

Adaptive Weighted Averaging

Aditya Bhaskara *University of Utah*

Ashok Cutkosky *Boston University*

Ravi Kumar *Google*

Manish Purohit *Google*

BHASKARAADITYA@GMAIL.COM

ASHOK@CUTKOSKY.COM

RAVI.K53@GMAIL.COM

PUROHITMANISH89@GMAIL.COM

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the problem of selecting the largest among n unknown values x_1, \dots, x_n given only a single unbiased estimate y_i for each x_i . We design strategies that are simultaneously admissible (not uniformly dominated by any other strategy) and also never worse than a given baseline such as uniform random selection. We provide an application to stochastic optimization, where we obtain online-to-batch conversion bounds with a desirable “no-compromise” guarantee: they are *never worse* than standard random iterate selection, and yet can be significantly better in benign settings.

1. Introduction

We consider a fundamental aggregation problem: given n unknown numbers $x = (x_1, \dots, x_n)$ and a single unbiased estimate y_i for each, how well can one identify the index of the maximum x_i ? This problem and its solutions arise throughout learning theory, with links to analysis of empirical risk minimization (Pollard, 2012; Vapnik, 2013), bandits (Bubeck et al., 2009; Audibert and Bubeck, 2010), experts, and stochastic optimization (Cesa-Bianchi et al., 2004; Cesa-Bianchi and Lugosi, 2006). We work with a standard generalization of this problem: rather than picking a specific index, we choose a categorical distribution over the indices $w \in \Delta_n$. Then, an intuitive measure of the quality of such a w is the inner product $\langle w, x \rangle$.

Although this problem is simple to state, there are surprising nuances to consider when designing a strategy. One intuitive approach is the simple *Empirical Risk Minimization* (ERM) rule that chooses the index i with the highest value of y_i . This is optimal in the sense of being “admissible” (which we will define presently), but it is widely recognized that one usually obtains better results by employing some form of regularization to move away from ERM. Alternatively, consider the naive approach of simply choosing an index uniformly at random. On the surface, this method might seem absurd: it completely ignores the estimates! Nevertheless, it is surprisingly effective in some applications. For example, this is the method employed by the so-called “online-to-batch conversion” that yields minimax optimal algorithms in many stochastic optimization problems.

1.1. No-compromise adaptivity

Given these two extremes, our focus is twofold: we attempt to build a method for choosing w that is (i) *never worse* than any arbitrary baseline “prior” over the choice of index (such as uniform sampling), and is (ii) also optimal in the sense that no other method can uniformly improve upon it.

By doing so, we are able to provide techniques that go beyond minimax optimality in certain stochastic optimization problems. Specifically, we obtain estimates that are never worse than the standard minimax optimal estimate, but provably obtain better loss in favorable settings. This goes

beyond the standard “adaptivity” story in stochastic optimization. Typical adaptive methods may do better in a benign setting, but sacrifice some performance in the worst-case setting. For example, adaptive gradient descent methods like AdaGrad (Duchi et al., 2011) perform better when the gradients are small, but their worst case performance is worse than the minimax rate. In contrast, we do better in benign settings without sacrificing anything in the hard settings.

1.2. Problem setup

Let $x = (x_1, \dots, x_n) \in [0, 1]^n$ be an unknown arbitrary vector and let $y = (y_1, \dots, y_n) \in [0, 1]^n$ be an unbiased estimate of x ; we assume that the y_i ’s are independent. We define a strategy S as an algorithm that uses the estimate y to output a distribution $S(y)$ over the indices $\{1, \dots, n\}$.

The *value* of a strategy S is $\mathbb{E}_{y \sim D}[\langle S(y), x \rangle]$. A strategy S *dominates* S' if its value is always at least that of S' no matter the value of x or the distribution of y . S *strictly dominates* S' if, in addition, there is some setting of x and distribution y for which the value of S is strictly larger than the value of S' . S is said to be *admissible* if it is not strictly dominated by any other strategy. We note that admissibility is a classic notion from Bayesian decision theory (e.g., see, (Robert, 2007)).

We also assume that we are given some *benchmark* strategy S_{det} , where S_{det} is a “constant” strategy that does not make use of the observation y at all. For example, one might consider the uniform benchmark $S_{\text{unif}}(y) = (1/n, \dots, 1/n)$. Our goal is to design a strategy that is simultaneously admissible and also dominates any given S_{det} .

1.3. Contributions

The S_{Bern} strategy. We first introduce a strategy S_{Bern} , derived via the multilinear extension of a base strategy optimized for Bernoulli observations (Section 5). (In that simpler case, the base strategy would select uniformly at random from among indices where $y_i = 1$.) We prove that S_{Bern} is admissible and strictly dominates the uniform strategy. Specifically, we show that the value of S_{Bern} improves over the mean $\mu(x)$ by a term related to the variance of the means $\sigma^2(x)$:

$$\text{value}(S_{\text{Bern}}; D) \geq \mu(x) + \left(1 - \prod_{k=1}^n (1 - x_k)\right) \cdot \frac{\sigma^2(x)}{\mu(x)}.$$

Arbitrary benchmarks. Moving beyond uniform strategies, we next consider arbitrary fixed deterministic strategies. For any such strategy S_{det} , we construct a new strategy (S_{peel}) that is both admissible and dominates S_{det} (Section 6).

Impossibility Results. We consider the harder setting where the observations y_i are not independent and in particular study the sequential setting where y_j (and the mean x_j) can depend on previous observations y_i for $i < j$. In this case, we construct a simple counterexample (Section 8) that shows uniform sampling is in fact admissible and cannot be dominated by any strategy. Even when the observations are independent, we show that for any two fixed benchmark strategies, there does not exist a strategy that simultaneously dominates both of them (Section 9).

Applications. We apply these results to stochastic optimization and establish a new online-to-batch conversion bound that uniformly improves upon the standard techniques by a term depending on amount of variation observed in the losses during the iteration of the online algorithm (Section 7). This provides a rigorous way to take advantage of the intuition that one should not put much weight on the early iterations of an optimization algorithm without sacrificing the worst-case optimality of uniform averaging. We also discuss other applications to ensemble methods and federated learning.

2. Related work

Optimal decision theory. Our setting can be precisely phrased in the language of Bayesian decision theory (see (Robert, 2007) for an excellent survey). In this language, our decision space is the n -dimensional simplex, and we seek to maximize the utility $\langle w, x \rangle$ (equivalently, minimize the loss $-\langle w, x \rangle$). Our goal is to design an estimator that is not only admissible, but also uniformly dominates any given benchmark strategy. This classical literature already suggests a natural approach for finding admissible strategies: start with any prior over decisions and choose a maximum a-posteriori strategy. This approach makes it fairly easy to see that the ERM rule is admissible. However, it is unclear how to design priors for which this approach actually dominates a given benchmark. In fact, it is surprisingly difficult to design a prior that does not simply yield the ERM strategy.

Prophet inequalities. Our setting has some conceptual similarity to the prophet-inequality literature, though the objectives differ. Prophet inequalities (Krengel and Sucheston, 1977; Hill and Kertz, 1982) study online selection from a sequence of independent random variables with the goal of approximating the expected maximum. Tight constant-factor bounds are known for i.i.d. and non-i.i.d. settings (Correa et al., 2017; Lachish, 2021) and some recent works combine prior samples with online decisions (Rubinstein et al., 2020; Immorlica et al., 2020). Our problem, on the other hand, does not involve online stopping or maximizing selection reward.

