x_i^1, y^1\big) - \mathrm{d}\big(x_j^1, y^1\big)\big) = \operatorname{Var}\big(\mathrm{d}\big(x_1^1, y^1\big) - \mathrm{d}\big(x_2^1, y^1\big)\big) = \frac{\mathrm{d}(y_0, y_0')}{2},$$

by Proposition 11.

□

For the setting of [13] and [15]'s Theorem 3.7, one can take, in Theorem 4, $Y = \mathbb{R}$, d as absolute loss, $y_0 = 0$, and $y_0' = 1$. Though the resulting constant $\frac{1}{16 \log_2 3}$ is worse than the asymptotically optimal $\frac{1}{\sqrt{2 \log_2(e)}}$, it applies non-asymptotically and in high dimensions. Moreover, this paper's theory is not designed to obtain optimal constants, but rather to be general and simple relative to its generality; obtaining the optimal constant requires more specialized techniques.

2.4.2 Online Linear Optimization with Box Constraints

Corollary 6 recovers, as follows, the minimax behavior of online linear optimization ([1]).[4]

To model linear loss, one can take X to be a singleton or disavow its presence and let $\mathrm{d}(f^t, y) := y(f^t)$.

Lemma 10 (Conditions Satisfaction)**.** Suppose

(C28) $N \in \mathbb{Z}_+$, $F := [-1,1]^N$, $F_* := \{-1,1\}^N$, d is the dot product,

and that there exist distinct $x_0, x_0' \in X$ of opposite signs. Then there exists x^t whose co-ordinates are identically distributed on $\{y_0, y_0'\}$ so that $\mathbb{E}\, y_1^1 = 0$, satisfying Conditions (C18), (C20), (C21), and (C23).

Proof. Re the prescription of y, given any two elements of opposite sign, one can construct a distribution on them having mean 0. The construction of y to have zero expectation guarantees (i) for all $w \in F$, $-\infty < \mathbb{E}\, w \cdot y^1 = \left(\sum_{k=1}^N w_k\right) \mathbb{E}\, y_1^1 = 0 < \infty$, satisfying Condition (C18); and (ii) (C23).

Condition (C28) guarantees (C20) and (C21).

□

Proposition 12 (Variance Concerning Losses)**.** Suppose $y_0, y_0' \in \mathbb{R}$ are nonzero and have opposite signs. Let x be distributed on $\{y_0, y_0'\}$ so that $\mathbb{E}\, x = 0$. Then

$$\mathrm{Var}(x) = -y_0 y_0'$$

Proof. Let $p = \mathbb{P}(x = y_0)$. $py_0 + (1-p)y_0' = 0$ implies $p = \frac{-y_0'}{y_0 - y_0'}$ and $1 - p = \frac{y_0}{y_0 - y_0'}$.

$$\begin{aligned}
\mathrm{Var}(x) &= \frac{1}{y_0 - y_0'} \left(-y_0' y_0^2 + y_0 {y_0'}^2 \right) \\
&= \frac{y_0 y_0'}{y_0 - y_0'} \left(-y_0 + y_0' \right) \\
&= -y_0 y_0'.
\end{aligned}$$

□

Theorem 5 (Value Lower Bound)**.** Suppose the conditions of Lemma 10. Then, for all $T \in \mathbb{Z}_+$,

$$\mathrm{Val}_T \geq \frac{-y_0 y_0'}{\sqrt{2} \max\{|y_0|, |y_0'|\}} N \sqrt{T}. \tag{A2.25}$$

[4] More generally, constrained online linear optimization is known to exhibit $\sqrt{T}$ behavior ([6, 56, 61]; and footnote 3 of [34]). In contrast, see [57] re unconstrained optimization.

Proof. Adopting the x of Lemma 10,

$$\begin{aligned}
&\mathbb{E}\sup_{w\in F_*}\sum_{t=1}^{T}\mathbb{E}[\mathrm{d}\big(wx^1,y^1\big)]-\mathrm{d}\big(wx^t,y^t\big)\\
=&\mathbb{E}\sup_{w\in\{-1,1\}^N}\sum_{t=1}^{T}0-wx^t\\
=&\mathbb{E}\sup_{w\in\{-1,1\}^N}w\sum_{t=1}^{T}-x^t\\
=&\mathbb{E}\sup_{w\in\{-1,1\}^N}\sum_{n=1}^{N}w_n\sum_{t=1}^{T}-x_n^t\\
=&\mathbb{E}\sum_{n=1}^{N}\sup_{w_n\in\{-1,1\}}w_n\sum_{t=1}^{T}-x_n^t\\
=&N\,\mathbb{E}\sup_{w_1\in\{-1,1\}}w_1\sum_{t=1}^{T}-x_1^t, \qquad \text{(Id2.26)}
\end{aligned}$$

because the components of x are identically distributed. Continuing with Identity Id2.26,

$$\begin{aligned}
&N\,\mathbb{E}\sup_{w_1\in\{-1,1\}}w_1\sum_{t=1}^{T}-x_1^t\\
=&N\,\mathbb{E}\sup_{w\in\{-1,1\}^1}\sum_{t=1}^{T}0-wx_1^t\\
=&N\,\mathbb{E}\max_{w\in\{-1,1\}^1}\sum_{t=1}^{T}\mathbb{E}[\mathrm{d}\big(wx_1^1,y^1\big)]-\mathrm{d}\big(wx_1^t,y^t\big).
\end{aligned}$$

Note, by Condition (C28), $\kappa=2^N$. Thus, Condition (C22) is satisfied with

$$\begin{aligned}
h\left(\kappa,T\right)&\equiv\log_2\left(\kappa\right)\\
g\left(\kappa,T\right)&\equiv T\\
f_*&=(1,0,\cdots,0)\\
f_*'&=(-1,0,\cdots,0).
\end{aligned}$$

By Corollary 5,

$$\begin{aligned}
C&=\frac{\mathrm{Var}\big(\mathrm{d}\big(f_*x^1,y^1\big)-\mathrm{d}\big(f_*'x^1,y^1\big)\big)}{4\sqrt{2}\sup_{\substack{w\in\{f_*,f_*'\},\\ x_*\in\mathrm{supp}(x),y_*\in\mathrm{supp}(y)}}|\mathrm{d}(wx_*,y_*)|}\\
&=\frac{\mathrm{Var}\big(2x_1^1\big)}{4\sqrt{2}\max\left\{|x_0|,|x_0'|\right\}}.
\end{aligned}$$

$$\mathrm{Var}\left(2x_1^1\right) = -4x_0' x_0,$$

by Proposition 12, yielding Approximation A2.25.

□

Infinite Dimensions

The following is a special case of Game 1 and a form of online linear optimization.

Given: $T \in \mathbb{Z}_+$, vector space $\mathcal{X}$, $X \subseteq \mathcal{X}$, $W \subseteq \mathcal{X}^*$
for $t = 1, \cdots, T$ **do**

1. Learner decides $w \in W$
2. Nature generates $x^t \in X$
3. Learner loses $(wx)^t$

end

Game 2: Select from Dual

Theorem 6 (Value under Measure Constraints)**.** Suppose X is a subset of $\mathbb{R}^U$ whose elements are integrable including constant functions $y_0 > 0$ and $y_0' < 0$, Σ_U is a σ-algebra on U, $C \geq 0$, and

$$W := \left\{ \begin{array}{c} x \mapsto \int x(u) d\mu(u) : \\ \mu \text{ signed measure on } \Sigma_U \wedge \int d|\mu| = C \end{array} \right\}.$$

Then

$$\mathrm{Val}_T \geq \frac{-y_0 y_0'}{\sqrt{2}\max\left\{|y_0|, |y_0'|\right\}} C\sqrt{T}. \tag{A2.27}$$

Proof. Letting each x^t be identically distributed on $\{x_0, x_0'\}$ so that $\mathbb{E}\, x_1 = 0$,

then

$$
\begin{aligned}
&\mathbb{E} \sup_{w \in W_*} \sum_{t=1}^{T} \mathbb{E}[\mathrm{d}\Big((wx)^{t\,1}, y^1\Big)] - \mathrm{d}\Big((wx)^{t\,t}, y^t\Big) \\
=&\mathbb{E} \sup_{w \in W_*} \sum_{t=1}^{T} 0 - (wx)^{t\,t} \\
=&\mathbb{E} \sup_{w W_*} w \sum_{t=1}^{T} -x^t \\
=&\mathbb{E} \sup_{\int |\mu| = C} \int \left[\sum_{t=1}^{T} -x^t(u) \right] d\mu(u) \\
=&\mathbb{E} \sup_{\int |\mu| = C} \int \left[\sum_{t=1}^{T} -x^t \right] d\mu(u) && x^t \text{ constant} \\
=&C\,\mathbb{E} \left| \sum_{t=1}^{T} -x^t \right| \\
\geq& C \frac{\mathrm{Var}\big(x^1\big)}{\sqrt{2}\|x^1\|_\infty} \sqrt{T} && \text{Corollary 1} \\
=&C \frac{-x_0 x_0'}{\sqrt{2} \max\{|x_0|, |x_0'|\}} \sqrt{T} && \text{Proposition 12.}
\end{aligned}
$$

Lemma 7 concludes.

