A Theory of Selective Prediction

Mingda Qiao
mqiao@stanford.edu

Gregory Valiant
valiant@stanford.edu*

Abstract

We consider a model of *selective prediction*, where the prediction algorithm is given a data sequence in an online fashion and asked to predict a pre-specified statistic of the upcoming data points. The algorithm is allowed to choose when to make the prediction as well as the length of the prediction window, possibly depending on the observations so far. We prove that, even without *any* distributional assumption on the input data stream, a large family of statistics can be estimated to non-trivial accuracy. To give one concrete example, suppose that we are given access to an arbitrary binary sequence $x_1, \ldots, x_n$ of length n. Our goal is to accurately predict the average observation, and we are allowed to choose the window over which the prediction is made: for some $t < n$ and $m \leq n - t$, after seeing t observations we predict the average of $x_{t+1}, \ldots, x_{t+m}$. This particular problem was first studied in [Dru13] and referred to as the "density prediction game". We show that the expected squared error of our prediction can be bounded by $O(\frac{1}{\log n})$ and prove a matching lower bound, which resolves an open question raised in [Dru13]. This result holds for any sequence (that is not adaptive to when the prediction is made, or the predicted value), and the expectation of the error is with respect to the randomness of the prediction algorithm. Our results apply to more general statistics of a sequence of observations, and we highlight several open directions for future work.

1 Introduction

Consider the following prediction problem: each day you observe the stock market, and at some point within the next n days, you must make a prediction about the average return, or average volatility, of the stock market over some (future) period of time. Crucially, *you get to choose both the timepoint within the n days when you make the prediction, as well as the interval over which your prediction spans.* Without any distributional assumptions on the daily movements in the stock market, is it possible to accurately make such a prediction about the future? As we show, the answer is "yes", and the expected error of the prediction tends to zero as n—the length of the window in which the prediction must occur—tends to infinity, assuming an absolute bound on the magnitude of daily fluctuations.

We consider several new angles to this age-old problem of making an accurate prediction about the future, given access to a sequence of observations. The setting we consider abstracts three crucial properties of the above prediction problem: 1) We make no distributional assumptions about the sequence of observations. 2) The sequence, while possibly adversarial, is not adaptive, and is chosen independently of our prediction and when we make it. 3) We decide both when to

*This work is supported by NSF awards CCF-1704417 and AF:1813049 and by an ONR Young Investigator award (N00014-18-1-2295).

make our prediction, as well as the duration over which our prediction spans (provided that both occur within some pre-specified horizon, denoted by n in the example above).

In some sense, this model can be viewed as an exploration of the power that comes with being able to decide when to make a prediction about the future, in a world which, while possibly adversarial and changeable, is indifferent to your predictions (i.e. adversarial but non-adaptive). As such, it captures a number of important and natural online prediction tasks, beyond the toy example of stock-market predictions.

A general formalization of this selective prediction problem can be framed as follows. We are given a family of n functions $(f_1, \ldots, f_n)$ where each $f_m : \mathcal{X}^m \to \mathbb{R}$. The prediction procedure proceeds as the following game. A sequence $x \in \mathcal{X}^n$ of length n is chosen adversarially at the beginning of the game. The prediction game proceeds in n rounds. At each time step $t \in \{0, \ldots, n-1\}$, the player can make a claim in the following form: the function value of the next m entries of the sequence ($1 \leq m \leq n-t$) is $\hat{\alpha}$. In this case, the game terminates immediately and the player incurs a loss of $\ell(\hat{\alpha}, \alpha)$, where $\alpha = f_m(x_{t+1}, \ldots, x_{t+m})$ is the actual function value on $x_{t+1}, \ldots, x_{t+m}$. Two natural loss functions that we focus on are the squared loss $\ell_2(\hat{\alpha}, \alpha) = (\hat{\alpha} - \alpha)^2$ and the absolute loss $\ell_1(\hat{\alpha}, \alpha) = |\hat{\alpha} - \alpha|$. If the player does not make a prediction at time t, the next data point x_{t+1} is revealed to the player and the game continues. The player must predict exactly once before the data sequence is entirely observed.

Facing an arbitrary and possibly adversarial data sequence, the predictor is only entitled the power of choosing the window over which the prediction is made. This power is indeed minimal in the sense that if the adversary knows in advance either the time step t at which a prediction is made or the window length m, the predictor cannot achieve a non-trivial loss even for the task of predicting the arithmetic mean; see Section 2.2 for more details.

This setting, and a related setting where one is must make a prediction about a single timestep, were first considered in [Dru13]. These models deviate significantly from many other prediction settings, which typically either make strong distributional assumptions on the sequence of observations (e.g., that they are drawn independently, or generated from a Markov model, Hidden Markov Model, or exchangeable sequence, etc.), or make no assumptions but quantify the accuracy in terms of some notion of "regret" with respect to a limited set of benchmarks. Additionally, most previously studied prediction settings assume that the predictor must make a prediction at a specified time, or must make predictions at *every* time step. We discuss these differences, and connections to other settings more in Section 1.2.

1.1 Overview of Results

Estimating the arithmetic mean. We first state our main results on the concrete task of predicting the average of a bounded real-valued sequence.

Theorem 1.1. *Suppose that $\mathcal{X} = [0, 1]$ and the function family (f_m) is the arithmetic mean, i.e.,*

$$f_m(x_1, \ldots, x_m) = \frac{1}{m} \sum_{i=1}^{m} x_i.$$

There exists a prediction algorithm that achieves an expected squared loss of $O(\frac{1}{\log n})$ on any sequence of length n. Moreover, this bound is tight: there is a distribution over sequences of length n for which no algorithm can achieve an expected loss better than $\Omega(\frac{1}{\log n})$.

The upper bound of $O(\frac{1}{\log n})$ was first given in [Dru13], and the matching lower bound resolves one of the main open questions posed in that work. At an intuitive level, the mean estimation algorithm follows from the observation that a sequence cannot have a high variance on both small and large scales: if an adversary generates a uniformly random sequence in $\{0,1\}^n$ in the hope that each single data point is hard to predict, the average of the whole sequence would concentrate around $\frac{1}{2}$ and thus be predictable.