Statistical estimation. Our setting also has a loose relationship to standard statistical estimation problems, e.g., empirical Bayes and shrinkage methods, which use information sharing across coordinates to improve mean estimation or attempt to estimate rankings in high probability (Robbins, 1956; Efron, 2012; Bechhofer, 1954). Classical results such as Stein’s phenomenon (Stein, 1956; James and Stein, 1961) typically assume Gaussian noise and seek to estimate the *entire* vector of means under squared-error loss, whereas in our setting the distributions are arbitrary, only a single sample is observed from each, and the target is a scalar that depends on the mean vector.

Iterate averaging and online-to-batch conversions. Our setting is also closely connected to the concept of iterate averaging in stochastic optimization. Formally, if $F : \mathbb{R}^n \rightarrow \mathbb{R}$ is some objective function, one might run some algorithm like stochastic gradient descent for N iterations to obtain iterates z_1, \dots, z_N in \mathbb{R}^n , and then set \bar{z} to be a uniformly selected z_i . This immediately implies:

$$\mathbb{E}[F(\bar{z}) - F(z^*)] \stackrel{\text{definition of } \bar{z}}{=} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[F(z_i) - F(z^*)] = \frac{\mathbb{E}[\text{regret}]}{N}.$$

This is usually justified by observing that it yields minimax optimal algorithms in certain cases. However, the worst case examples for minimax optimality often somewhat degenerate. For example, the standard minimax adversary for non-smooth convex optimization forces $F(z_i) = F(z_j)$ for all i, j (Bubeck, 2015), so that there is no “information” available in the actual observed values of the $F(z_i)$. In practice, however, one expects some variation in the values. We would like an alternative weighted iterate averaging approach that can take advantage of such variation.

Prior literature frequently employs some limited forms of non-uniform averaging: non-adaptive polynomial weights are often used when the losses are known to have some additional smooth and convex structure (Joulani et al., 2020; Kavis et al., 2019; Lacoste-Julien et al., 2012; Cutkosky, 2019), while more advanced “adaptive” weighting schemes attempt to improve on uniform averaging more generally (Dekel, 2008; Levy, 2017; Defazio et al., 2023). However, these approaches

are usually restricted to the causal regime, where the weight assigned to y_t is a function solely of the historical sequence $\{y_i\}_{i=1}^{t-1}$ and they all under-perform uniform averaging in difficult settings in expectation. In contrast, our framework relaxes the causality constraint, allowing weights for y_t to incorporate information from future iterates. Crucially, we also provide guarantees that our method is never outperformed by uniform averaging.

3. Preliminaries

Let $[n] = \{1, \dots, n\}$. Let D_1, \dots, D_n be distributions supported on $[0, 1]$. We observe a single sample $y_i \sim D_i$, independently drawn, for each $i \in [n]$. Let $x_i = \mathbb{E}[y_i]$; we assume that x_i is not known to the algorithm. Let $y = (y_1, \dots, y_n)$, $x = (x_1, \dots, x_n) = (\mathbb{E}[y_1], \dots, \mathbb{E}[y_n]) =: \mathbb{E}[y]$, and $D = D_1 \times \dots \times D_n$. Let $\text{Bern}(x)$ denote the Bernoulli distribution with mean x .

Let Δ_n denote the probability simplex in \mathbb{R}^n and for a vector, its i th coordinate is accessed using the subscript i . Let $\mu(x) = \frac{1}{n} \sum_{i=1}^n x_i$ and let $\sigma^2(x) = (\frac{1}{n} \sum_{i=1}^n x_i^2) - \mu^2(x)$.

Definition 1 (Strategy and value) A strategy is a function $S : [0, 1]^n \rightarrow \Delta_n$. The value of a strategy S for a product distribution D is the expected value of the mean outcome, i.e., $\text{value}(S; D) = \mathbb{E}_{y \sim D}[\langle S(y), \mathbb{E}[y] \rangle] = \mathbb{E}_{y \sim D}[\sum_{i=1}^n x_i \cdot S(y)_i]$.

Definition 2 (Benchmark strategy) A strategy $S : [0, 1]^n \rightarrow \Delta_n$ is called a benchmark if it maps every outcome to a fixed probability distribution, i.e., $S(y) = p, \forall y \in [0, 1]^n$ for some fixed $p \in \Delta_n$. For ease, we abuse notation and use $S = p$ to denote a benchmark strategy.

We define S_{unif} to be the benchmark strategy that places uniform mass on $[n]$ regardless of its input, i.e., $S_{\text{unif}}(y) = (1/n, \dots, 1/n)$. Clearly, $\text{value}(S_{\text{unif}}; D) = \mu(x)$.

Definition 3 (Dominance and admissibility) A strategy S dominates strategy S' , denoted $S \succeq S'$, iff $\text{value}(S; D) \geq \text{value}(S'; D)$, for all product distributions D . S strictly dominates S' , denoted $S \succ S'$, iff both $S \succeq S'$ and there is a product distribution D such that $\text{value}(S; D) > \text{value}(S'; D)$. A strategy S is admissible iff it is not strictly dominated by any other strategy S' , i.e., there is no S' such that $S' \succ S$.

3.1. Failure of simple approaches

Before describing our algorithms, we take a moment to appreciate how one intuitively reasonable approach might fall short—even if we are satisfied with simply dominating a benchmark without considering admissibility. The reader familiar with standard optimization and learning theory might imagine that taking a single step of exponential weights with an appropriately small learning rate η , initialized at a benchmark, would suffice. This strategy is essentially taking one step of SGD on the value function. However, this strategy fails to dominate the benchmark, even for very small learning rates η . Indeed, we verify this numerically in Figure 1, where we

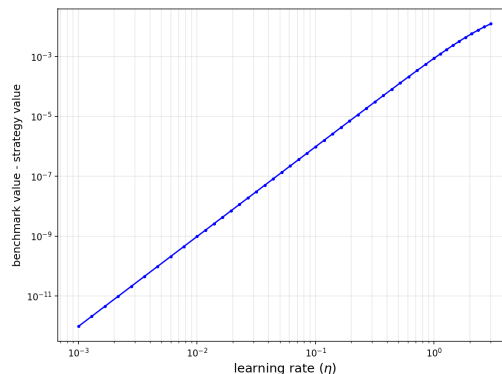


Figure 1: Naive single-step exponential weights based strategies fail to dominate the benchmark.

consider the case $D = (\text{Bern}(x_1), \text{Bern}(x_2))$. For any learning rate η , we perform a grid search over all tuples (x_1, x_2, p) where $p \in [0, 1]$ specifies a benchmark strategy $S_{\text{det}} = (p, 1 - p)$. We plot the maximum over all such tuples of the benchmark value $x_1 p + (1 - x_2) \cdot (1 - p)$ minus the expected value obtained by taking one step of exponential weights with learning rate η initialized at $(p, 1 - p)$, i.e., $S_{\text{exp}} = \left(\frac{p \exp(\eta y_1)}{p \exp(\eta y_1) + (1-p) \exp(\eta y_2)}, \frac{(1-p) \exp(\eta y_2)}{p \exp(\eta y_1) + (1-p) \exp(\eta y_2)} \right)$. Figure 1 shows that, no matter the η value, there is always an (x_1, x_2, p) tuple for which S_{exp} has smaller value than S_{det} .