□

Remark: Theorem 6 generalizes Theorem 5.

Remark: Under additional assumptions, W is the "Riesz representation" of those continuous linear functionals whose operator norm is C.

2.5 Upper Bounds

In designing algorithms, it is usually sufficient to consider the following variant of Game 1, where the Learner decides its prediction function based solely on the past and not current features. This allows the generation turns of the game

to be compressed into a single step:

Given: $T \in \mathbb{Z}_+$; vector space $\mathcal{X}$; sets $X \subseteq \mathcal{X}$, Y, $W \subseteq \mathcal{X}^*$;
$\mathrm{d} : X \times Y \to \mathbb{R}$
for $t = 1, \cdots, T$ **do**

1. Learner decides $w \in W$
2. Nature generates $(x, y)^t \in X \times Y$
3. Learner loses $\mathrm{d}(wx, y)^t$

end

Game 3: Single Generation Step

In proving an upper bound for an adversarial process, one may begin by fixing a deterministic x, y, an element $w_* \in W$, and formulating an algorithm for w. In that way, the regret becomes a deterministic sum

$$\sum_{t=1}^{T} \mathrm{d}(wx, y)^t - \mathrm{d}\big(w_* \cdot x^t, y^t\big) \tag{Id2.28}$$

whose terms have a memory only through w.

The foregoing sum can be abstracted in the form

$$\sum_{t=1}^{T} [f_t(w^t) - f_t(w_*)] = \sum_{t=1}^{T} g_t(w^t, w^{t+1}).$$

The addition of w^{t+1} provides an avenue to bounding each term by those of the form $h_T(w_*, w) - h_T(w_*, w^{t+1}) + c_T(w_*)$, so that the presence of w telescopes away, and with it, memory (except for a fixed number, 2, of terms). That bound

$$f_t(w^t) - f_t(w_*) \leq h_T(w_*, w) - h_T(w_*, w^{t+1}) + c_T(w_*),$$

can be interpreted as approximating a difference between (i) w^t and w_* with that of (ii) the former term and w^{t+1}; $c_T(w_*)$ is then an (approximate) error.

How might finding such a bound be easier than the original problem? By restricting w^{t+1} to be of the form $\Phi(w, f_t)$, the problem can be further abstracted as determining h, c, and Φ so that

$$f(w) - f(w_*) \leq h_T(w_*, w) - h_T(w_*, \Phi(w, f)) + c_T(w_*); \tag{A2.29}$$

for all $f \in \mathcal{F}, w, w_* \in W$ for some function class $\mathcal{F}$. For fixed T, there are no more processes.

$$c_T(w_*) := \sup_{f \in \mathcal{F}, w \in W} [f(w) - h_T(w_*, w) + h_T(w_*, \Phi(w, f)) - f(w_*)]$$

is the optimal solution for c_T. Solving the inequality may be even more straightforward if Φ is given and h_T is designed to minimize $\sup_{w_*} c_T(w_*)$. To accommodate more algorithms, the inequality could be generalized to allow the variables

such as Φ and c_T to be processes, and if their dependence on time is simple, solving the inequality could remain tractable.

If Approximation A2.29 holds, then for all sequences $f_1, \cdots, f_T \in \mathcal{F}$,

$$\begin{aligned}\inf_{w_1}\sup_{w_*}\sum_{t=1}^{T} f_t(w^t) - f_t(w_*) &\leq \inf_{w_1}\sup_{w_*}\left[Tc_T(w_*) - h_T(w_*, w^{T+1}) + h_T(w_*, w^1)\right] \\ &\leq \sup_{w_*}\left(Tc_T(w_*) + \sup_{w_1,w_2\in W}[h_T(w_*, w_1) - h_T(w_*, w_2)]\right).\end{aligned} \tag{A2.30}$$

The following condition on f allows linearization of $f(w)-f(w_*)$, by its very definition.

Definition 9 (Sub-gradient)**.** Suppose W is a subset of a vector space $\mathcal{W}$ (not necessarily a dual space). A **sub-gradient** ∇ of $f \in \mathbb{R}^W$ is an element of $(\mathcal{W}^*)^W$ such that, for all $w_0, w \in W$,

$$\nabla f(w_0)(w - w_0) \leq f(w) - f(w_0),$$

(equivalently,

$$\nabla f(w_0)(w_0 - w) \geq f(w_0) - f(w).)$$

A sub-gradient of f exists iff f is **sub-differentiable**.

2.5.1 Experts

In this section, adopt the following generalized notion of regret:

$$\mathrm{Reg}_T\,(w, x, y) := \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{d}(wx, y)^t - \inf_{w_*\in W_*}\sum_{t=1}^{T}\mathrm{d}\big(w_*\cdot x^t, y^t\big)\right], \tag{Id2.31}$$

for some $W_* \subseteq W$, that is, the constraint set for the infimum is now a parameter. This may or may not coincide with the notion of W_* as a carefully chosen subset of W_{opt} (used in the context of lower bounds). If $\inf_{w_*\in W_*}\sum_{t=1}^{T}\mathrm{d}(w_*\cdot x^t, y^t) < \infty$, Identity Id2.31 is equal to

$$\begin{aligned}&\mathbb{E}\sup_{w_*\in W_*}\left[\sum_{t=1}^{T}\mathrm{d}(wx, y)^t - \sum_{t=1}^{T}\mathrm{d}\big(w_*\cdot x^t, y^t\big)\right] \\ =&\,\mathbb{E}\sup_{w_*\in W_*}\left[\sum_{t=1}^{T}\mathrm{d}(wx, y)^t - \mathrm{d}\big(w_*\cdot x^t, y^t\big)\right].\end{aligned}$$

Theorem 7 (Value Upper Bound)**.** Suppose Condition (C24), d is convex in its first argument (for all values of its second), and non-negative. Then, for all $N, T \in \mathbb{Z}_+$, the following value bound holds.

$$\mathrm{Val}_T \leq \sup_{x_*\in X, y_*\in Y}\mathrm{d}(x_*, y_*)\min\left\{\sqrt{T}, \sqrt{\ln(N)/2}\right\}\sqrt{T}.$$

Proof. One can normalize d without affecting its convexity and then apply Corollary 2.2 of [15].

To prove that corollary, fix an arbitrary sequence of $(X \times Y)^T$ and adopt Hedge with an appropriate learning rate.

For all $n \in \{1, \cdots, N\}$, $t \in \{1, \cdots, T\}$, and given sequence $y^1, \cdots, y^T \in Y$, let $\ell_n^t := \mathrm{d}(x_n, y)^t$ and $\mathrm{Lum}_n^T := \sum_{t=1}^T \ell_n^t$. Hedge mixes expert i with a normalization of the weight w_n^t:

$$w_n^t = e^{-\eta \mathrm{Lum}_n^{t-1}},$$

in which $\eta > 0$ is the **learning rate**. $\eta = \sqrt{8 \ln(N)/T}$ suffices.

The proof is then simply an application of convexity followed by Hoeffding's lemma.

□

Corollary 8 (Competitiveness)**.** Suppose Conditions (C24), (C26)–(C27), that d is convex in its first argument, and that $\mathrm{d}(y_0, y_0') = \sup_{x_* \in X, y_* \in Y} \mathrm{d}(x_*, y_*)$. Then, for all $N, T \in \mathbb{Z}_+$,

$$\mathrm{CD}_T \leq 16 \log 3.$$

Proof. Recalling this generalized notion of regret applies to the lower bound of the previous section, by Theorems 4 and 7,

$$\begin{aligned}
&\frac{\sup_{x_* \in X, y_* \in Y} \mathrm{d}(x_*, y_*) \min\left\{\sqrt{T}, \sqrt{\ln(N)/2}\right\} \sqrt{T}}{\frac{\mathrm{d}(y_0, y_0')}{16 \log 3} \min\left\{T, \sqrt{T \log N}\right\}} \\
&= 16 \log 3 \frac{\sup_{x_* \in X, y_* \in Y} \mathrm{d}(x_*, y_*) \min\left\{\sqrt{T}, \sqrt{\ln(N)/2}\right\} \sqrt{T}}{\mathrm{d}(y_0, y_0') \min\left\{T, \sqrt{T \log N}\right\}} \\
&= 16 \log 3 \frac{\sup_{x_* \in X, y_* \in Y} \mathrm{d}(x_*, y_*) \min\left\{\sqrt{T}, \sqrt{\ln(N)/2}\right\}}{\mathrm{d}(y_0, y_0') \min\left\{\sqrt{T}, \sqrt{\log N}\right\}} \\
&= 16 \log 3 \frac{\sup_{x_* \in X, y_* \in Y} \mathrm{d}(x_*, y_*) \min\left\{\sqrt{T}, \sqrt{\log N/(2 \log e)}\right\}}{\mathrm{d}(y_0, y_0') \min\left\{\sqrt{T}, \sqrt{\log N}\right\}} \\
&\leq 16 \log 3 \frac{\sup_{x_* \in X, y_* \in Y} \mathrm{d}(x_*, y_*) \min\left\{\sqrt{T}, \sqrt{\log N}\right\}}{\mathrm{d}(y_0, y_0') \min\left\{\sqrt{T}, \sqrt{\log N}\right\}} \\
&= 16 \log 3 \frac{\sup_{x_* \in X, y_* \in Y} \mathrm{d}(x_*, y_*)}{\mathrm{d}(y_0, y_0')} \\
&= 16 \log 3.
\end{aligned}$$

□

2.5.2 Online Linear Optimization with Box Constraints

Consider Game 2. Note X can be regarded as a subset of W^*.