The lower bound proof amounts to constructing a sequence with moderate variance at all different scales, simultaneously. Hence, no matter when the prediction algorithm chooses to make a prediction, and no matter the chosen time window, there will be a significant amount of variance in the values, conditioned on the sequence up to the time of prediction. Consequently, the prediction algorithm has no hope in achieving too small a loss.

Estimating smooth functions. The positive result extends to other function families beyond the arithmetic mean. One such function family is the collection of all Lipschitz functions with respect to the earth mover's distance defined as follows. For a real sequence $(x_i)_{i=1}^m$ of length m, let $\mathcal{U}(x)$ denote the uniform distribution on the multiset $\{x_1, \ldots, x_m\}$, i.e., $\mathcal{U}(x)$ assigns probability mass $\frac{1}{m}\sum_{i=1}^m \mathbb{I}[x_i = x]$ to each x.

Definition 1.2. *The earth mover's distance* $\mathrm{EMD}(x, y)$ *between two real sequences* x *and* y *is defined as the Wasserstein distance between* $\mathcal{U}(x)$ *and* $\mathcal{U}(y)$ *with respect to the metric* $d(a, b) = |a-b|$.

Definition 1.3. *A function* $f : \mathbb{R}^m \to \mathbb{R}$ *is* L*-smooth if and only if it is* L*-Lipschitz in earth mover's distance, i.e.,* $|f(x) - f(y)| \le L \cdot \mathrm{EMD}(x, y)$ *for any* $x, y \in \mathbb{R}^m$.

We show that on bounded sequences, smooth functions can be estimated up to an absolute loss of $O\left(\frac{L}{\sqrt{\log n}}\right)$, where L is the smoothness parameter and n is the length of the input sequence.

Theorem 1.4. *Suppose* $\mathcal{X} = [0, 1]$ *and every function in* (f_m) *is* L*-smooth. There exists a prediction algorithm that achieves an expected absolute loss of* $O\left(\frac{L}{\sqrt{\log n}}\right)$ *on any sequence of length* n.

Estimating concatenation-concave functions. In addition to the positive result on smooth functions, which only applies to functions on $\mathbb{R}^m$, we consider the following class of *concatenation-concave* functions that admit a more general domain.

Definition 1.5. *A function family* $(f_m : \mathcal{X}^m \to \mathbb{R})_{m=1}^n$ *is concatenation-concave if and only if for any* $x \in \mathcal{X}^{m_1}$ *and* $y \in \mathcal{X}^{m_2}$ *with* $m_1 + m_2 \le n$, *it holds that*

$$f_{m_1+m_2}(x, y) \ge \frac{m_1}{m_1 + m_2} f_{m_1}(x) + \frac{m_2}{m_1 + m_2} f_{m_2}(y),$$

where $f_{m_1+m_2}(x, y)$ *is a shorthand for* $f_{m_1+m_2}(x_1, \ldots, x_{m_1}, y_1, \ldots, y_{m_2})$.

Note that the arithmetic mean is concatenation-concave, with all inequalities in the above definition being equalities. Another family of concatenation-concave functions of practical importance is the following "learnability" function. Suppose that $\mathcal{L}$ is a given model class, which can be equivalently viewed as a family of bounded loss functions mapping $\mathcal{X}$ to $[0, 1]$. The learnability of a data sequence $(x_1, \ldots, x_m)$ is defined as $\inf_{\ell \in \mathcal{L}} \frac{1}{m}\sum_{i=1}^m \ell(x_i)$, the minimum average loss when we fit the sequence using a model in class $\mathcal{L}$.

The learnability function is not captured by the family of smooth functions in the previous paragraph—in fact, $\mathcal{X}$ may not even be associated with a non-trivial metric. On the other hand, it can be easily verified that the learnability function is concatenation-concave.

Our positive result for concatenation-concave functions states that any bounded concatenation-concave function can be estimated with an expected squared loss of $O(\frac{1}{\log n})$. This result is especially striking when considered in the context of estimating learnability, as the prediction accuracy is independent of the complexity of model class $\mathcal{L}$.

Theorem 1.6. *Assuming that the function family (f_m) is concatenation-concave and bounded in $[0,1]$, there exists a prediction algorithm that achieves an expected squared loss of $O(\frac{1}{\log n})$ on any sequence of length n.*

Fitting unseen data. Given that we can accurately estimate the learnability of future data with respect to *any* model class, it is natural to ask whether we can identify a model that actually fits the unseen data well. To this end, we consider the following generalization of our prediction model: instead of predicting $f_m(x_{t+1}, \ldots, x_{t+m})$, the predictor is required to output a model $\hat{\ell}$ in $\mathcal{L}$ that fits $x_{t+1}, \ldots, x_{t+m}$ well. The setting remains selective in the sense that t and m are still chosen by the prediction algorithm. The loss of the prediction is defined as the excess risk

$$\frac{1}{m}\sum_{i=1}^{m} \hat{\ell}(x_{t+i}) - \inf_{\ell \in \mathcal{L}} \frac{1}{m}\sum_{i=1}^{m} \ell(x_{t+i}).$$

By our results on mean estimation and a standard uniform convergence argument over $\mathcal{L}$, we can easily obtain an $O\left(\sqrt{\frac{|\mathcal{L}|}{\log n}}\right)$ upper bound on the optimal excess risk. Note that in this prediction task, the loss bound indeed depends on the cardinality of $\mathcal{L}$. In classic learning theory, however, the dependence of the excess risk on $|\mathcal{L}|$ is typically logarithmic. It remains a compelling open question whether the excess risk can be further improved to $O\left(\sqrt{\frac{\log|\mathcal{L}|}{\log n}}\right)$ as classical learning theory suggests, or whether a polynomial dependence on $|\mathcal{L}|$ is inevitable in the worst case.