4. Warm up: Two items

We first consider a simple setup with just two items (i.e., $n = 2$) and further let $D = D_1 \times D_2$ where $D_1 = \text{Bern}(x_1)$ and $D_2 = \text{Bern}(x_2)$. Since D is completely characterized by $x = (x_1, x_2)$, we define $\text{value}(S; x) := \text{value}(S; D)$ for convenience. In this setting, $\text{value}(S_{\text{unif}}; x) = \frac{x_1 + x_2}{2}$. Our goal is to design a strategy that is admissible and strictly dominates S_{unif} .

Since each $y_i \in \{0, 1\}$, we explicitly construct such a strategy $S_{n=2}$ as follows: $S_{n=2}((0, 0)) = S_{n=2}((1, 1)) = (\frac{1}{2}, \frac{1}{2})$, $S_{n=2}((0, 1)) = (0, 1)$ and $S_{n=2}((1, 0)) = (1, 0)$. In other words, the strategy deterministically chooses the item with the higher outcome, but reverts to the uniform distribution when the outcomes are equal. We omit the proof that $S_{n=2}$ is admissible here since we prove a more general result in Section 5.2 but we observe that $S_{n=2} \succeq S_{\text{unif}}$ as follows.

$$\begin{aligned} \text{value}(S_{n=2}; x) &= \mathbb{P}[y_1 = y_2] \cdot \left(\frac{x_1 + x_2}{2} \right) + \mathbb{P}[y = (0, 1)] \cdot (x_2) + \mathbb{P}[y = (1, 0)] \cdot (x_1) \\ &= (x_1 x_2 + (1 - x_1)(1 - x_2)) \left(\frac{x_1 + x_2}{2} \right) + (1 - x_1)x_2(x_2) + x_1(1 - x_2)(x_1) \\ &= \frac{x_1 + x_2}{2} + \frac{(x_1 - x_2)^2}{2} = \text{value}(S_{\text{unif}}; x) + \frac{(x_1 - x_2)^2}{2}. \end{aligned}$$

4.1. Extension to $n > 2$

When D is a product of n Bernoulli distributions, where $n > 2$, one can extend the $S_{n=2}$ strategy to dominate S_{unif} . Let S_{pair} be a strategy that first chooses two indices $i < j$ uniformly at random and then uses the $S_{n=2}$ strategy to pick among items i and j . Formally, for any y , we have $S_{\text{pair}}(y) = \frac{1}{\binom{n}{2}} \cdot \sum_{i < j} S_{n=2}((y_i, y_j))$. Once again, we observe that $S_{\text{pair}} \succeq S_{\text{unif}}$ as follows.

$$\begin{aligned} \text{value}(S_{\text{pair}}; x) &= \frac{1}{\binom{n}{2}} \cdot \sum_{i < j} \left(\frac{x_i + x_j}{2} + \frac{(x_i - x_j)^2}{2} \right) \\ &= \frac{\sum_{i=1}^n x_i}{n} + \frac{1}{\binom{n}{2}} \cdot \sum_{i < j} \frac{(x_i - x_j)^2}{2} = \text{value}(S_{\text{unif}}; x) + \frac{n}{n-1} \cdot \sigma^2(x). \end{aligned}$$

Thus, while S_{pair} shows that improvement over S_{unif} is possible for product of Bernoulli distributions, it is unclear if something better exists. We therefore turn our attention to designing a strategy that is never worse than S_{unif} , is admissible, and works for all product distributions. This leads us to a new general strategy, described in Section 5.

5. An admissible strategy: S_{Bern}

We now consider the general setting when the distributions D_i are arbitrary and hence we have $y_i \in [0, 1]$. In this section, we design a new strategy S_{Bern} that dominates S_{unif} and is also admissible.

It is easiest to describe the strategy S_{Bern} as a randomized process (although all strategies are deterministic by nature). Let $y = (y_1, \dots, y_n)$ be a fixed outcome. First, for each item i , we sample a new independent Bernoulli random variable z_i with mean y_i . Next, we choose an item uniformly at random from the items that have $z_i = 1$; if all $z_i = 0$, then we choose a uniformly random item. The strategy S_{Bern} is defined as the expectation of this process.

Let $f : \{0, 1\}^n \rightarrow \Delta_n$ be a function defined over vertices of the Boolean hypercube as $f(z) = (\frac{z_1}{\sum_j z_j}, \dots, \frac{z_n}{\sum_j z_j})$ for $z \neq 0$ and $f(0, \dots, 0) = (\frac{1}{n}, \dots, \frac{1}{n})$. Then formally, the strategy S_{Bern} is defined as the unique multilinear extension of f over the domain $[0, 1]^n$. Mathematically, we can express $S_{\text{Bern}}(y) = (q_1(y), \dots, q_n(y))$ where we have:

$$q_i(y) = \frac{1}{n} \prod_{j=1}^n (1 - y_j) + \sum_{S:i \in S} \frac{1}{|S|} \cdot \prod_{j \in S} y_j \cdot \prod_{k \notin S} (1 - y_k), \quad (1)$$

and, equivalently, using inclusion-exclusion¹, this can be expressed as:

$$= \frac{1}{n} \prod_{j=1}^n (1 - y_j) + \sum_{S:i \in S} \frac{(-1)^{|S|-1}}{|S|} \prod_{j \in S} y_j. \quad (2)$$

It is easy to see that the strategy is valid, i.e., $S_{\text{Bern}}(y) \in \Delta_n$ for all $y \in [0, 1]^n$.

5.1. Value

First, notice that since $q_i(y)$ is linear in each y_j and all the y 's are independent, we have $\mathbb{E}[q_i(y)] = q_i(x)$. Hence, we have $\text{value}(S_{\text{Bern}}; D) = \sum_{i=1}^n x_i \cdot \mathbb{E}[q_i(y)] = \sum_{i=1}^n x_i \cdot q_i(x)$. Next, we focus on the properties of $q_i(x)$.

Lemma 1 *For any two indices i and j , $q_i(x) - q_j(x) = (x_i - x_j) \cdot K_{i,j}(x)$ where*

$$K_{i,j}(x) = \int_{t=0}^1 \left(\prod_{k \notin \{i,j\}} (1 - tx_k) \right) dt.$$

Further, $K_{i,j}(x) \geq 0$, so $x_i > x_j \iff q_i(x) > q_j(x)$.

This immediately leads to a simple but loose lower bound on $K_{i,j}(\cdot)$.

Corollary 2 *For any pair $i, j \neq i$ of distinct indices and any vector x , we have $K_{i,j}(x) \geq \frac{1}{n-1}$.*

Proof For any x_k , we have $1 - tx_k \geq 1 - t$. So we have $K_{i,j}(x) \geq \int_{t=0}^1 (1 - t)^{n-2} dt = \frac{1}{n-1}$. ■

We now obtain a stronger lower bound. Towards this, we define the multivariate auxiliary function:

$$F(x) := \frac{1 - \prod_{k=1}^n (1 - x_k)}{\sum_{k=1}^n x_k}.$$

We first argue that $F(x)$ is non-increasing in each of its coordinates.