Proposition 13 (All Elements of X are Sub-Differentiable)**.** $\nabla : x \mapsto (w \mapsto wx)$ is a sub-gradient.

Proof. As defined, for all $w_0, w \in W$,

$$\nabla w_0 x(w - w_0) = (w - w_0)x = wx - w_0 x.$$

□

Using sub-gradients, it is relatively simple to obtain upper bounds in the finite-dimensional case.

Theorem 8 (Upper Value Bound)**.** Suppose the conditions of Lemma 10 and $|y| \leq \max\{|y_0|, |y_0'|\}$ for all $y \in Y$. Then, for all $N, T \in \mathbb{Z}_+$,

$$\mathrm{Val}_{N,T} \leq \left(2 + \max\left\{y_0^2, {y_0'}^2\right\}\right) N\sqrt{T}.$$

Proof. In the context of [61], W_N plays the role of F, and, as required in that paper, is bounded, closed, and non-empty. $\max_{w,w' \in W_N} \|w - w'\|_2 = \|(2, \cdots, 2)\|_2 = 2\sqrt{N}$ and $\max_{x \in X^N, w \in W_N} \|\nabla x(w)\|_2 = \max\{|y_0|, |y_0'|\}\sqrt{N}$. Thus, by Theorem 1 of [61],

$$\mathrm{Val}_{N,T} \leq \frac{\left(2\sqrt{N}\right)^2 \sqrt{T}}{2} + \left(\sqrt{T} - \frac{1}{2}\right)\left(\max\{|y_0|, |y_0'|\}\sqrt{N}\right)^2.$$

□

Corollary 9 (Competitiveness)**.** Suppose the conditions of Theorem 8 and that $|y_0| = 1 = |y_0'|$. Then, for all $N, T \in \mathbb{Z}_+$,

$$\mathrm{CD}_{N,T} \leq \frac{3}{\sqrt{2}}.$$

Proof. By Theorems 5 and 8,

$$\begin{aligned}\mathrm{CD}_{N,T} &\leq \frac{\left(2 + \max\left\{y_0^2, {y_0'}^2\right\}\right) N\sqrt{T}}{\frac{-y_0 y_0'}{\sqrt{2}\max\{|y_0|, |y_0'|\}} N\sqrt{T}} \\ &\leq \frac{3}{\sqrt{2}}.\end{aligned}$$

□

2.6 Computation

Algorithms that achieve the upper bounds of the previous section—Hedge and gradient descent—can be considered linear time in both N and T. That dependence on the horizon is essentially optimal against adversarial data, yet sub-linearity in the horizon is possible for learning IID data. Consequently, being statistically competitive comes at a computational cost.

2.7 Conclusion

In the most general finite-dimensional case considered herein, the lower bounds are of the form in Approximation A2.19, where κ is a measure of a problem's complexity; and, specifically for the examples, C is independent of κ and

$$h(\kappa, T)\sqrt{g(\kappa, T)}$$

simplifies to bounds of the form

$$O\left(\sqrt{T}\min\left\{\log\kappa, (\log\kappa)^{\alpha}\sqrt{T}\right\}\right),$$

where $\alpha \in \{0, 1\}$ is problem dependent.

The examples were proven constructively in that they prescribed IID strategies (which apply to both the IID and adversarial regimes) for Nature. In certain (convex) cases, the bounds are optimal. For example, in Lemma 9, when the metric is a norm and X, Y are each bounded, descent methods provide a matching upper bound, even when the data is adversarial. The non-convex case is less understood and worth further inquiry. Optimal constants for competitiveness (Corollaries 8 and 9) is an open question.

Chapter 3

A Case Study: Macroeconomic Forecasting

In time series analysis, considerable effort may be put into transforming the data to make it appear more stationary. In macroeconomics, for example, data is typically de-trended and de-seasonalized. Moreover, observables are subject to different transformations from one another, usually on an ad hoc basis. From the perspective of Chapter 2, at best, there may be modestly faster learning. At worst, stationarity might not be achieved and the assumption of it may thwart learning at all.

An alternative ensemble approach is to train the model on different subsets of data, accepting that some data may be too stale or irregular to be of use to the model. Moreover, the ensemble weights can be trained in accordance with the theory of Chapter 2.

A standard assumption in empirical economics is stationarity, either of (transformed) observables, for instance in autoregressive models [60, 10] or, of model variables, for instance in state space models [20]. Stationarity assures all available data is representative of what is to come in the future. However, if one naïvely applies an estimator designed for stationarity to data generated by a non-stationary process, then unrepresentative data may be detrimental to the performance of the estimator.

A problem common to existing approaches to non-stationarity is their lack of adaptability, particularly to frequent changes in the data generating process. I therefore propose simultaneously learning multiple estimation problems that collectively aim to capture the possible nature of the data generating process, including multiple parametric models, the possibility of drift rather than a limited number of breaks, and a data generating process that does not conform to existing distributional models of break point processes (or is otherwise misspecified). This guarantee is the sense in which the mixture is optimal.

The dynamic stochastic general equilibrium (DSGE) U.S. macroeconomy model of [52] (hereafter SW) presents an ideal case study. SW is the first DSGE

model to be observed to be competitive with Bayesian vector autoregression [29], and it did so for the U.S. macroeconomy. Now it is an off-the-shelf standard model and one that is extensible to newer models, such as [25]. Of the cited papers, four are based on SW. Thus, SW is a natural case study of DSGE forecasting. Moreover, its application to post-WWII data for which it was designed is plagued with the problems of FOT selection. SW found the first ten years of their data to be "unrepresentative" and were therefore burdened with coping with this purportedly unrepresentative time frame. Yet, representativeness or lack thereof was not explicitly defined in the paper, nor was their an explanation of how the first-ten-years cutoff was selected, as opposed to another, possibly similar, cutoff. Ultimately, SW discarded the first ten years, restricting the time span of analysis despite initially aiming to analyze all post-WWII data.

Though the mixture automatically handles the entire data set whereas the pure recursive estimator requires manual data deletion, there are other considerations in choosing an estimator, such as parsimony and interpretability of the underlying model upon which the design and guarantees of the estimator are based, statistical-parameter stability, goodness of fit, out of sample performance (such as regret or generalizability) over the entire data set, and out of sample performance post an in-sample-selected FOT. In the interest of time, this document considers only out of sample performance[1], namely squared l2-norm error (loss function) of one-step (quarter)-ahead, point forecasts (prediction task). Out of sample performance is of more transparent significance to model evaluation and to practitioners than the other considerations, and is relatively easy to measure. SW can be used to simulate forecast trajectories of the seven observables on which it is estimated—quarterly GDP, consumption, investment, hours worked, wages, prices, and a short-term interest rate—from which a one-step-ahead distributional forecast can be read off, and the mean taken as the point forecast. As expected, given the absence of manual guidance of the recursive estimator's FOT, the mixture outperforms it. Moreover, this outperformance is dramatic (Figure 3.1); and the mixture even outperforms the recursive estimator estimated from the SW-in-sample-determined FOT (Figure 3.2).

3.1 Case Study: Recursive DSGE Estimation

The components of the mixture estimator, as well as the estimators used in SW, are recursive estimators of a single DSGE model or DSGE for short. DSGEs not only models the macro behavior of an economy (whose internal workings are highly complex and not fully knowable), it models the equilibrium behavior of intertemporally profit and utility maximizing agents that determine this macro behavior. To make the model tractable, the economy is reduced to one with a few eternal agents and goods. Each agent is **representative** of a segment of the economy, for instance assuming the segment is made up of a set of homo-

[1] Out of sample performance is based on a prediction task and a loss function.

[2] The Great Moderation, a period of relatively low sample variance in observables, is taken to begin in 1984, as in SW; and is chosen to end at the beginning of the Financial Crisis.