1.2 Related Work

Most closely related to this paper is the work of [Dru13], which studies several prediction problems in the setting where we are given access to an arbitrary (adversarial) infinite binary sequence, and attempt to predict the value of a single index, or predict the fraction of 1's in a future interval. Crucially, the predictor is also allowed to choose the prediction window selectively. [Dru13] shows that given a horizon of length $2^{O(1/\epsilon)}$, one can achieve a squared error of at most ϵ in expectation, which translates into an expected squared loss of $O(\frac{1}{\log n})$ in our setting. Our work recovers this result as a special case, and proves a matching lower bound which implies that this exponential dependence on $1/\epsilon$ is necessary.

The recent work [FKT17] proves a *local repetition lemma*, which states that a sufficiently long sequence must exhibit a certain level of pattern at some time scale. The difference from our work is the interpretation of this observation: while [FKT17] addresses the online learning setting where the regret is defined with respect to a set of "stateful" policies that can be represented by state machines, we consider the problem of directly predicting an arbitrary sequence and aim to generalize this observation to a broader class of prediction and learning tasks.

More broadly, sequential prediction and decision making is a major subject of research in many different fields. Early study on this problem dates back to the pioneering work of [Han57] in the 1950s. This problem, along with many of its extensions, is addressed under various terminologies in different communities, including "universal prediction" in information theory [FMG92], "universal portfolios" in mathematical finance [Cov91, CO96, BK99] and "online learning" in machine learning theory [LW94, CBFH+97, CBL06]. In particular, our approach is closely related to yet different from the online learning formulation. In online learning, the predictor has access to a class of strategies (also known as "experts"). The prediction algorithm leverages the expert advice and makes sequential prediction on *every* time step. The performance of the predictor is measured in terms of the regret, defined as the difference between the incurred loss and the loss of the best expert in hindsight.

There is also a large body of work on "conformal prediction" in the online setting where datapoints are revealed one at a time (see e.g. the book [VGS05]). This body of work is largely concerned with understanding how confidently one can make a prediction about the label, y_t, given x_t and a sequence of labeled data $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$. In general, strong positive results exist in the independent setting where data is drawn independently from a fixed distribution, and also in the more general setting where the sequence of data is assumed to be exchangeable.

The selective prediction model we consider is significantly different from the above two settings. In contrast to the regret minimization framework, we do not restrict ourselves to a specified family of experts; instead, we evaluate the predictor solely based on the expected loss rather than the loss relative to the best expert. In contrast to work on conformal prediction, our results hold without any distributional assumptions on the sequence of data. Crucially, to enable these strong results, in our model the prediction algorithm is allowed to be *selective* in the sense that its prediction may not necessarily cover the entire time horizon, and the prediction can be made over an interval of arbitrary length instead of a single observation.

We note that the recent work of [SKLV18] addresses the problem of predicting the distribution of the next observation in the data sequence from a different perspective. The focus of their work is whether accurate prediction can be made using a small memory, and their results apply to the scenario where the data stream is drawn from a distribution with bounded mutual information between the past and the future (for example, a sequence generated by a hidden Markov model). In contrast, our model captures the prediction of a more general family of statistics of the upcoming observations, and we make no distributional assumptions on the sequence.

Another related line of research concerns the estimation of learnability given limited data. In more detail, given labeled data drawn i.i.d. from an underlying distribution, we are asked to estimate how well a given model class can fit the distribution. It is shown that for linear models, a sample of size $O(\sqrt{d})$ is sufficient for accurate estimation [Dic14, KV18], and this is much less than the amount of data needed to learn a linear model. Our work is incomparable to this line of research, since our results apply to the more general setting where the data are not assumed to be i.i.d. and the model class $\mathcal{L}$ can be arbitrary.

2 Tight Loss Bounds for Mean Estimation

We start by studying a special case of the general prediction problem: estimating the mean of a bounded sequence. Without loss of generality, we assume that the instance space is $\mathcal{X} = [0, 1]$. The function value on a subsequence of numbers is simply the arithmetic mean, i.e., $f_m(x_1, \ldots, x_m) =$

$\frac{1}{m}\sum_{i=1}^{m} x_i$.

2.1 Selective Predictor with Vanishing Loss

We begin by presenting the simple prediction scheme from [Dru13] that achieves an error which goes to zero as n tends to infinity, and include a slightly simpler proof of the $O(\frac{1}{\log n})$ loss. In the following, we assume that the sequence length n is a power of two. Let $\mathcal{U}(S)$ denote the uniform distribution over the finite set S.

Algorithm 1: Selective Prediction

Input: Sequence $x \in \mathcal{X}^n$ of length $n = 2^k$.

$k' \leftarrow \mathcal{U}([k])$;

$t \leftarrow \mathcal{U}(\{0, 2^{k'}, 2 \cdot 2^{k'}, \ldots, n - 2^{k'}\})$;

Observe $x_1, x_2, \ldots, x_{t+2^{k'-1}}$;

$\hat{\alpha} \leftarrow f_{2^{k'-1}}(x_{t+1}, \ldots, x_{t+2^{k'-1}})$;

Predict that $f_{2^{k'-1}}(x_{t+2^{k'-1}+1}, \ldots, x_{t+2^{k'}})$ equals $\hat{\alpha}$;

Algorithm 1 chooses the prediction window by drawing k' and t randomly at the beginning. Then, at time $t + 2^{k'-1}$, the algorithm predicts that the average of the next $2^{k'-1}$ numbers is close to that of the most recent $2^{k'-1}$ numbers. We prove in the following that Algorithm 1 achieves a squared loss of $O(\frac{1}{\log n})$.

Lemma 2.1. *Suppose that the instance space is $\mathcal{X} = [0, 1]$ and the function family f is the arithmetic mean. For any integer $k \geq 1$, Algorithm 1 achieves an expected squared loss of at most $\frac{1}{k}$ on any sequence of length 2^k.*

Remark 2.2. *Lemma 2.1 directly implies that $O(\frac{1}{\log n})$ squared loss can be achieved in the general case that n is not a power of two, thus proving the upper bound part of Theorem 1.1. Indeed, choosing $k = \lfloor \log_2 n \rfloor$ and running Algorithm 1 as if the sequence is of length 2^k gives an expected squared loss of at most $\frac{1}{\lfloor \log_2 n \rfloor} = O(\frac{1}{\log n})$.*

Proof. For integer $k \geq 1$ and $\mu \in [0, 1]$, let $L(k, \mu)$ denote the maximum expected squared loss that Algorithm 1 incurs on a sequence of 2^k numbers between 0 and 1 with average μ. We prove by induction on k that $L(k, \mu) \leq \frac{4\mu(1-\mu)}{k}$, which directly implies the proposition.