1. Lemma 13 in the Appendix contains a full proof of the equivalence.

Lemma 3 (Monotonicity) $F(x) = F(x_1, \dots, x_n)$ is a monotonically decreasing function of each of its variables $x_k \in [0, 1]$.

We use this to show an improved lower bound on $K_{i,j}(\cdot)$.

Lemma 4 For any pair $i, j \neq i$ of distinct indices and any vector x , we have $K_{i,j}(x) \geq F(x)$.

Proof To simplify notation, let $\bar{N}_{i,j} = [n] \setminus \{i, j\}$ and let $\bar{P}_{i,j} = \prod_{k \in \bar{N}_{i,j}} (1 - x_k)$. We will first show that

$$K_{i,j}(x) \geq \frac{1 - \prod_{k \in \bar{N}_{i,j}} (1 - x_k)}{\sum_{k \in \bar{N}_{i,j}} x_k} = \frac{1 - \bar{P}_{i,j}}{\sum_{k \in \bar{N}_{i,j}} x_k}.$$

Observe that for $t \in [0, 1]$, the function $(1 - y)^t$ is concave and hence it lies below its tangent, i.e., $(1 - y)^t \leq 1 - ty$. Applying this, we obtain

$$K_{i,j}(x) = \int_{t=0}^1 \left(\prod_{k \in \bar{N}_{i,j}} (1 - tx_k) \right) dt \geq \int_{t=0}^1 \left(\prod_{k \in \bar{N}_{i,j}} (1 - x_k) \right)^t dt = \int_{t=0}^1 \bar{P}_{i,j}^t dt = \frac{\bar{P}_{i,j} - 1}{\log \bar{P}_{i,j}}. \quad (3)$$

Next we claim $\log \bar{P}_{i,j} \leq -\sum_{k \in \bar{N}_{i,j}} x_k$. Indeed, by concavity as before, we have for all $t \in [0, 1]$, $\log(1 - y) \leq \frac{\log(1 - ty)}{t}$ and hence $\log(1 - y) \leq \lim_{t \rightarrow 0} \frac{\log(1 - ty)}{t} = -y$. The claim follows by applying this inequality pointwise to each term in $\log \bar{P}_{i,j}$.

Using this claim in (3), we obtain

$$\begin{aligned} K_{i,j}(x) &\geq \frac{\bar{P}_{i,j} - 1}{-\sum_{k \in \bar{N}_{i,j}} x_k} = \frac{1 - \prod_{k \in \bar{N}_{i,j}} (1 - x_k)}{\sum_{k \in \bar{N}_{i,j}} x_k} = F(x_1, \dots, x_i = 0, \dots, x_j = 0, \dots, x_n) \\ &\geq F(x_1, \dots, x_n), \text{ using the monotonicity of } F \text{ (Lemma 3)}. \quad \blacksquare \end{aligned}$$

Finally, we apply the lower bounds on $K_{i,j}(x)$ to bound the value of S_{Bern} .

Theorem 5 We have the following two lower bounds:

- $\text{value}(S_{\text{Bern}}; D) \geq \mu(x) + \frac{n}{n-1} \cdot \sigma^2(x)$,
- $\text{value}(S_{\text{Bern}}; D) \geq \mu(x) + (1 - \prod_{k=1}^n (1 - x_k)) \cdot \frac{\sigma^2(x)}{\mu(x)}$.

Proof As before, let $S_{\text{Bern}}(y) = (q_1(y), \dots, q_n(y))$. Recall that

$$\text{value}(S_{\text{Bern}}; D) = \mathbb{E}_y \left[\sum_i x_i \cdot q_i(y) \right] = \sum_i x_i \cdot \mathbb{E}[q_i(y)] = \sum_i x_i \cdot q_i(x).$$

We begin with the following identity

$$\sum_i a_i b_i = \frac{1}{n} \sum_i a_i \sum_i b_i + \frac{1}{n} \sum_{i < j} (a_i - a_j)(b_i - b_j).$$

Substituting $a_i = q_i(x)$ and $b_i = x_i$, using $\sum_i q_i(x) = 1$, and using Lemma 1,

$$\sum_i x_i \cdot q_i(x) = \frac{1}{n} \sum_i x_i + \frac{1}{n} \sum_{i < j} (x_i - x_j)(q_i(x) - q_j(x)) = \mu(x) + \frac{1}{n} \sum_{i < j} (x_i - x_j)^2 K_{i,j}(x).$$

Now, lower bounding $K_{i,j}(x)$ using Corollary 2, we obtain:

$$\sum_i x_i \cdot q_i(x) \geq \mu(x) + \frac{1}{n(n-1)} \sum_{i < j} (x_i - x_j)^2 = \mu(x) + \frac{n}{n-1} \cdot \sigma^2(x).$$

Instead, lower bounding $K_{i,j}(x)$ using Lemma 4, we obtain:

$$\sum_i x_i \cdot q_i(x) \geq \mu(x) + \frac{1 - \prod_{k=1}^n (1 - x_k)}{n \sum_{k=1}^n x_k} \sum_{i < j} (x_i - x_j)^2 = \mu(x) + \left(1 - \prod_{k=1}^n (1 - x_k)\right) \cdot \frac{\sigma^2(x)}{\mu(x)}. \blacksquare$$

Theorem 5 shows that S_{Bern} *dominates* the S_{pair} strategy we discussed in Section 4. Note that the two bounds in Theorem 5 are incomparable as they depend on the sparsity of x . This trade-off can be seen from the coefficient multiplying $\sigma^2(x)$ in each bound. The first bound has a coefficient of $\frac{n}{n-1} \approx 1$ while the second bound has a coefficient of $(1 - \prod_{k=1}^n (1 - x_k))/\mu(x) \geq 1$ (Lemma 14).

Indeed, consider the setting when ℓ of the x_k 's is ≈ 1 and the others are ≈ 0 ; hence, the product $\prod_k (1 - x_k) \approx 0$ and $\mu(x) \approx \ell/n$. When $\ell = 1$, the second coefficient $\approx n$, making the second bound better than the first. When $\ell = n$, the second coefficient ≈ 1 , making it slightly weaker than the first bound's coefficient of $\frac{n}{n-1}$. Thus, while the second bound can be asymptotically better (by a factor of n), it is never worse than the first bound by more than a small factor of $\approx n/(n-1)$.

5.2. Admissibility

In this section, we show that S_{Bern} is admissible. To this end, we define the notion of symmetric strategies. Let \mathbb{S}_n be the set of all permutations on $[n]$. A strategy S is *symmetric* if $S(\pi(y)) = \pi(S(y))$ for any permutation $\pi \in \mathbb{S}_n$; here for a vector y , $\pi(y)$ is the vector y with elements permuted according to π . By construction, S_{Bern} is a symmetric strategy. Note that for a symmetric strategy S and for any $\pi \in \mathbb{S}_n$, we then have that $\text{value}(S; D) = \text{value}(S; D \circ \pi)$, where $D \circ \pi$ denotes the product distribution given by coordinates of D , permuted by π .