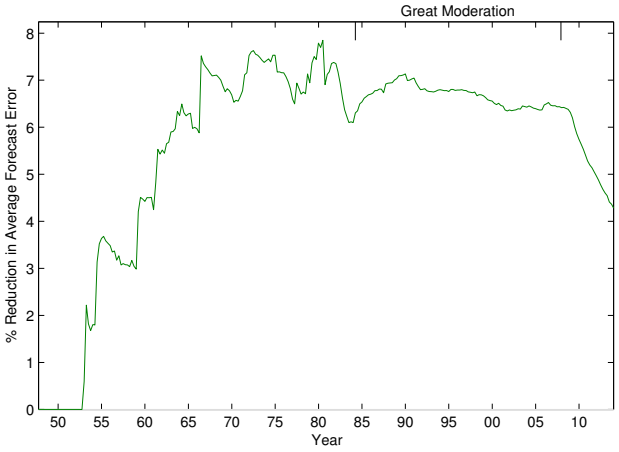

Figure 3.1: **Effect of Mixing**: The mixture's reduction of average forecast error, relative to the recursive estimator estimated from the beginning of the data set, plotted over time. The reduction steadies during the Great Moderation[2], before dipping slightly to a final reduction of 12%.

geneous agents. In SW, these agents are homogeneous final and intermediate goods producers, homogeneous households or consumers who like consuming and dislike working, a representative labor union that seeks higher wages, homogeneous labor packers who purchase the right to labor services and sell it to intermediate goods producers, a central government, and a central bank or monetary authority that sets the interest rate in response to changes in inflation and output.

The various goods are analogously grouped according to type, such as consumables and production inputs, and **aggregated**. In the SW DSGE model, the aggregator functions used are .

The solution of a DSGE gives rise to forward looking rational expectation equations, which are transformed into Markovian backward looking evolution and observation equations.[3] Without a source of random variation (so-called schoks) in the model, the observables would have to obey a deterministic relationship (a so-called stochastic singularity) that is unlikely to be borne out in the data (in fact, if the data were to behave as prescribed by a DSGE, it would almost surely not behave in the stochastically singular manner prescribed by replacing shocks with deterministic functions). In particular, the DSGE equations must contain at least as many shocks as observables. In SW, the model has seven shocks (to go along with its seven observables)—shocks to productivity, the cost of investing in the latest technology, the risk aversion of consumers, wages, prices, government spending (and thus borrowing or tax collection), and the interest rate. Each shock is driven by its own vector autoregression mov-

[3]In SW, the model equations are log linearized.

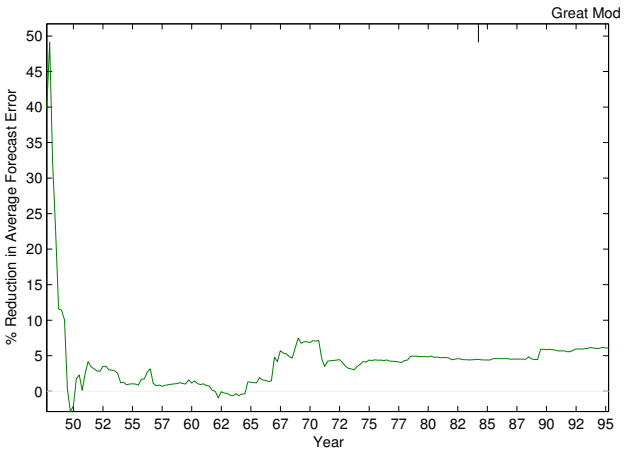

Figure 3.2: **No performance loss in using mixture**. The bar set for the mixture whose training sets include the unrepresentative data was to perform nearly as well as the recursive estimator with FOT 1957. In fact, the mixture outperforms the recursive estimator over the time span used in SW, and beyond. The performance metric is out of sample forecasting squared error.

ing average model with white noise innovations. In particular, the model is stationary.

The equations include real valued statistical parameters, representing aspects of the economy that must be estimated. In particular, SW is of the New Keynsianism or Neoclassical Synthesis family of models, which incorporate economic "frictions" in an otherwise idealized economy, such as sticky prices and wages, and adjustment costs in investment. As static values, the statistical parameters are in particular assumed to not depend on interventions in the economy.

The statistical parameters are endowed with a prior, herein taken from SW. The likelihood is estimated via a Kalman filter, initialized with one year of data. The mode of a model is estimated via optimization of the posterior kernel. The posterior mode for forecasting was previously used in [29] for around 300 estimations performed over two months, which they considered prohibitive for re-running their experiments. Here we must grapple with approximately 1900 estimations per sample path of data. Recursive estimation is implemented in the dynare Matlab/Octave package [2].

3.1.1 Data

The SW observables are based on quarterly GDP, consumption, investment, hours worked, wages, prices, and a short-term interest rate from 1947 – 2004. Eight time series are used to obtain the seven observables (with respect to

which the DSGE models and mixture are estimated): real GDP, personal consumption expenditures, fixed private investment, nonfarm total weekly hours, nonfarm real hourly wages, the GDP deflator, the federal funds rate, and civilian population. GDP defines boundaries of an economy geographically as opposed to by citzenship as in gross national income (GNI). Geography is a physically a more concrete, static criterion, is easier to account for, and consequently GDP measurements will tend to be more meaningful and accurate. Consumption and fixed private investment allows one to distinguish the decision problems of households and producers, which measures a trade off between consumption now and having something better to consume in the future. Fixed private investment (FPI) is of particular importance to the producers' decision problem because it has long-term effects. Consumption and hours worked are complements. Hours worked serves as a productivity input, and because of the representative agent assumption, implicitly captures unemployment. Wages reflect the competition between labor and production for shares of profit. Inflation is used to obtain real dollar values, and is included in the model to account for the fact agents may not just think in terms of real dollars but also nominal dollars. The federal funds rate serves as a riskless discount factor of payoffs in the next quarter. Population affects economic activity and growth, but because it is not part of the model, it is figuratively and literally divided out of the data

The latest revisions are used. Much of the data can be automatically acquired via the FRED API. Otherwise the data can be acquired from the BEA, BLS, or from the SW online data appendix. The civilian population is based on the BLS LNS10000000 series of civilian noninstitutional population; 1976 – 2013 were acquired directly from the BLS, while years prior were obtained from the SW appendix.

Transformations

In the hopes of improving the performance of the recursive estimator and inline with SW, the data is transformed to approximate a data set representative of a stationary process with time homogeneous statistical parameters. The transformations are not prescribed by the model, but guidance is given in SW. Except for the interest rate, the series used are seasonally adjusted. All but the interest rate are log transformed. GDP is first expresed per capita, log differenced, and finally expressed as a percent. Consumption and investment are likewise transformed after first being deflated by the GDP deflator. Wages as reported by the BLS are already averaged over persons (employees). The wages are then GDP deflated, log differenced, and expressed as a percent. To obtain inflation, the GDP deflator is log differenced and expressed as a percent. Hours per capita are adjusted by the employment to population ratio (per [16]), log transformed, and then expressed as a percent. Unlike previous series, transformed hours is not differenced. Hence, to obtain a unitless measure, the initial value of the transformed hours is subtracted from the series. The federal funds rate, which is already expressed as a percent, is simply divided by four.

To ensure that differencing is done on only the data deemed "representative"

by SW and so that exactly ten years pass from the beginning of the differenced data set, the SW in-sample-selected era is taken to begin in 1957Q2. (Aside from the issue about differencing, the removal of data is a transformation of the data.)

3.2 Growing Mixture Estimators

If a DSGE were correctly specified, then an optimal estimator would give equal weight to all available data, or, in other words, from the FOT that coincides with the beginning of available data. However, given the complexity of DSGE modeling, it is impractical for a DSGE to be strictly correctly specified. In particular, the (transformed) data may be non-stationary. SW offered no formal guidelines for moving the FOT forward, or, by how much, despite it requiring removal of over 10% of their data and restricting their time span of analysis by over 10%. Such drastic changes raise questions about their necessity—for instance, even if the time span of analysis should have been moved forward, why is the initial segment of data so much less representative or informative than the prior? As a starting point for analyzing multiple FOTs, consider evaluating the performance of a FOT $t_0 \leq T_0$ given a prediction or estimation task that begins with period T_0 and ends with period T. A natural performance measure is that of the performance of the recursive estimator with FOT t_0, for the given task (over the period T_0 to T).