When $k = 1$, Algorithm 1 reduces to predicting that $x_2 = x_1$, and the squared loss can be bounded as follows:

$$L(1, \mu) = \sup_{\substack{x_1, x_2 \in [0,1] \\ x_1 + x_2 = 2\mu}} (x_1 - x_2)^2 = \min(4\mu^2, 4(1-\mu)^2) \leq 4\mu(1-\mu).$$

For $k \geq 2$, we note that with probability $\frac{1}{k}$, Algorithm 1 chooses $k' = k$ and predicts that the last 2^{k-1} numbers have the same average as the first 2^{k-1} numbers. Let μ_1 and μ_2 denote the averages of the first and last 2^{k-1} numbers, respectively. Then, the squared loss in in this case is given by $(\mu_1 - \mu_2)^2$. With probability $\frac{k-1}{k}$, the algorithm chooses some $k' < k$ and the algorithm is equivalent to running the same algorithm either on either the first 2^{k-1} numbers or the last 2^{k-1} numbers. By the induction hypothesis, the conditional expected squared loss is upper bounded by

$$\frac{L(k-1, \mu_1) + L(k-1, \mu_2)}{2} \leq \frac{2\mu_1(1-\mu_1) + 2\mu_2(1-\mu_2)}{k-1}.$$

Based on the above analysis, we have

$$L(k,\mu) \le \sup_{\substack{\mu_1,\mu_2\in[0,1]\\ \mu_1+\mu_2=2\mu}} \left[\frac{1}{k}\cdot(\mu_1-\mu_2)^2 + \frac{k-1}{k}\cdot\frac{2\mu_1(1-\mu_1)+2\mu_2(1-\mu_2)}{k-1}\right]$$
$$= \frac{1}{k}\cdot \sup_{\substack{\mu_1,\mu_2\in[0,1]\\ \mu_1+\mu_2=2\mu}} \left[2(\mu_1+\mu_2)-(\mu_1+\mu_2)^2\right] = \frac{4\mu(1-\mu)}{k},$$

which completes the proof. □

2.2 Selectivity is Necessary

Algorithm 1 is selective in the sense that it randomly chooses the time step t as well as the window length m for its prediction. Such selectivity is crucial to achieving a sub-constant loss. Intuitively, if t is known to the adversary, the data stream can be chosen such that the first t elements are independent of the rest, rendering any meaningful prediction unfeasible. Likewise, if the prediction window is of fixed length m, the data sequence can be constructed as blocks of size $m/2$, which also leads to a constant lower bound on the prediction loss. Finally, if the time, t, of the prediction can be chosen, but the window must contain the remaining $n-t$ observations, a constant lower bound also exists. The formal proof of the following proposition is deferred to Appendix A.

Proposition 2.3. *Suppose that prediction algorithm $\mathcal{A}$, when running on a sequence of length n, either: (1) always predicts at the same time t, (2) always chooses the same window length m, or (3) chooses t, but must make a prediction over the entire window of $n-t$ remaining timesteps. Then, there exists a binary sequence of length n on which $\mathcal{A}$ incurs an expected squared loss of at least $\frac{1}{64}$.*

2.3 Matching Lower Bound

The prediction scheme in Algorithm 1 may appear not to leverage all the power of the predictor; indeed, the algorithm chooses the prediction window at the beginning of the algorithm, while the model in general allows the algorithm to make the decision adaptively. Nevertheless, we show in the following that such adaptivity brings little marginal gain—the upper bound in Lemma 2.1 is optimal up to a constant factor.

The key in our lower bound proof is to construct a sequence that simultaneously satisfies an anti-concentration property on both small and large timescales. Such a sequence guarantees that even after the predictor observes a prefix of the sequence, the average of the future data sequence still has a large conditional variance given the prefix. This implies a lower bound on the expected squared error achievable by any prediction algorithm.

Again, we focus on the case that $n = 2^k$ is a power of two, as the proof can be extended to the general case (losing at most a constant factor) by the same argument as in Remark 2.2. Consider a perfect binary tree with n leaves. In the following, we assign a real value between 0 and 1 to each node in the tree recursively, and the sequence $x \in [0,1]^n$ is chosen as the values on the n leaves. Let $\delta = \frac{1}{2\sqrt{k}}$. The value of the root is defined as $\frac{1}{2}$. Then, for each node at the j-th level of the tree (the root being at level 0 and leaves at level k), we choose its value randomly and independently

from $\frac{1}{2} \pm \sqrt{j} \cdot \delta$ such that the expectation of the value equals the value of its parent. In particular, if the parent has value $\frac{1}{2} + \sqrt{j-1} \cdot \delta$, the node takes value

$$\begin{cases} \frac{1}{2} + \sqrt{j} \cdot \delta, & \text{with probability } \frac{\sqrt{j}+\sqrt{j-1}}{2\sqrt{j}}, \\ \frac{1}{2} - \sqrt{j} \cdot \delta, & \text{with probability } \frac{\sqrt{j}-\sqrt{j-1}}{2\sqrt{j}}; \end{cases}$$

the probabilities are switched in the other case. Note that by our choice of δ, all leaves will be assigned values in $[0, 1]$ and thus the resulting sequence is bounded. See Figure 1 for a realization of the construction when $k = 3$. Also see Figure 2 for plots of a sample sequence for $k = 20$. Note that after taking the moving average at different scales, the sequence still exhibits strong anti-concentration. (In contrast, the moving average of a uniformly random bit string would concentrate around $\frac{1}{2}$ at larger scales.)