Lemma 6 *If $S \succeq S_{\text{Bern}}$, then there is a symmetric strategy $S' \succeq S_{\text{Bern}}$.*

Proof Let S' be a symmetrized version of S , i.e., $S'(y) = (1/n!) \sum_{\pi \in \mathbb{S}_n} \pi^{-1}(S(\pi(y)))$. For any product distribution D , we have

$$\text{value}(S'; D) = \frac{1}{n!} \sum_{\pi \in \mathbb{S}_n} \text{value}(S; D \circ \pi) \geq \frac{1}{n!} \sum_{\pi \in \mathbb{S}_n} \text{value}(S_{\text{Bern}}; D \circ \pi) = \text{value}(S_{\text{Bern}}; D),$$

where the inequality follows since $S \succeq S_{\text{Bern}}$ and the last step used the symmetry of S_{Bern} . \blacksquare

Lemma 7 *If $S \succeq S_{\text{Bern}}$, then $S(y) = S_{\text{Bern}}(y)$ for all binary outcomes $y \in \{0, 1\}^n$.*

Proof By Lemma 6, we can assume that S is symmetric. For $y = 0$, any symmetric strategy must have $S(y) = S_{\text{Bern}}(y) = (1/n, \dots, 1/n)$. Now for any other binary outcome $y \in \{0, 1\}^n$, consider the distribution D_y that is point mass at y in the binary hypercube; hence, $\text{value}(S_{\text{Bern}}; D_y) = 1$. Given $S \succeq S_{\text{Bern}}$, it can only depend on indices i such that $y_i = 1$. But since S is symmetric, it must be uniform over the ones in y , i.e., identical to the ones strategy for binary outcomes, which by construction coincides with S_{Bern} for binary outcomes. \blacksquare

The proof proceeds by contradiction, assuming there is a strategy S that strictly dominates S_{Bern} . For an outcome y , let $\#_0(y) = |\{i : y_i = 0\}|$ (resp., $\#_1(y) = |\{i : y_i = 1\}|$) denote the number of zeroes (resp., ones) in y . Consider the ordering \succ of the y 's induced by the natural lexicographic ordering of $(\#_0(y), \#_1(y))$. The main idea is a backward induction on this ordering of the y 's; the base of the induction is Lemma 7. For the inductive step, we select an outcome y^* that is highest in the ordering among those where S and S_{Bern} differ. We then construct a specific product distribution D that exploits the difference at outcome y^* . If S places less weight on an item j compared to S_{Bern} where $y_j^* = 1$, we assign that item a high mean in D , making S_{Bern} better than S on D . Conversely, if S places less weight on an item j where $y_j^* = 0$, we treat j as a ‘‘rare but high-value’’ item, assigning it a small probability p of being 1 but a high value, in D . By carefully choosing parameters, we can once again ensure that the gain from S_{Bern} correctly weighting item j in the rare event $y_j = 1$ is more than the loss that can arise in the common event $y_j = 0$. Both contradict that S dominates S_{Bern} .

Theorem 8 *If $S \succeq S_{\text{Bern}}$, then $S = S_{\text{Bern}}$.*

Proof Suppose for the sake of contradiction that $S \succeq S_{\text{Bern}}$ but $S \neq S_{\text{Bern}}$. Let y^* be an outcome such that $S(y^*) \neq S_{\text{Bern}}(y^*)$ and y^* is the highest in the ordering \succ . Inductively, for any outcome $y \succ y^*$, we know that $S(y) = S_{\text{Bern}}(y)$.

For ease, let $\delta(y) := S_{\text{Bern}}(y) - S(y)$. Since $S(y), S_{\text{Bern}}(y) \in \Delta_n$ and since $S(y^*) \neq S_{\text{Bern}}(y^*)$, let $j \in [n]$ be an index such that $\delta(y^*)_j > 0$. Let $Z = \{i \mid y_i^* = 0\}$ be the set of zero indices in y^* . We consider two cases depending on whether y_j^* is zero or non-zero.

Case 1: $[j \notin Z]$. We construct a product distribution D , parametrized by a small ϵ , over the outcomes as follows:

- D_j is a point mass at y_j^* ; hence $x_j = y_j^* > 0$.
- For each $i \in Z$, D_i is a point mass at 0; hence $x_i = 0, \forall i \in Z$
- For every other index k , $D_k = y_k^* \cdot \text{Bern}(\epsilon)$; hence $x_k = \epsilon \cdot y_k^*, \forall k \in [n] \setminus Z \setminus \{j\}$.

Note that under this distribution D , the only outcomes are either y^* or y with $\#_0(y) > \#_0(y^*)$. For the latter outcome, we have $y \succ y^*$ and by the inductive hypothesis, $S(y) = S_{\text{Bern}}(y)$. Hence, $\text{value}(S_{\text{Bern}}; D) > \text{value}(S; D)$ if and only if $x_j \delta(y^*)_j + \sum_{k \neq j} x_k \delta(y^*)_k \geq 0$ which can be ensured by choosing ϵ small enough.

Case 2: $[j \in Z]$. We construct a product distribution D , parametrized by small $p \in (0, 1)$, over the outcomes as follows:

- $D_j = \text{Bern}(p)$; hence $x_j = p$.
- For each $i \in Z \setminus \{j\}$, D_i is a point mass at 0; hence $x_i = 0, \forall i \in Z \setminus \{j\}$.
- For every other index k , $D_k = y_k^* \cdot \text{Bern}(q)$ for $q = p^2$; hence $x_k = q \cdot y_k^*, \forall k \in [n] \setminus Z$.

We first argue that $y_k^* < 1, \forall k \in [n] \setminus Z$, so that $\#_1(y^*) = 0$. Indeed, otherwise we'll have $y_k^* = 1$ and $y_j^* = 0$, which implies that $S_{\text{Bern}}(y^*)_j = 0$ and contradicts the choice of j . Now, consider the outcomes y from the distribution D . First, suppose $y_k = 0$ for any $k \notin Z$. Then since $y_j = 0$ for all $k \in Z \setminus \{j\}$, we must have $\#_0(y) \geq \#_0(y^*)$. Moreover, the only way to have $\#_0(y) = \#_0(y^*)$ is for $y_j = 1$, in which case $\#_1(y) > 0 = \#_1(y^*)$. In either case, we have $y \succ y^*$ and hence by the inductive hypothesis, $S(y) = S_{\text{Bern}}(y)$. Hence, the only class of outcomes y from D to consider is when $y_k = y_k^*, \forall k \notin Z$.