To make things simple, assume that an estimator makes an estimation or prediction $y_t \in \mathcal{Y}$ entering each period t, subsequently suffering loss $\ell(x_t, y_t)$ on the following observation $x_t \in \mathcal{X}$, as it would for one-step-ahead forecasting. Hence, an estimator recieves immediate, complete feedback. Its performance is then $\sum_{t=T_0}^{T} \ell(x_t, y_t)$. The in-sample optimal FOT t_0 is the one whose sequence of forecasts $y_1^*, \cdots, y_{T_*}^*$ has the minimum $\sum_{t=T_0}^{T} \ell(x_t, y_t^*)$. If one can perform as well as or better than recursive estimator with FOT t_*, then one is doing as well as one could hope for in handling multiple FOTs (at least with respect to the proposed performance measure). The objective of the mixture estimator over different FOTs is therefore to provably approximately achieve or perform better than this in sample optimal forecast error. The recursive estimator with FOT t_* can thus be said to be the "rival" to the mixture estimator.

However, like the statistical parameters of the recursive estimators, the mixture weights are trained with data that precedes the current period. Hence, the rival is not known to the mixture estimator while its weights are being trained. Each candidate FOT corresponds to a potential rival or, in language more familiar to those familiar with no regret learning, an expert. When a new potential rival enters the ensemble, it is said to be born or awakened, and a new **epoch** is said to begin. For a period on or subsequent to its birth, a potential rival is said to be awake. The first observation time used by a potential rival is its birth period. Because the number of first observations (and thus the number of potential rivals to the mixture) grows linearly in the number of observations, only a subset of first observations (and thus potential rivals) is considered.

Given a horizon T, the τth epoch's regret is

$$\sum_{t=t_\tau}^{T} \ell(x_t, y_t) - \min_{e \in \{1,\cdots,\tau\}} \sum_{t=t_\tau}^{T} \ell(x_t, y_t^e)$$

in which ℓ is the loss suffered as a function of the realization and prediction, x_t and y_t respectively are the former and latter in period t, y_t^e is the prediction of the recursive estimator with first observation t_e, $e \in \{1, \cdots, \tau\} =: E_\tau$, t_τ is the beginning of τth epoch, and $\underset{e \in E_\tau}{\operatorname{argmin}} \sum_{t=t_\tau}^{T} \ell(x_t, y_t^e)$ is the τth epoch rival. By bounding a particular epoch's regret, one can account for the number of periods the mixture had full access to E_τ. Moreover, simultaneously bounding the epochs' regrets is sufficient (and, because it is not known which epoch will be closest to binding, necessary) to bound the underperformance of the mixture relative to each rival its considers. If the optimal first observations do not change too rapidly, then the simultaneous bound is sufficient to bound the mixture's underperformance of a recursive estimator with a FOT determined by in sample analysis.

For simplicity, in the implementation there is but a single epoch length that determines the spacings between possible first observations. In the absence of other considerations, we somewhat arbitrarily chose five years for the epoch length. Though limited, a new possible first observation every five years is more flexible than a single first observation for what is over a half century of data.

3.2.1 Algorithm

The mixture weights are learned according to recursive multiplicative weight training (Algorithm 1)[4]. Given an online learning problem specified by a realization space $\mathcal{X}$, a convex subset of a vector space over the reals or prediction space $\mathcal{Y}$, a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and a finite number of experts N; the algorithm is parameterized by a $\rho > 0$ and an $\varepsilon \in (0, 1/2)$; smaller ρ and larger ε result in a faster learning rate, at the expense of a higher likelihood of overfitting.

Algorithm 1 (Multiplicative Weight Training)
Given: $\mathcal{X}$, $\mathcal{Y}$, ℓ, N
Parameters: ρ, ε
Input: $(x_t^1, \cdots, x_t^N) \in \mathcal{X}^N$, $(y_t^1, \cdots, y_t^N) \in \mathcal{Y}^N$, $(w_t^1, \cdots, w_t^N) \in \Delta^{N-1}$
Output: For $i \in \{1, \cdots, N\}$

$$w_{t+1}^i := (1 - \varepsilon)^{\ell(x_t^i, y_t^i)/\rho} w_t^i$$

[4]This form of multiplicative of weight training is slightly different than that of [4], which sets $w_{t+1}^i := (1 + \varepsilon)^{\ell(x_t^i, y_t^i)/\rho} w_t^i$ if $\ell(x_t^i, y_t^i) < 0$ (and is the same as Algorithm 1 otherwise). It has the advantages of a simpler, more intuitive representation; path independence of the realized losses; and invariance under constant shifts of the loss function.

$$\overline{w}_{t+1}^{i} := \frac{w_{t+1}^{i}}{\sum_{j=1}^{N} w_{t+1}^{j}}$$

$$y_{t+1} := \sum_{j=1}^{N} \overline{w}_{t+1}^{j} y_{t+1}^{j}$$

If the elements of $\mathcal{Y}$ are represented by an ordered basis, the mixture forecast of Algorithm 1 can be succinctly written as the matrix product $[y_t^1, \cdots, y_t^N]\overline{w}_t$.

3.2.2 Guarantees

The epochs' regrets can be simultaneously bounded via a dynamic programming approach—in each epoch, the learner merely tries to bound that epoch's regret. Because the number of potential rivals or experts is fixed, this is merely no regret learning. Therefore, the dynamic programming approach can be regarded as a meta-algorithm that takes as an input a regret minimization algorithm parameterized by the number of experts. Under mild assumptions, a regret minimization algorithm for N experts can guarantee

$$\sum_{t=1}^{T} \ell(x_t, y_t) - \max_{e \in \{1,\cdots,n\}} \sum_{t=1}^{T} \ell(x_t, y_t^e) \leq 4\rho \ln(N)\sqrt{T}$$

in which ρ is assumed to bound the loss. In particular, multiplicative weight training realizes this bound.

No regret learning guarantees that as time goes on, the expert's performance approaches or exceeds the performance of the best expert. Merely assuming bounded losses, one obtains a sublinear bound on the regret. (Bounded losses corresponds to an economy not indefinitely collapsing or growing at a super-exponential rate.) The bound is distribution agnostic, in particular about the nature of the stationarity or lack thereof, and therefore applies to tracking the optimal FOT process without knowledge about its distribution.

The assumptions for the estimation of the mixture weights are weaker than the assumptions used for the recursive estimator, which is what makes no regret learning feasible in a non-stationary and possibly adversarial (minimax) setting. To ensure learnability in the form of all epochs' regrets growing sublinearly and to similarly control the quadratic computational complexity of computing the ensemble estimations, it is necessary to expand the epoch lengths for longer horizons (otherwise the epochs' regrets will grow logarithmically).

3.3 Re-estimation over Time, Forecasting, and the Loss Function

Once a DSGE is estimated, it furnishes a distributional forecast via simulation. Taking the mean thereof yields a point forecast, which will be taken as the estimator's forecast. Dealing with point forecasts rather than distributional forecasts skirts the issue of proper scoring.

If $\mathcal{Y}$ and $\mathcal{X}$ are subsets of the same normed vector space, then the loss ℓ can be expressed as a function of the norm of the difference between its arguments. For SW, $\mathcal{Y} = \mathbb{R}^{7\times 1} = \mathcal{X}$ and ℓ is squared l^2 norm error[5]. The periods or steps are quarters and T is 266.

The filtered variables, model statistical parameters, and mixture weights are updated every period. When there is no data from which to estimate a posterior kernel (for each model, the quarter which represents its future first observation), the prior can be used as the estimate. However, if there is also no data upon which to apply the Kalman filter, then the forecast is based on the prior of the steady state. Because the steady states of some observables in the measurement equations are with respect to differenced data as opposed to the levels that are being forecast, the forecasting is begun in the second period of the data set.

3.4 Discussion

The simultaneous regret bounds guarantee convergence at a rate that could be inadequate for the horizons economists consider. On the other hand, the bound is pessimistic in that it is loose and assumes a perfect adversary. Therefore, it is necessary to assess the mixture's empirical performance before any conclusions can be drawn about whether it is a stronger estimator than a pure recursive estimator, and, if so, by how much. To wit, the mixture reduces the cumulative loss of the recursive estimator by 12%, as already shown in Figure 3.1.

Correctly specifying a model for a particular era, and in particular selecting the correct first observation, is impractical. No regret learning is a promising approach for averaging (mixing) among first observations, as opposed to selecting a single one. Estimation of the mixture weights does not require piecewise or pseudo stationarity or an explicit accounting of concept drift or change points, making it robust to misspecification. Robust estimators are particularly important for complex models of complex phenomena, which are harder to estimate and less likely to satisfy standard assumptions, and practitioners, whose decisions can affect millions of risk averse or vulnerable persons. Empirically, a mixture improves prediction accuracy of the SW DSGE model by 12%. Important virtues of the DSGE approach are the interpretability of its statistical parameters, and its ability to capture uncertainty. Mixing DSGEs preserves these attributes.

[5]Squared error is popular for its mathematical simplicity and its accounting of variance.

Experts need not be based on FOTs. For instance, for contexts in which a rolling window of fixed length (or, more concisely, a fixed memory) is used, the experts could use different memories rather than different FOTs. If it is unclear which of a recursive window or a rolling window is more likely to be optimal, the experts could encompass both or even more general combinations of windows. Finally, if the optimal window switches among candidate windows, no regret with respect to experts who switch among windows is possible, based on the switching experts idea from [36], and implemented in [51].