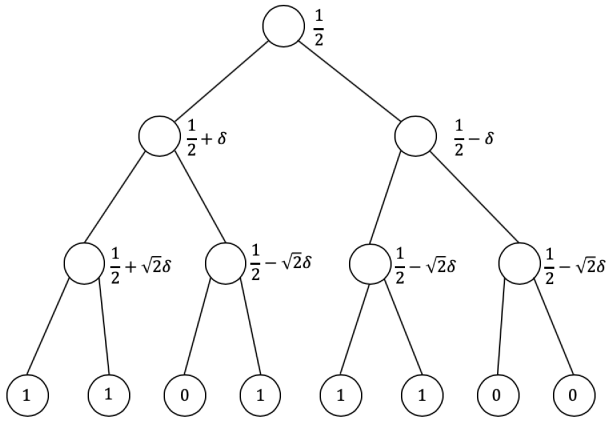


Figure 1: Sample construction for $k = 3$, where $\delta = \frac{1}{2\sqrt{3}}$.

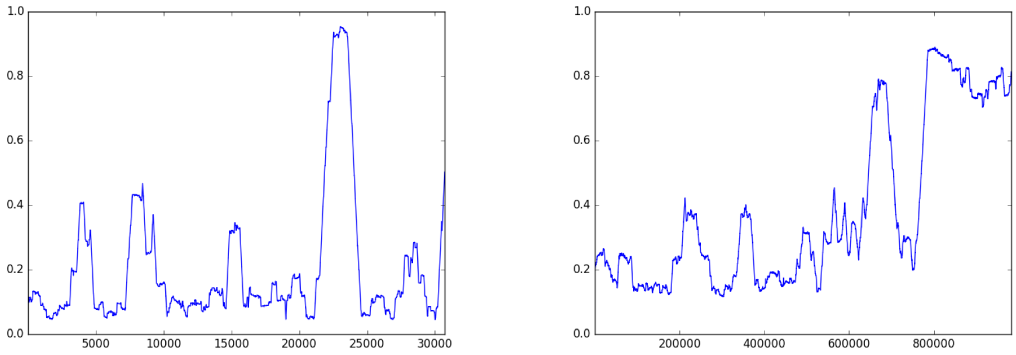
Let $\mathcal{D}_n$ denote the distribution of the sequence that we defined as above. We show that any algorithm will incur an $\Omega(\frac{1}{\log n})$ squared loss in expectation given a random sequence drawn from $\mathcal{D}_n$. By an averaging argument, there exists a sequence on which the algorithm incurs an $\Omega(\frac{1}{\log n})$ squared loss. This proves the lower bound part of Theorem 1.1.

Lemma 2.4. *For any integer $k \geq 1$ and $n = 2^k$, any prediction algorithm for the arithmetic mean incurs an expected squared loss of at least $\frac{1}{64k}$ on a random sequence drawn from $\mathcal{D}_n$.*

Remark 2.5. *Since the arithmetic mean is both 1-smooth (Definition 1.3) and concatenation-concave (Definition 1.5), Lemma 2.4 implies that an $\Omega(\frac{1}{\log n})$ squared loss is inevitable for these two function families.*

Proof. We show that for any $t \in \{0, \ldots, n-1\}$ and $m \in [n-t]$, conditioned on any prefix $x_1, \ldots, x_t$ of the sequence, the variance in $\frac{1}{m}\sum_{i=1}^{m} x_{t+i}$ is at least $\frac{1}{64k}$. The theorem follows from the observation that this is the smallest expected squared loss that can be achieved when the player decides to predict the average of $x_{t+1}, \ldots, x_{t+m}$.

Fix t and m. By our construction of $\mathcal{D}_n$, there exists integer k' such that $2^{k'} \geq \frac{m}{4}$ and $(x_{t+1}, \ldots, x_{t+m})$ contains a contiguous subsequence $(x_{t'+1}, \ldots, x_{t'+2^{k'}})$ of length $2^{k'}$ that exactly



(a) Window length 2^{10}

(b) Window length 2^{15}

Figure 2: Depiction of our lower bound construction for $k = 20$. For one sequence constructed according to our construction, the two plots each depict the average value in a moving window of a certain length, w, over the first 30 (non-overlapping) windows of size w. In the left plot, the window length is $w = 2^{10}$ and in the right plot, $w = 2^{15}$. The fact that both plots look like similar stochastic processes, and in particular exhibit similar anti-concentration behaviors illustrates the main property of our construction: a single underlying sequence that is anti-concentrated, simultaneously, at all timescales.

corresponds to the $2^{k'}$ leaves in a subtree of height k'. Let u denote the root of the subtree. We can actually prove a stronger claim: the variance in $\frac{1}{m}\sum_{i=1}^{m} x_{t+i}$ is lower bounded by $\frac{1}{64k}$, even when conditioned on the values of *all* nodes in the binary tree except the subtree rooted at u.

Let v be the parent of u. Let p_u and p_v denote the values of u and v respectively. It can be verified from the construction that $\mathrm{Var}\left[p_u | p_v\right] = \delta^2 = \frac{1}{4k}$. Since the subtree rooted at u has $2^{k'} \geq \frac{m}{4}$ leaves, the value of node u contributes at least a $\frac{1}{4}$ fraction to the average $\frac{1}{m}\sum_{i=1}^{m} x_{t+i}$. It follows that the conditional variance in $\frac{1}{m}\sum_{i=1}^{m} x_{t+i}$ is lower bounded by $\left(\frac{1}{4}\right)^2 \cdot \mathrm{Var}\left[p_u | p_v\right] = \frac{1}{64k}$. □

3 Estimating General Functions

We extend the positive results for mean estimation to more general function families. It turns out that Algorithm 1 has a stronger guarantee beyond mean estimation: we will show that exactly the same algorithm also achieves a vanishing loss on smooth functions and concatenation-concave functions.

3.1 Smooth Functions

Recall that Algorithm 1 chooses k' and t randomly, and then uses $f_{2^{k'-1}}(x_{t+1}, \ldots, x_{t+2^{k'-1}})$ as an estimate for $f_{2^{k'-1}}(x_{t+2^{k'-1}+1}, \ldots, x_{t+2^{k'}})$. We show in the following that the sequences $x_{t+1}, \ldots, x_{t+2^{k'-1}}$ and $x_{t+2^{k'-1}+1}, \ldots, x_{t+2^{k'}}$ are close in earth mover's distance defined as in Definition 1.2. The prediction loss can then be bounded using the smoothness of f.