There are only two outcomes in this class: either when $y_j = 0$ or $y_j = 1$. When $y_j = 0$, the outcome is exactly y^* ; call the outcome z^* when $y_j = 1$. Let $d = |[n] \setminus Z|$.

$$\begin{aligned} & \text{value}(S_{\text{Bern}}; D) - \text{value}(S; D) \\ &= \Pr[y = y^*] \cdot \left(x_j \delta_j(y^*) + \sum_{k \notin Z} x_k \delta_k(y^*) \right) + \Pr[y = z^*] \cdot \left(x_j \delta_j(z^*) + \sum_{k \notin Z} x_k \delta_k(z^*) \right) \\ &= (1-p)q^d \cdot \left(p\delta_j(y^*) + \sum_{k \notin Z} qy_k^* \delta_k(y^*) \right) + pq^d \cdot \left(p\delta_j(z^*) + \sum_{k \notin Z} qy_k^* \delta_k(z^*) \right) \end{aligned}$$

We want to show that the above expression is positive for some appropriate value of p and q . We can loosely bound the term in the second parenthesis above by $-p$ as in the worst case S_{Bern} gets zero value and S gets value p .

$$\geq (1-p)q^d \cdot \left(p\delta_j(y^*) + \sum_{k \notin Z} qy_k^* \delta_k(y^*) \right) - p^2q^d$$

Similarly $\sum_{k \notin Z} qy_k^* \delta_k(y^*) \geq -q$ since again in the worst case S_{Bern} gets zero value but S gets the best item in Z which has value at most q .

$$\begin{aligned} & \geq (1-p)q^d \cdot (p\delta_j(y^*) - q) - p^2q^d = (1-p)pq^d\delta_j(y^*) - (1-p)q^{d+1} - p^2q^d \\ & = (1-p)p^{2d+1}\delta_j(y^*) - (1-p)p^{2d+2} - p^{2d+2} > 0, \end{aligned}$$

since $\delta_j(y^*) > 0$ and for p small enough, contradicting the dominance of S . ■

6. Arbitrary benchmarks

We now consider the general problem of designing strategies that dominate an arbitrary benchmark strategy rather than just the uniform strategy S_{unif} . Let $S_{\text{det}} = (p_1, \dots, p_n)$ be an arbitrary benchmark strategy. We design a new strategy S_{peel} that dominates S_{det} and is also admissible.

Analog of the uniform-on-ones strategy. Given a benchmark S_{det} , a natural way to extend the uniform-on-ones strategy is the following: observe the values y , and let J be set of indices with $y_j = 1$. Then we output the vector $p_J / \|p_J\|_1$, where p_J is the vector (p_1, \dots, p_n) restricted to the indices J (by setting the entries in $[n] \setminus J$ to 0). As before, if $J = \emptyset$, we output (p_1, \dots, p_n) .

Interestingly, this approach fails for non-uniform choices of S_{det} . Consider the following example with $n = 2$: $S_{\text{det}} = (\frac{1}{3}, \frac{2}{3})$. Suppose $(x_1, x_2) = (\frac{1}{2}, \frac{1}{2} + \delta)$, for some $\delta > 0$. In this case, the strategy above outputs:

$$\begin{cases} (\frac{1}{3}, \frac{2}{3}) & \text{with prob. } \frac{1}{2} \\ (1, 0) & \text{with prob. } \frac{1}{2}(\frac{1}{2} - \delta) \\ (0, 1) & \text{with prob. } \frac{1}{2}(\frac{1}{2} + \delta). \end{cases}$$

For a small value of δ (e.g., in $(0, 0.1)$), we can observe that in expectation, this strategy does not dominate S_{det} . Essentially, the last two cases create a move towards $(\frac{1}{2}, \frac{1}{2})$, which is worse for the given (x_1, x_2) than the baseline of $(\frac{1}{3}, \frac{2}{3})$.

6.1. Peeling decomposition

Without loss of generality, let $p_1 \leq \dots \leq p_n$. Let $w_k = (n - k + 1)(p_k - p_{k-1}), \forall 1 \leq k \leq n$ (where p_0 is defined to be zero). We note that $w \in \Delta_n$ and observe that the benchmark strategy S_{det} can be decomposed into a weighted combination of uniform strategies on nested subsets of items. Formally, let S_{unif}^k denote the strategy that places equal probability mass on items k or higher, i.e., for any outcome y we have, $S_{\text{unif}}^k(y)_j = \frac{1}{n-k+1}$ if $j \in \{k, \dots, n\}$ and zero otherwise. Then,

$$S_{\text{det}} = \sum_{k=1}^n w_k \cdot S_{\text{unif}}^k.$$

This decomposition suggests a natural way to construct a dominating strategy since each of the uniform strategies in the decomposition can be replaced with a strategy that dominates it. Let S_{Bern}^k denote the strategy defined in Section 5 that is restricted to items k or higher. Formally we have $S_{\text{Bern}}^k(y) = (q_1(y), \dots, q_n(y))$ where:

$$q_i(y) = \begin{cases} 0, & \text{if } i < k \\ \frac{1}{n-k+1} \prod_{j=k}^n (1 - y_j) + \sum_{\substack{S \subset \{k, \dots, n\} \\ i \in S}} \frac{(-1)^{|S|-1}}{|S|} \prod_{j \in S} y_j, & \text{if } i \geq k. \end{cases} \quad (4)$$

Then, we define the strategy S_{peel} as:

$$S_{\text{peel}}(y) = \sum_{k=1}^n w_k \cdot S_{\text{Bern}}^k(y).$$

6.2. Value and admissibility

The following theorem specifies the gain of S_{peel} over the benchmark strategy S_{det} .

Theorem 9 *We have the following two lower bounds:*

- $\text{value}(S_{\text{peel}}; D) \geq \text{value}(S_{\text{det}}; D) + \sum_{k=1}^n w_k \cdot \frac{n-k+1}{n-k} \cdot \sigma^2(x; k),$
- $\text{value}(S_{\text{peel}}; D) \geq \text{value}(S_{\text{det}}; D) + \sum_{k=1}^n w_k \cdot \left(1 - \prod_{j=k}^n (1 - x_j)\right) \cdot \frac{\sigma^2(x; k)}{\mu(x; k)},$

where $\mu(x; k) = \frac{\sum_{j=k}^n x_j}{n-k+1}$ and $\sigma^2(x; k) = \left(\frac{1}{n-k+1} \cdot \sum_{j=k}^n x_j^2\right) - \mu(x; k)^2$ are defined to be the mean and variance of the components of x with indices k or higher.

We also show that S_{peel} is admissible. The proof again proceeds by induction and relies on carefully constructing product distributions that guarantee that any strategy that dominates S_{peel} must in fact be identical to S_{peel} .

Theorem 10 *If $S \succeq S_{\text{peel}}$, then $S = S_{\text{peel}}$. Hence S_{peel} is admissible.*

7. Application: Improved online-to-batch conversions

The classical online-to-batch conversion provides a simple technique for converting regret bounds in online optimization into stochastic optimization guarantees (Cesa-Bianchi et al., 2004). We briefly recapitulate the classical argument here: suppose that $f(x, z) \in \mathbb{R}$ is some loss function taking

as input a parameter x to be optimized and a data point z . Our goal is to minimize the function $F(x) = \mathbb{E}_Z[f(x, Z)]$ where Z follows some unknown distribution. Let z_1, \dots, z_N be an i.i.d. data set. Then we run an online learning algorithm to obtain iterates x_1, \dots, x_N such that x_i is independent of z_j for $j \geq i$. Then if \bar{x} is chosen uniformly at random from x_1, \dots, x_N , we have for any x_* (usually $x_* \in \operatorname{argmin} F$):

$$\mathbb{E}[F(\bar{x}) - F(x_*)] = \frac{\sum_{i=1}^N \mathbb{E}[f(x_i, z_i) - f(x_*, z_i)]}{N}.$$

The numerator of this quantity is exactly the *regret* that an online learning algorithm keeps under control (typically growing as \sqrt{N}). Thus, $\mathbb{E}[F(\bar{x})]$ approaches the optimal value.