The success of no regret learning suggests further inquiry into estimators that are robust to model misspecification. For instance, regardless of the ambition to develop or faith in stationary models [51], if non-stationary methods dramatically improve the performance of macroeconomic modeling, then they are the de facto state of the art. Indeed, this work demonstrates the utility of non-stationary methods in the analysis of economic data and its complentarity to traditional methods. While the theory here emphasized guarantees for non-stationary data, guarantees for stationary data are likewise of interest.

3.5 Appendix: Implementation Validation

To validate the estimation of the models, the parameter mode estimates were checked to be finite, real numbers and forecasts generated by the cumulative estimations are compared to the forecasts of SW and alternative methods. Checking parameter estimates identified five superfluous model statistical parameters from the original SW model file.

To validate the proper ordering, and more generally assignment, of models to their temporal position in predicting the next quarter, the cumulative and mixture estimators are applied to shifted versions of the data. One would expect that a shift backward one quarter would improve predictive performance, and further shifts in either direction would degrade performance monotonically, whether or not the mixture is estimated on the original or transformed data.

To validate the fixed shares algorithm, it was applied to two sets of transformed predictions for which the behavior of the algorithm is known and then that behavior confirmed. Firstly, such that all but the first model receives the minimum weight. Secondly, such that the predictions are the same and thus the weights gravitate toward 1 over the number of awake models.

3.6 Appendix: Data Plots

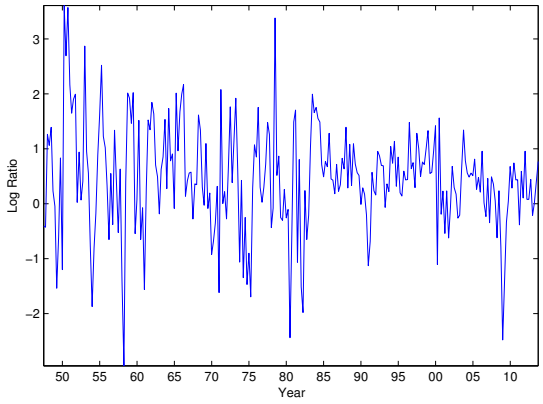

Figure 3.3: Market value of all officially recognized final goods and services in a given quarter. Source: Bureau of Economic Analysis (BEA).

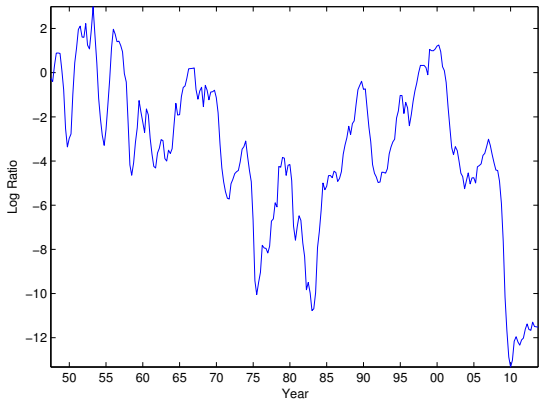

Figure 3.4: Non-farm average hours worked. Source: BLS.

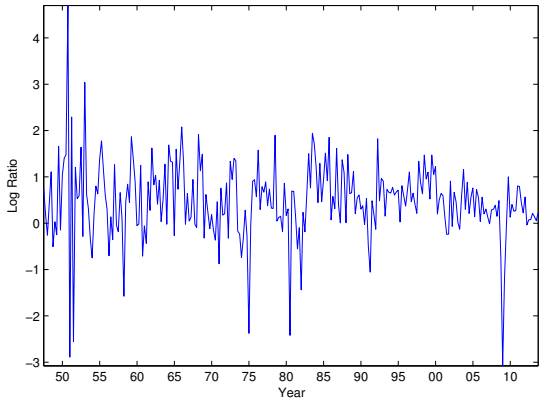

Figure 3.5: Consumer expenditures excluding fixed residential investments. Source: BEA.

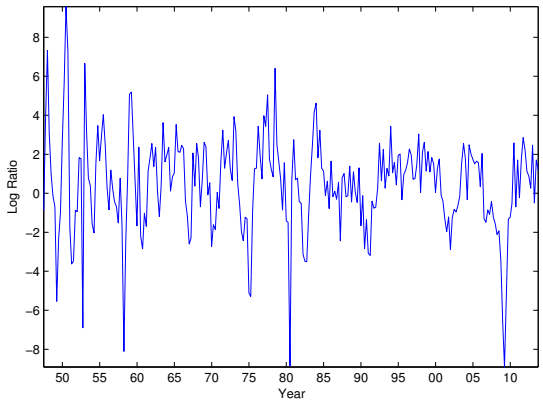

Figure 3.6: Expenditures on reusable production facilities. Source: BEA.

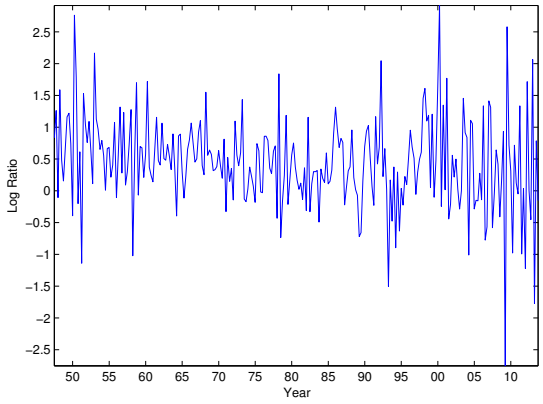

Figure 3.7: Non-farm wages. Source: Bureau Labor Statistics (BLS).

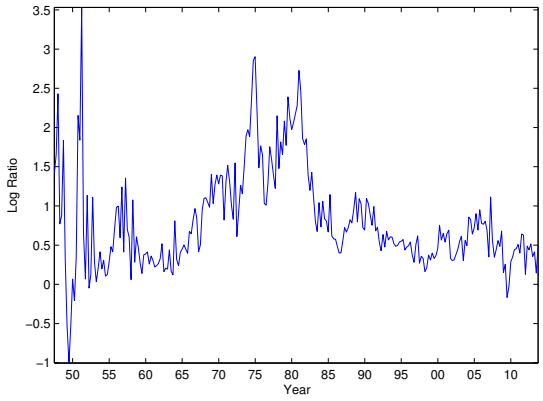

Figure 3.8: Deflator by which to divide to obtain real values. Source: BEA.

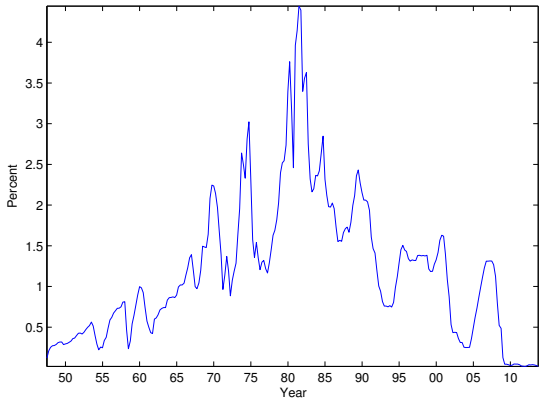

Figure 3.9: Approximately riskless interest rate. Source: Federal Reserve Board.

Chapter 4

Conclusion

Chapter 2 was the main (technical) contribution and focus of this thesis. The other examples of Chapter 1 therefore were presented in only enough detail and context, mostly with historical references, to help motivate competitive analysis as a framework for handling incomplete information and thereby a lens through which to view Chapter 2 and *future* research. Empirical implications were explored in Chapter 3.

4.1 Other Relevant Literature

In Chapter 2, I reviewed the asymptotic lower bound of $O\left(\sqrt{T \ln N}\right)$ ([14, 15]). I provided a non-asymptotic bound of $O\left(\min\left\{T, \sqrt{T \ln N}\right\}\right)$, relying on the novelty of Lemma 5. An analogous recurrence bound for adversarial data is given in [15]. For the sum over a independent and identically-distributed vector process, the usual approach to approximating an extreme sum value is Gaussian approximation, either the entire distribution ([18, 19]) or its moments ([14, 15]). Note this Gaussian approximation need not have a limiting extreme value ([18]). If the sums themselves have a non-Gaussian approximation one could try to exploit that. One might also approximate an extreme via a limiting distribution, when it is guaranteed by the max central limit theorem ([44]). Finally one may compare two different extremes without a full characterization of what either is ([26]). All these techniques are overkill for obtaining growth information on the first moment, and are not known to have validity in as high dimensions. Indeed, if the max sum mean is treated as the object of study, there is a well known matching upper bound for sub-Gaussians for $N = O(e^T)$, which is of more general applicability than the required $N = O\left(e^{T^{1/5}}\right)$ of [26]. The "parametric rate" ([46]), that is the rate for Berry-Esseen, allows at best $N = O\left(e^{T^{1/2}}\right)$; thus, when one wants a Gaussian approximation of the entire distribution in the Kolmogorov distance, that is the highest dimensionality one could hope

for. ([46, 7] come arbitrarily close to reaching this restricted order, in the form $N = O\left(e^{T^{1/2-\varepsilon}}\right)$ for some $\varepsilon > 0$). This $e^{T^{1/2}}$ restriction can also be seen in more abstract, non-Gaussian settings (for example, Proposition 4.1(b) of ([43])). In some settings, stronger restrictions are believed necessary ([19, 17]).