Lemma 3.1. *Suppose that $\mathcal{X} = [0, 1]$ and every function in (f_m) is L-smooth. For any integer $k \geq 1$, Algorithm 1 achieves an expected absolute loss of at most $\frac{L}{\sqrt{k}}$ on any sequence of length 2^k.*

Lemma 3.1 implies Theorem 1.4 by the argument in Remark 2.2.

Proof. Let $\mathcal{I}_{k',t}$ and $\mathcal{J}_{k',t}$ denote subsequences $x_{t+1}, \ldots, x_{t+2^{k'-1}}$ and $x_{t+2^{k'-1}+1}, \ldots, x_{t+2^{k'}}$. In the following, we prove the an upper bound on the expected earth mover's distance between $\mathcal{I}_{k',t}$ and $\mathcal{J}_{k',t}$:

$$\mathbb{E}\left[\mathrm{EMD}(\mathcal{I}_{k',t}, \mathcal{J}_{k',t})\right] \le \frac{1}{\sqrt{k}},$$

where the expectation is taken over the randomness in k' and t.

It is well-known that the earth mover's distance between two distributions on $[0,1]$ can be rewritten as

$$\mathrm{EMD}(\mathcal{I}_{k',t}, \mathcal{J}_{k',t}) = \int_0^1 \left|\mathcal{U}(\mathcal{I}_{k',t})([0,\tau]) - \mathcal{U}(\mathcal{J}_{k',t})([0,\tau])\right| \,\mathrm{d}\tau.$$

Recall that $\mathcal{U}(\mathcal{I}_{k',t})$ (resp. $\mathcal{U}(\mathcal{J}_{k',t})$) denotes the uniform distributions naturally defined by $\mathcal{I}_{k',t}$ (resp. $\mathcal{J}_{k',t}$), i.e., $\mathcal{U}(\mathcal{I}_{k',t})([0,\tau]) = \frac{1}{2^{k'-1}} \sum_{i=1}^{2^{k'-1}} \mathbb{I}\left[x_{t+i} \in [0,\tau]\right]$.

Fix $\tau \in [0,1]$ and consider an auxiliary sequence $x^{(\tau)}$ defined as follows:

$$x_i^{(\tau)} = \mathbb{I}\left[x_i \in [0,\tau]\right].$$

Then, $\mathcal{U}(\mathcal{I}_{k',t})([0,\tau])$ and $\mathcal{U}(\mathcal{J}_{k',t})([0,\tau])$ are exactly the means of subsequences $x_{t+1}^{(\tau)}, \ldots, x_{t+2^{k'-1}}^{(\tau)}$ and $x_{t+2^{k'-1}+1}^{(\tau)}, \ldots, x_{t+2^{k'}}^{(\tau)}$, respectively. Since $x^{(\tau)}$ is bounded in $[0,1]$, by Lemma 2.1,

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathcal{U}(\mathcal{I}_{k',t})([0,\tau]) - \mathcal{U}(\mathcal{J}_{k',t})([0,\tau])\right|\right] \\
\le &\sqrt{\mathbb{E}\left[\left(\mathcal{U}(\mathcal{I}_{k',t})([0,\tau]) - \mathcal{U}(\mathcal{J}_{k',t})([0,\tau])\right)^2\right]} && \text{(concavity of } \sqrt{x}\text{)} \\
\le &\frac{1}{\sqrt{k}}. && \text{(Lemma 2.1)}
\end{aligned}$$

Taking an integral over $\tau \in [0,1]$ proves that

$$\mathbb{E}\left[\mathrm{EMD}(\mathcal{I}_{k',t}, \mathcal{J}_{k',t})\right] = \int_0^1 \mathbb{E}\left[\left|\mathcal{U}(\mathcal{I}_{k',t})([0,\tau]) - \mathcal{U}(\mathcal{J}_{k',t})([0,\tau])\right|\right] \,\mathrm{d}\tau \le \frac{1}{\sqrt{k}},$$

which completes the proof, since the expected absolute loss is upper bounded by

$$\mathbb{E}\left[\left|f_{2^{k'-1}}(\mathcal{I}_{k',t}) - f_{2^{k'-1}}(\mathcal{J}_{k',t})\right|\right] \le \mathbb{E}\left[L \cdot \mathrm{EMD}(\mathcal{I}_{k',t}, \mathcal{J}_{k',t})\right] \le \frac{L}{\sqrt{k}}$$

due to the L-smoothness of (f_m). □

3.2 Concatenation-Concave Functions

Algorithm 1 also applies to the case where the function family to be predicted is concatenation-concave. The proof resembles that of Lemma 2.1, yet a slightly different induction hypothesis is used. Again, Lemma 3.2 readily extends to the general case where the sequence length is not a power of two and thus proves Theorem 1.6.

Lemma 3.2. *Suppose that the function family (f_m) is concatenation-concave and bounded in $[0,1]$. For any integer $k \geq 1$, Algorithm 1 achieves an expected squared loss of at most $\frac{4}{k}$ on any sequence of length 2^k.*

Proof. For integer $k \geq 1$ and $\mu \in [0,1]$, let $L(k,\mu)$ denote the maximum expected squared loss that Algorithm 1 incurs on a sequence of length 2^k with function value $f_{2^k}(x_1,\ldots,x_{2^k}) = \mu$. Let $\mu_1 = f_{2^{k-1}}(x_1,\ldots,x_{2^{k-1}})$ and $\mu_2 = f_{2^{k-1}}(x_{2^{k-1}+1},\ldots,x_{2^k})$. By the concatenation-concavity of (f_m), we have $\mu_1+\mu_2 \leq 2\mu$. In the following, we prove by induction that $L(k,\mu) \leq \frac{4\mu(2-\mu)}{k}$, which further implies that $L(k,\mu) \leq \frac{4}{k}$ for any $\mu \in [0,1]$.