However, with our techniques we can do better: suppose that $f(x, z) \in [0, 1]$. Now, for $i \leq N/2$, Define $X_i = 1 - F(x_i)$ and $Y_i = 1 - f(x_i, z_{i+N/2})$. For all $i \leq N/2$, we have x_i and $z_{j+N/2}$ are independent. Therefore, Y_i and Y_j are independent given $(X_1, \dots, X_{N/2})$ for all $i, j \leq N/2$. Now, let $w_1, \dots, w_{N/2}$ be the weighting obtained by our algorithm applied to $Y_1, \dots, Y_{N/2}$. Clearly,

$$\mathbb{E} \left[\sum_{i=1}^{N/2} w_i X_i \right] \geq \mathbb{E} \left[\frac{2}{N} \sum_{i=1}^{N/2} X_i \right].$$

Set $\hat{x} = x_i$ with probability $w_i/2$ if $i \leq N/2$ and $2/N$ if $i > N/2$. Then, we will have a value that is never worse than the standard “random iterate selection”, but may be better in the reasonable case that the X_i values are not all equal to each other. Thus, by Lemma 5, we have (with formal statement in Theorem 16):

$$\mathbb{E}[F(\hat{x}) - F(x_*)] \leq \mathbb{E}[F(\bar{x}) - F(x_*)] - \mathbb{E} \left[O \left(\frac{\sum_{i=1}^{N/2} X_i^2 - \left(\sum_{i=1}^{N/2} X_i \right)^2}{\left(\sum_{i=1}^{N/2} X_i \right)} \right) \right].$$

Essentially, we obtain a significant gain if there is a large variance in the population loss values in the first half of the online learning procedure. Intuitively, this is exactly where we should expect the largest change in population loss. However, one seemingly lossy step in the above is the partitioning of data and the use of some points twice, once to generate iterates and once to generate our adaptive weighting. In Section 8 we show that in general, a natural single-pass alternative is impossible.

8. Sequentially dependent x_i 's

As discussed, in Section 7, we use many of the data points in our training sample twice: once to generate iterates for the online algorithm, and once to generate our adaptive weighting. One might wonder if this is necessary: can we simply use the single-pass of observations generated during the online optimization procedure to also set the averaging weights? In this section we see that without some further assumptions or relaxation, the deterministic uniform random strategy is already admissible.

To formalize the situation, we consider the case that the x_i are not fixed up-front, but are in fact dependent on the y_i in a sequential manner. That is, the x_i are random variables satisfying $x_i = F_i(y_1, \dots, y_{i-1})$. This models certain sequential problems in which the x_i represent decisions that are themselves generated via some adaptive process. This might be the case if the x_i are

iterates of some online optimization procedure, for which x_{t+1} depends on the previously observed loss value y_t . In this case, we show that it is impossible to dominate the uniform strategy.

Theorem 11 *Let $N = 2$ and let $S : [0, 1]^N \rightarrow \Delta_2$, and suppose S is not equal to $(1/2, 1/2)$ for all inputs. Then there is a $x_1 \in [0, 1]$ and a function $x_2 : \{0, 1\} \rightarrow [0, 1]$ such that with random variables $y_1 \sim \text{Bern}(x_1)$ and $y_2 \sim \text{Bern}(x_2(y_1))$, we have:*

$$\mathbb{E}[\langle S(y_1, y_2), x \rangle] < \mathbb{E}[\langle (1/2, 1/2), x \rangle].$$

At first glance, this suggests that our adaptive weighting schemes cannot be extended to improve the classical online-to-batch conversion. However, in Section 7 we already showed that in fact the stochastic optimization setup has slightly more structure, which does enable us to improve.

9. Impossibility of Simultaneous Dominance

In this section, we show that it is not possible to design a strategy that simultaneously dominates more than one benchmark strategy.

Theorem 12 *Let $n = 2$ and let $S_1 = (p, 1 - p)$ and $S_2 = (p', 1 - p')$ be two distinct benchmark strategies. There exists no strategy $S : [0, 1]^2 \rightarrow \Delta_2$ that simultaneously dominates S_1 and S_2 .*

Proof Suppose for contradiction that there exists $S : [0, 1]^2 \rightarrow \Delta_2$ that dominates S_1 and S_2 . By definition, we must have $\text{value}(S; D) \geq \text{value}(S_i; D)$, for all product distributions $D = D_1 \times D_2$. We restrict our attention to Bernoulli distributions, $D_i = \text{Bern}(x_i)$ for arbitrary $x_1, x_2 \in [0, 1]$. Under this restriction, the samples y_1 and y_2 only take values in $\{0, 1\}$. So the strategy S is only evaluated at four points. For $i, j \in \{0, 1\}$, let $a_{i,j} := S(i, j)_1 \in [0, 1]$ be the four constants that fully determine strategy S .

Let $q(x_1, x_2)$ be the expected probability that S chooses item 1. Because the distributions are Bernoulli, this is exactly:

$$q(x_1, x_2) = a_{0,0}(1 - x_1)(1 - x_2) + a_{0,1}(1 - x_1)(x_2) + a_{1,0}(x_1)(1 - x_2) + a_{1,1}(x_1)(x_2).$$

Crucially, note that $q(x_1, x_2)$ is a polynomial in (x_1, x_2) and is thus continuous with respect to (x_1, x_2) . We now compute the value of strategy S and the two benchmark strategies S_1 and S_2 .

$$\begin{aligned} \text{value}(S; D) &= x_2 + (x_1 - x_2)q(x_1, x_2) \\ \text{value}(S_1; D) &= x_2 + (x_1 - x_2)p \\ \text{value}(S_2; D) &= x_2 + (x_1 - x_2)p'. \end{aligned}$$

Since $\text{value}(S; D) \geq \text{value}(S_i; D)$ for all product distributions, we have for all $(x_1, x_2) \in [0, 1]^2$:

$$(x_1 - x_2)(q(x_1, x_2) - p) \geq 0, \quad \text{and} \quad (x_1 - x_2)(q(x_1, x_2) - p') \geq 0.$$

Since $q(x_1, x_2)$ is continuous, taking the limit $x_1 \rightarrow x_2$ requires that $q_1(x, x) = p$ to satisfy the first inequality, and similarly requires that $q_1(x, x) = p'$ to satisfy the second inequality. But since $p \neq p'$, this is a contradiction. \blacksquare

10. Conclusion

The problem we study in this paper is both fundamental and pleasingly simple: given a single observation for each of n values, how well can you pick the largest value? Simply selecting a value at random is minimax optimal, but is intuitively unsatisfying because it does not depend on the data. On the other hand, the ERM rule, which is admissible, may be too brittle. We take a principled approach to resolve this issue: we ask for admissibility subject to the constraint that our strategy must dominate a benchmark.

While we study perhaps the simplest and most basic setting, our work raises natural follow-up questions. For example, what should be done if more than one sample is available? What if the observations are not bounded? What is an admissible strategy for stochastic convex optimization that dominates SGD with iterate averaging? In general, we hope our dual focus on admissibility and dominating a given benchmark may inspire further work in the machine learning theory community.