4.1.1 Literature on Other Multiple Regularity-Settings

[5] considers both adversarial and independent and identically-distributed data in the bandit setting. [15, 40] consider online Bayesian algorithms.

Bounds that Scale in the Regularity

One theoretical innovation not covered at length within this dissertation are (adaptive) bounds that scale in regularity measures of the realized data itself, rather than the generating process. A simple example is an algorithm that need not know by what constant an aspect of the learning problem is bounded, as with the gradient descent algorithm used in Theorem 8, whose performance depends on the gradients' magnitudes observed.

Other forms of regularity for which adaptive bounds are available include small losses ([15]), predictability ([48]), curvature ([39]), and low rank ([35])

The algorithm covered in the next section also enjoys adaptive bounds.

Regime Detection

Though in general it is impossible to learn that data is not adversarial, it is possible to learn that it is not perfectly predictable. The FlipFlop algorithm of ([24]) is in the spirit of regime detection in that it switches its behavior according to past data, first presuming that bounded regret is possible. In contrast to the work here, this paper does not treat the high-dimensional setting and (probably necessarily) requires more complicated algorithms than ones merely trying to achieve low regret. [22] also modifies its behavior according to what it perceives to be the regime, but for online convex optimization.

Regret Lower Bounds in the IID Setting

Often lower bounds are proven in the independent and identically-distributed setting out of analytical convenience rather than as an end separate to an adversarial lower bound. This proof technique occurs in, for example, [14, 15, 58, 59, 61, 6, 56].

4.1.2 Luckiness Principle

The informal luckiness principle ([33]) states that a learner should allow itself the opportunity to be lucky. Given possible states of Nature that afford such opportunities, the luckiness principle could be formalized and operationalized by the learner seeking to be competitive when $\mathcal{D}$ contains those states. This

principle could perhaps be further captured by supplementing worst-case analyses with best-case ones. An idea similar to the principle is be "optimistic" ([48]).

4.1.3 Online-to-Batch Conversion

Online-to-batch conversion relates the online setting considered in this dissertation to more classical batch learning. A low-regret algorithm is run on the data and its predictions or hypotheses are combined to form a prediction or hypothesis that generalizes. For instance, [50] considers combinations via (i) randomization and (ii) averaging. These combinations enjoy an expected generalization bound under independent and identically-distributed data that depends on the cumulative loss of the online algorithm. In contrast, generalization guarantees based on a single, terminal instantaneous loss do not exist against an adversary. Nonetheless one could hope to provide simultaneous guarantees of generalizablity under independent and identically-distributed data and low regret under adversarially generated data. Since online-to-batch conversion is designed specifically for batch learning, I pose the question and leave open whether the combinations used in the literature are sub-optimal for obtaining simultaneous online guarantees.

4.2 Future Work

One could hope given the connections among problems in Chapter 1 that there would be results of common applicability. Proposition 2 and the discussion following are steps in that direction. This line of work can be pursued further, though another issue that must be addressed is how that particular list sorting problem to which the theorem applies relates to the modern literature.

Though this dissertation is suggestive of the merits (and computational limitations) of competitive algorithms (as summarized in the next section), its implications do not necessarily generalize to the many important settings not explicitly considered. A viable albeit extensive research program is the competitive analyses of the extension to unknown regularity of every existing online learning setting in the literature that has not yet been analyzed so, or at least for which no lower bounds under IIDness are known. The next sub-section discusses one avenue of generalizing to new settings. Also of interest is studying the theoretical properties of certain heuristic techniques, such as product of experts ([37, 38]) and other methods from deep learning, in the online adversarial setting.

4.2.1 Unbounded Case

Chapter 2 was based on boundedness, beginning with Proposition 3.

The following is an unbounded analog and generalization of that proposition to begin an extension to the unbounded case.

Proposition 14 (Expectation of Square Root: Unbounded Case)**.** Suppose $\mathbb{P}(L \geq 0) = 1$ and $\mathbb{P}(L = 0) < 1$. Then $\|L\|_\infty > 0$ and, for all $C \geq 0$,

$$\mathbb{E}\sqrt{L} \geq \frac{\mathbb{P}(L \leq C)\mathbb{E}\left(L \middle| L \leq C\right)}{\sqrt{C}} + \sqrt{C}\,\mathbb{P}(L > C).$$

The following is an analog of Corollary 1.

Proposition 15. Suppose the conditions of Theorem 2 and that L_1 is not almost surely 0. Then $\|L_1\|_\infty > 0$ and, for all $C \geq 0$,

$$\mathbb{E}\left|\sum_{t=1}^{T} L_t\right| \geq \frac{\mathbb{P}\left(\sum_{t=1}^{T} L_t^2 \geq C\right)}{\sqrt{2}}\sqrt{C}.$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left|\sum_{t=1}^{T} L_t\right| \\
&\geq \frac{1}{\sqrt{2}}\,\mathbb{E}\sqrt{\sum_{t=1}^{T} L_t^2} && \text{Theorem 2} \\
&\geq \frac{1}{\sqrt{2}}\,\mathbb{P}\left(\sum_{t=1}^{T} L_t^2 \geq C\right)\sqrt{C} && \text{Markov.}
\end{aligned}$$

□

Corollary 1 helped prove the $\sqrt{T}$ lower bounds in Chapter 2, and Proposition 15 suggests the necessary techniques extend to the unbounded case. Further generalizations to acquire the necessary machinery for proving a generalized value lower bound is future work.

4.2.2 Upper Bounds

A rationale for focusing on lower bounds was given in Section 2.1. Nonetheless, to establish, competitiveness for infinite-dimensional online linear optimization, an upper bound is necessary (and one that matches Theorem 6 would suffice).

It would be nice to have matching upper bounds even in the non-convex case for experts under metric loss, in some sense giving a comprehensive solution to that problem. This is conceivable iff Learner is allowed to randomize. If there is such a matching with the main assumption be metric loss, then I would expect a geometric characterization.

One can stay in the deterministic decision framework and allow randomization by regarding the instantaneous loss as an expectation over decisions. (The resulting notion of regret might be called pseudo-regret, since it results in an expectation inside the infimum.) Doing so would simplify the analysis and be a reasonable first step.

4.2.3 Empirical Studies

Based on empirical forecasting tests, low-regret learning appears to help cope with non-stationarity. Nonetheless, a more persuasive case for the popular use of Chapter 3's methodology could be made if it offered state-of-the-art performance for DSGE estimation. This could only be ascertained with a more up-to-date literature review and might require implementing a newer DSGE model than SW, if only as a basis of comparison.

On the other hand, recent work on DSGEs for the U.S. macroeconomy seem to be focused on accounting for the Great Recession ([42, 45]). The resulting models may perform better on historical data but (i) would not seem to address the underlying issue of macroeconomists' fallibility in constructing predictive models and (ii) would introduce the possibility of over-fitting or at least their performance not generalizing. Testing estimation methodologies, ones that are based on statistical, machine learning and data science principles rather than any particular problem to which they are applied, would give a less biased evaluation of relative predictive performance. It may be preferable these tests are performed with an old economic model so that it can be tested on data that succeeded its creation. This perspective was not adopted in Chapter 3 but might be in future work.

4.2.4 Probability Charges

Even when a measurable space is countable, there can be dramatic differences in the behavior of so-called probability charges and standard probabilities ([54]). A charge replaces the axiom of countably additivity with the finite kind and thus need only behave like a probability on finite and co-finite sets. One theoretical curiosity is whether the additional freedom afforded to Nature in the use of charges could increase a competitive difference such as minimax regret (or other learning objective value).

4.3 Lessons Learned

Competitive analysis can be used to design algorithms or systems to be adaptive or otherwise competitive. That this analysis does not require statistical uncertainty quantification or other full problem specification lends itself to being more robust than methods which require a fully specified model or unique solution.[1] For learning under unknown regularity, competitiveness (i) corresponds to being protected against adversarial data yet doing commensurately better on regular data and (ii) is achievable. Namely, there exist competitive algorithms for (a) learning expert ensembles under metric loss and (b) online convex optimization. Ultimately, the adoption and deployment of competitive analysis and algorithms lead to better performance, as illustrated in macroeconomic forecasting. The

[1] Regularization is also known to address the latter, and when used in conjunction with competitive analysis, the former.

only apparent caveat is computation. If an algorithm with sublinear complexity in the horizon is required, it may be necessary to design for more regular data.