When $k=1$, the squared loss is upper bounded by

$$L(1,\mu) = \max_{\substack{\mu_1,\mu_2\in[0,1]\\ \mu_1+\mu_2\leq 2\mu}} (\mu_1-\mu_2)^2 \leq \min(4\mu^2, 4(1-\mu)^2) \leq 4\mu(2-\mu).$$

Suppose that $k \geq 2$. With probability $\frac{1}{k}$, Algorithm 1 chooses $k' = k$ and the loss is given by $(\mu_1-\mu_2)^2$. With probability $\frac{k-1}{k}$, the algorithm chooses $k' \neq k$, and the algorithm is equivalent to running the same algorithm on either the first or last 2^{k-1} entries of the sequence. The conditional expected loss in this case is upper bounded, thanks to the induction hypothesis, by

$$\frac{L(k-1,\mu_1)+L(k-1,\mu_2)}{2} \leq \frac{2\mu_1(2-\mu_1)+2\mu_2(2-\mu_2)}{k-1}.$$

To sum up, we have

$$\begin{aligned} L(k,\mu) \leq & \sup_{\substack{\mu_1,\mu_2\in[0,1]\\ \mu_1+\mu_2\leq 2\mu}} \left[\frac{1}{k}(\mu_1-\mu_2)^2 + \frac{k-1}{k}\cdot\frac{2\mu_1(2-\mu_1)+2\mu_2(2-\mu_2)}{k-1}\right] \\ = & \frac{1}{k}\cdot \sup_{\substack{\mu_1,\mu_2\in[0,1]\\ \mu_1+\mu_2\leq 2\mu}} \left[4(\mu_1+\mu_2)-(\mu_1+\mu_2)^2\right] = \frac{4\mu(2-\mu)}{k} \end{aligned}$$

as desired, where the last step follows from $0 \leq \mu_1+\mu_2 \leq 2\mu \leq 2$ and the monotonicity of $4x-x^2$ on $[0,2]$. □

4 Fitting Unseen Data

In this section, we study the problem of finding a model that fits the upcoming data points with a small excess risk. We consider a finite model class $\mathcal{L}$, each element of which can be viewed as a loss function $\ell : \mathcal{X} \to [0,1]$. The goal of the player is to choose some time step t and window length m and output a model $\hat{\ell}$ that minimizes the excess risk defined as follows:

$$\frac{1}{m}\sum_{i=1}^{m} \hat{\ell}(x_{t+i}) - \inf_{\ell\in\mathcal{L}} \frac{1}{m}\sum_{i=1}^{m} \ell(x_{t+i}).$$

A natural approach to this problem is to follow the strategy in Algorithm 1 and output the "empirical risk minimizer" (ERM) of observed data. We formally state the algorithm as follows. The excess risk of Algorithm 2 can be bounded by a uniform convergence argument over all models in $\mathcal{L}$.

Algorithm 2: Empirical Risk Minimization

Input: Model class $\mathcal{L}$ and sequence $x \in \mathcal{X}^n$ of length $n = 2^k$.

$k' \leftarrow \mathcal{U}([k])$;

$t \leftarrow \mathcal{U}(\{0, 2^{k'}, 2 \cdot 2^{k'}, \ldots, n - 2^{k'}\})$;

Observe $x_1, x_2, \ldots, x_{t+2^{k'-1}}$;

$\hat{\ell} \leftarrow \operatorname{argmin}_{\ell \in \mathcal{L}} \frac{1}{m} \sum_{i=1}^{m} \ell(x_{t+i})$;

Predict that $\hat{\ell}$ minimizes the risk on $(x_{t+2^{k'-1}+1}, \ldots, x_{t+2^{k'}})$;

Proposition 4.1. *For any integer $k \geq 1$ and finite model class $\mathcal{L}$, Algorithm 2 achieves an expected excess risk of at most $O\left(\sqrt{\frac{|\mathcal{L}|}{k}}\right)$ on any sequence of length 2^k.*

Proof. Let $\mathcal{I}_{k',t}$ and $\mathcal{J}_{k',t}$ denote sequences $(x_{t+1}, \ldots, x_{t+2^{k'-1}})$ and $(x_{t+2^{k'-1}+1}, \ldots, x_{t+2^{k'}})$. For $\ell \in \mathcal{L}$, let $\ell(\mathcal{I}_{k',t})$ denote the average loss of ℓ on sequence $\mathcal{I}_{k',t}$. By a standard uniform convergence argument, the expected excess risk of Algorithm 2 is upper bounded by

$$
\begin{aligned}
& \mathbb{E}\left[2 \max_{\ell \in \mathcal{L}} \left|\ell(\mathcal{I}_{k',t}) - \ell(\mathcal{J}_{k',t})\right|\right] \\
= \; & 2\mathbb{E}\left[\sqrt{\max_{\ell \in \mathcal{L}} \left[\ell(\mathcal{I}_{k',t}) - \ell(\mathcal{J}_{k',t})\right]^2}\right] \\
\leq \; & 2\sqrt{\mathbb{E}\left[\max_{\ell \in \mathcal{L}} \left[\ell(\mathcal{I}_{k',t}) - \ell(\mathcal{J}_{k',t})\right]^2\right]} && \text{(concavity of } \sqrt{x}\text{)} \\
\leq \; & 2\sqrt{\sum_{\ell \in \mathcal{L}} \mathbb{E}\left[\left(\ell(\mathcal{I}_{k',t}) - \ell(\mathcal{J}_{k',t})\right)^2\right]} = O\left(\sqrt{\frac{|\mathcal{L}|}{k}}\right). && \text{(Lemma 2.1)}
\end{aligned}
$$

□

Falling short of proving a lower bound that matches Proposition 4.1, we show that further improving the excess risk would require a more sophisticated prediction scheme than Algorithm 2. In particular, Proposition 4.2 states that when $|\mathcal{L}| = \Theta(\log n)$, Algorithm 2 incurs a constant excess risk in expectation and thus the upper bound in Proposition 4.1 is almost tight for Algorithm 2.

Proposition 4.2. *For any integer $k \geq 2$, there exists a model class $\mathcal{L}$ of size k and a sequence (x_t) of length 2^k such that Algorithm 2 incurs an expected excess risk of at least $\frac{1}{8}$ on (x_t).*

Proof. Let $\mathcal{X} = \{0, 1, \ldots, 2^k - 1\}$ and $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_k\}$. Each $\ell_i(x)$ is defined as:

$$
\ell_i(x) = \begin{cases} 1, & \left\lfloor \frac{x}{2^{i-1}} \right\rfloor \text{ is odd,} \\ i \cdot \epsilon, & \text{otherwise,} \end{cases}
$$

where $\epsilon = \frac{1}{4k}$. The input sequence is defined as $x_t = t - 1$. An example of the construction with $k = 3$ is shown in Table 1.

x_t	0	1	2	3	4	5	6	7
$\ell_1(x_t)$	ϵ	1	ϵ	1	ϵ	1	ϵ	1
$\ell_2(x_t)$	2ϵ	2ϵ	1	1	2ϵ	2ϵ	1	1
$\ell_3(x_t)$	3ϵ	3ϵ	3ϵ	3ϵ	1	1	1	1

Table 1: Construction for $k = 3$.

Let $\mathcal{I}_{k',t}$ and $\mathcal{J}_{k',t}$ denote subsequences $x_{t+1}, \ldots, x_{t+2^{k'-1}}$ and $x_{t+2^{k'-1}+1}, \ldots, x_{t+2^{k'}}$. For $\ell \in \mathcal{L}$, let $\ell(\mathcal{I}_{k',t})$ denote the average loss of ℓ on sequence $\mathcal{I}_{k',t}$. It can be verified that

$$\ell_i(\mathcal{I}_{k',t}) = \begin{cases} \frac{i\epsilon+1}{2}, & i < k', \\ k' \cdot \epsilon, & i = k', \\ i\epsilon \text{ or } 1, & i > k' \end{cases} \quad \text{and} \quad \ell_i(\mathcal{J}_{k',t}) = \begin{cases} \frac{i\epsilon+1}{2}, & i < k', \\ 1, & i = k', \\ i\epsilon \text{ or } 1, & i > k'. \end{cases}$$

By our choice of $\epsilon = \frac{1}{4k}$, $\ell_{k'}$ is always the unique minimizer of $\ell(\mathcal{I}_{k',t})$. Thus, Algorithm 2 always outputs $\ell_{k'}$. Moreover, when $k' \neq 1$ (which happens with probability $1 - \frac{1}{k} \geq \frac{1}{2}$), the resulting excess risk is at least $\ell_{k'}(\mathcal{J}_{k',t}) - \ell_1(\mathcal{J}_{k',t}) = 1 - \frac{\epsilon+1}{2} \geq \frac{1}{4}$. This proves the lower bound of $\frac{1}{8}$ on the expected excess risk incurred by Algorithm 2. □

References

[BK99] Avrim Blum and Adam Kalai. Universal portfolios with and without transaction costs. *Machine Learning*, 35(3):193–205, 1999.

[CBFH+97] Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CO96] Thomas M Cover and Erik Ordentlich. Universal portfolios with side information. *Transactions on Information Theory (TIT)*, 42(2):348–363, 1996.

[Cov91] Thomas M Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.

[Dic14] Lee H Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014.

[Dru13] Andrew Drucker. High-confidence predictions under adversarial uncertainty. *Transactions on Computation Theory (TOCT)*, 5(3):12, 2013.

[FKT17] Uriel Feige, Tomer Koren, and Moshe Tennenholtz. Chasing ghosts: competing with stateful policies. *SIAM Journal on Computing*, 46(1):190–223, 2017.

[FMG92] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *Transactions on Information Theory (TIT)*, 38(4):1258–1270, 1992.

[Han57] James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

[KV18] Weihao Kong and Gregory Valiant. Estimating learnability in the sublinear data regime. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5460–5469, 2018.

[LW94] Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[SKLV18] Vatsal Sharan, Sham Kakade, Percy Liang, and Gregory Valiant. Prediction with a short memory. In *Symposium on Theory of Computing (STOC)*, pages 1074–1087, 2018.

[VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Conformal prediction*. Springer, 2005.

A Proof of Proposition 2.3

Proof of Proposition 2.3. In the first case that t is known to the adversary, we simply construct a binary sequence such that $x_{t+1} = \cdots = x_n$, and x_{t+1} is randomly drawn from $\{0, 1\}$ with equal probability. When $\mathcal{A}$ makes a prediction at time t, the actual average of the sequence is either 0 or 1 with equal probability. It can be verified that any algorithm must achieve an expected squared loss of at least $\frac{1}{4} \geq \frac{1}{64}$.

Now we consider the second case, where the window length m is fixed. We choose $m' = \lceil \frac{m}{2} \rceil$ and construct a sequence consisting of blocks of length m'. Each block consists of the same value, which is chosen from $\{0, 1\}$ uniformly and independently at random. Whenever Algorithm $\mathcal{A}$ makes a prediction, by our choice of m', the prediction window of size m must contain an entire block. Since the variance in the average of the block is $\frac{1}{4}$ and the block contributes an $\frac{m'}{m}$ fraction to the average that $\mathcal{A}$ aims to predict, the variance in the arithmetic mean is then lower bounded by $\left(\frac{m'}{m}\right)^2 \cdot \frac{1}{4} \geq \frac{1}{64}$. This implies a lower bound of $\frac{1}{64}$ on the squared loss.

In the third case, the prediction algorithm chooses t, but is forced to make a prediction over the entire remaining $n - t$ timesteps. In this case, consider constructing an adversarial distribution over sequences of length n such that the first block of $n/2$ values are all identical and are chosen to either all be 0 or all be 1 with probability 1/2 of each choice, then next block of $n/4$ are identical and randomly selected to be either 0 or 1, and similarly for the next block of $n/8$, $n/16$, $n/32$, etc. Let t denote the time at which the prediction algorithm makes its prediction. There will always some i for which the block of size $n/2^i$ is contained within the final $n - t$ timesteps, and for which $n/2^i$ is at least a 1/4 fraction of $n - t$. Hence the variance in the average value due to that block alone implies a lower bound of at least $\frac{1}{64}$ on the expected squared loss of any prediction. □