Bibliography

[1] Abernethy, J., Bartlett, P. L., Rakhlin, A., and Tewari, A. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the nineteenth annual conference on computational learning theory* (2008).

[2] Adjemian, S., Bastani, H., Juillard, M., Mihoubi, F., Perendia, G., Ratto, M., and Villemot, S. Dynare: Reference manual, version 4. Tech. rep., Dynare Working Papers 1, CEPREMAP, 2011.

[3] Altschuler, J., and Talwar, K. Online learning over a finite action set with limited switching. *arXiv preprint arXiv:1803.01548* (2018).

[4] Arora, S., Hazan, E., and Kale, S. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing 8*, 1 (2012), 121–164.

[5] Auer, P., and Chiang, C.-K. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Annual Conference on Learning Theory* (2016), pp. 116–120.

[6] Balandat, M., Krichene, W., Tomlin, C., and Bayen, A. Minimizing regret on reflexive banach spaces and nash equilibria in continuous zero-sum games. In *Advances in Neural Information Processing Systems* (2016), pp. 154–162.

[7] Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *The Annals of Statistics 46*, 6B (2018), 3643–3675.

[8] Bentley, J. L., and McGeoch, C. C. Worst-case analyses of self-organizing sequential search heuristics. Tech. rep., Carnegie Mellon University, 1983.

[9] Bentley, J. L., and McGeoch, C. C. Amortized analyses of self-organizing sequential search heuristics. *Communications of the ACM 28*, 4 (1985), 404–411.

[10] Box, G. E., Jenkins, G. M., and Reinsel, G. C. *Time series analysis: forecasting and control*. John Wiley & Sons, 2011.

[11] Boyar, J., Favrholdt, L. M., and Larsen, K. S. Relative worst-order analysis: A survey. *arXiv:1802.07080* (2018).

[12] Candes, E. J. Modern statistical estimation via oracle inequalities. *Acta numerica 15* (2006), 257–325.

[13] Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., and Warmuth, M. K. How to use expert advice. *Journal of the ACM (JACM) 44*, 3 (1997), 427–485.

[14] Cesa-Bianchi, N., Gaillard, P., Lugosi, G., Stoltz, G., et al. Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems* (2012), pp. 989–997.

[15] Cesa-Bianchi, N., and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

[16] Chang, Y., Gomes, J. F., and Schorfheide, F. Learning-by-doing as a propagation mechanism. *The American Economic Review 92*, 5 (2002), 1498–1520.

[17] Chen, X. Gaussian and bootstrap approximations for high-dimensional u-statistics and their applications. *The Annals of Statistics 46*, 2 (2018), 642–678.

[18] Chernozhukov, V., Chetverikov, D., and Kato, K. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics 41*, 6 (2013), 2786–2819.

[19] Chernozhukov, V., Chetverikov, D., and Kato, K. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability 45*, 4 (2017), 2309–2352.

[20] Commandeur, J. J., and Koopman, S. J. *An introduction to state space time series analysis*. Oxford University Press, 2007.

[21] Cover, T. Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory 14*, 1 (1968), 50–55.

[22] Cutkosky, A., and Boahen, K. A. Stochastic and adversarial online learning without hyperparameters. In *Advances in Neural Information Processing Systems* (2017), pp. 5059–5067.

[23] Cutkosky, A., and Orabona, F. Black-box reductions for parameter-free online learning in banach spaces. *arXiv preprint arXiv:1802.06293* (2018).

[24] De Rooij, S., Van Erven, T., Grünwald, P. D., and Koolen, W. M. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research 15*, 1 (2014), 1281–1316.

[25] Del Negro, M., and Schorfheide, F. Dsge model-based forecasting. *Available at SSRN 2018451* (2012).

[26] Deng, H., and Zhang, C.-H. Beyond gaussian approximation: bootstrap for maxima of sums of independent random vectors. *arXiv preprint arXiv:1705.09528* (2017).

[27] Donoho, D. L. Unconditional bases are optimal. Tech. Rep. 410, Stanford, 1992.

[28] Donoho, D. L., and Johnstone, J. M. Ideal spatial adaptation by wavelet shrinkage. *biometrika 81*, 3 (1994), 425–455.

[29] Edge, R. M., Gürkaynak, R. S., Reis, R., and Sims, C. A. How useful are estimated dsge model forecasts for central bankers?[with comments and discussion]. *Brookings Papers on Economic Activity* (2010), 209–259.

[30] Efron, B. Tweedies formula and selection bias. *Journal of the American Statistical Association 106*, 496 (2011), 1602–1614.

[31] Ferger, D. Optimal constants in the marcinkiewicz-zygmund inequalities. *Statistics & Probability Letters 84* (2014), 96–101.

[32] Gofer, E., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Regret minimization for branching experts. In *Conference on Learning Theory* (2013), pp. 618–638.

[33] Grünwald, P. D. *The minimum description length principle*. MIT Press, 2007.

[34] Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning 69*, 2-3 (2007), 169–192.

[35] Hazan, E., Koren, T., Livni, R., and Mansour, Y. Online learning with low rank experts. In *Conference on Learning Theory* (2016), pp. 1096–1114.

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

[37] Hinton, G. E. Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN)* (1999), IEE.

[38] Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation 14*, 8 (2002), 1771–1800.

[39] Huang, R., Lattimore, T., György, A., and Szepesvari, C. Following the leader and fast rates in linear prediction: Curved constraint sets and other regularities. In *Advances in Neural Information Processing Systems* (2016), pp. 4970–4978.

[40] Kakade, S. M., and Ng, A. Y. Online bounds for bayesian algorithms. In *Advances in neural information processing systems* (2005), pp. 641–648.

[41] Karlin, A. R., Manasse, M. S., Rudolph, L., and Sleator, D. D. Competitive snoopy caching. *Algorithmica 3*, 1-4 (1988), 79–119.

[42] Kehoe, P. J., Midrigan, V., and Pastorino, E. Evolution of modern business cycle models: Accounting for the great recession. *Journal of Economic Perspectives 32*, 3 (2018), 141–66.

[43] Koike, Y. Mixed-normal limit theorems for multiple skorohod integrals in high-dimensions, with application to realized covariance. *arXiv preprint arXiv:1806.05077* (2018).

[44] Leadbetter, M. R., Lindgren, G., and Rootzén, H. *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 2012.

[45] Lindé, J. Dsge models: still useful in policy analysis? *Oxford Review of Economic Policy 34*, 1-2 (2018), 269–286.

[46] Lopes, M. E., Lin, Z., and Mueller, H.-G. Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional data analysis. *arXiv preprint arXiv:1807.04429* (2018).

[47] Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. *Algorithmic game theory*. Cambridge University Press, 2007.

[48] Rakhlin, S., and Sridharan, K. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. 2013, pp. 3066–3074.

[49] Rosenkrantz, D. J., Stearns, R. E., and Lewis, P. M. Approximate algorithms for the traveling salesperson problem. In *Switching and Automata Theory, 1974., IEEE Conference Record of 15th Annual Symposium on* (1974), IEEE, pp. 33–42.

[50] Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning 4*, 2 (2012), 107–194.

[51] Shalizi, C. R., Jacobs, A. Z., Klinkner, K. L., and Clauset, A. Adapting to non-stationarity with growing expert ensembles. *arXiv preprint arXiv:1103.0949* (2011).

[52] Smets, F., and Wouters, R. Shocks and frictions in us business cycles: A bayesian dsge approach. *The American Economic Review 97*, 3 (2007), 586–606.

[53] Sofman, B. Online learning techniques for improving robot navigation in unfamiliar domains. Tech. rep., Carnegie Mellon University, 2010.

[54] Spece, M., and Kadane, J. B. Prime residue class of uniform charges on the integers. *Journal of Theoretical Probability* (2018), 1–21.

[55] Srebro, N., Sridharan, K., and Tewari, A. On the universality of online mirror descent. In *Advances in neural information processing systems* (2011), pp. 2645–2653.

[56] Sridharan, K., and Tewari, A. Convex games in banach spaces. In *COLT* (2010).

[57] Streeter, M., and Mcmahan, B. No-regret algorithms for unconstrained online convex optimization. In *Advances in neural information processing systems* (2012), pp. 2402–2410.

[58] Vovk, V. Competitive on-line linear regression. *Advances in Neural Information Processing Systems* (1998), 364–370.

[59] Vovk, V. Competitive on-line statistics. *International Statistical Review/Revue Internationale de Statistique* (2001), 213–248.

[60] Whittle, P. Hypothesis testing in time series analysis-hafner. *New York, 1951.–136 p* (1951).

[61] Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (2003), pp. 928–936.