References

- Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.
- Robert E Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Stat.*, pages 16–39, 1954.
- Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.
- Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *ALT*, pages 23–37, 2009.
- Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *TOIT*, 50:2050–2057, 2004.
- José R. Correa, Pedro Foncea, and Leah Epstein. The prophet inequality for independent random variables with unknown distributions. In *SODA*, pages 194–211, 2017.
- Ashok Cutkosky. Anytime online-to-batch, optimism and acceleration. In *ICML*, pages 1446–1454, 2019.
- Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv*, 2310.07831, 2023.
- Ofer Dekel. From online to batch learning with cutoff-averaging. In *NIPS*, 2008.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12(7), 2011.
- Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2012.

- Theodore P. Hill and Robert P. Kertz. Comparisons of stop rule and supremum expectations of i.i.d. random variables. *Ann. Prob.*, 10(2):336–345, 1982.
- Nicole Immorlica, Robert Kleinberg, Jennifer Lewi, Thanh Nguyen, and Morteza Zadimoghaddam. Prophet inequalities with limited information. In *EC*, pages 845–863, 2020.
- W. James and Charles Stein. Estimation with quadratic loss. *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, 1:361–379, 1961.
- Pooria Joulani, Anant Raj, Andras Gyorgy, and Csaba Szepesvari. A simpler approach to accelerated optimization: iterative averaging meets optimism. In *ICML*, pages 4984–4993, 2020.
- Ali Kavis, Kfir Y Levy, Francis Bach, and Volkan Cevher. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *NeurIPS*, 2019.
- Ulrich Krengel and Louis Sucheston. Semiamarts and finite values. *Bull. Amer. Math. Soc.*, 83(4):745–747, 1977.
- Oren Lachish. Tight prophet inequalities for identically distributed random variables. *Ann. Appl. Prob.*, 31(3):1200–1216, 2021.
- Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $o(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv*, 1212.2002, 2012.
- Kfir Levy. Online to offline conversions, universality and adaptive minibatch sizes. In *NIPS*, 2017.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agueria y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS*, pages 1273–1282, 2017.
- David Pollard. *Convergence of Stochastic Processes*. Springer Science & Business Media, 2012.
- Herbert Robbins. An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp. Math. Stat. Prob.*, 1:157–163, 1956.
- Christian P Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2007.
- Aviad Rubinstein, Sahil Singla, and Jingyan Wang. Beyond the prophet: When waiting is not an option. In *SODA*, pages 2194–2213, 2020.
- Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Symp. Math. Stat. Prob.*, 1:197–206, 1956.
- Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2013.

Appendix A. Equivalence is not equality

In this section, we study the notions of equivalence for strategies.

Definition 4 (Equivalence) *Strategies S and S' are equivalent, denoted $S \equiv S'$, if $\text{value}(S; D) = \text{value}(S'; D)$ for all product distributions D .*

In this section we show that $S \equiv S'$ does not imply $S = S'$. We will construct an explicit example; let $n = 3$. Let S be the uniform strategy, i.e., $S(y) = (1/3, 1/3, 1/3)$ for all $y \in \{0, 1\}^3$ and D be the product of Bernoulli distributions. Define S' via the following table for any $c \in (0, 1/3]$:

y	$\Pr_{Y \sim \text{Bern}(x)}[Y = y]$	S'
000	$(1 - x_1)(1 - x_2)(1 - x_3)$	$(1/3, 1/3, 1/3)$
001	$(1 - x_1)(1 - x_2)x_3$	$(1/3 - c, 1/3 + c, 1/3)$
010	$(1 - x_1)x_2(1 - x_3)$	$(1/3 + c, 1/3, 1/3 - c)$
011	$(1 - x_1)x_2x_3$	$(1/3, 1/3 + c, 1/3 - c)$
100	$x_1(1 - x_2)(1 - x_3)$	$(1/3, 1/3 - c, 1/3 + c)$
101	$x_1(1 - x_2)x_3$	$(1/3 - c, 1/3, 1/3 + c)$
110	$x_1x_2(1 - x_3)$	$(1/3 + c, 1/3 - c, 1/3)$
111	$x_1x_2x_3$	$(1/3, 1/3, 1/3)$

Clearly S and S' are different strategies. Now we only need to check $S \equiv S'$, i.e., we need to check if $\forall x \in (0, 1)^3$, it holds that

$$\sum_{y \in \{0,1\}^N} \Pr_{Y \sim \text{Bern}(x)}[Y = y] \cdot \langle (S(y) - S'(y)), x \rangle = 0. \quad (5)$$

Indeed, since S' and S agree on two of the y 's we only focus on the six where they differ and dropping c for simplicity, the left-hand side of (5) evaluates to

$$\begin{aligned} & (1 - x_1)(1 - x_2)x_3 \cdot (-x_1 + x_2) + (1 - x_1)x_2(1 - x_3) \cdot (x_1 - x_3) \\ & + (1 - x_1)x_2x_3 \cdot (x_2 - x_3) + x_1(1 - x_2)(1 - x_3) \cdot (-x_2 + x_3) \\ & + x_1(1 - x_2)x_3 \cdot (-x_1 + x_3) + x_1x_2(1 - x_3) \cdot (x_1 - x_2) \\ & = (1 - x_1)(x_1x_2 - x_1x_3) + x_1(x_3 - x_2 - x_1x_3 + x_1x_2) \\ & = 0. \end{aligned}$$

Appendix B. Missing proofs

B.1. Proofs from Section 5

Lemma 13 *Equations (1) and (2) used to define the S_{Bern} strategy are equivalent.*

Proof Since the first term in both the equations is identical, we need to prove the identity

$$\sum_{S:i \in S} \frac{1}{|S|} \prod_{j \in S} y_j \prod_{k \notin S} (1 - y_k) = \sum_{S:i \in S} \frac{(-1)^{|S|-1}}{|S|} \prod_{j \in S} y_j.$$

First, the term $\prod_{k \notin S} (1 - y_k)$ can be expanded using standard binomial expansion:

$$\prod_{k \notin S} (1 - y_k) = \sum_{T \subseteq [n] \setminus S} (-1)^{|T|} \prod_{\ell \in T} y_\ell.$$

Substituting this into the left-hand side and simplifying yields:

$$\text{LHS} = \sum_{S:i \in S} \sum_{T \subseteq [n] \setminus S} \frac{(-1)^{|T|}}{|S|} \prod_{\ell \in S \cup T} y_\ell.$$

Since S and T are disjoint, the union $U = S \cup T$ is a set that contains i . We can rewrite the double summation by first summing over the combined set U .

$$\text{LHS} = \sum_{U:i \in U} \sum_{S:i \in S, S \subseteq U} \frac{(-1)^{|U|-|S|}}{|S|} \prod_{\ell \in U} y_\ell.$$

Let $k = |S|$ and $m = |U|$. Since i must be in S , the number of such sets S is exactly $\binom{m-1}{k-1}$. So the inner sum simplifies as:

$$\sum_{S:i \in S, S \subseteq U} \frac{(-1)^{|U|-|S|}}{|S|} = \sum_{k=1}^m \binom{m-1}{k-1} \frac{(-1)^{m-k}}{k}$$

Using the binomial identity $m \cdot \binom{m-1}{k-1} = k \cdot \binom{m}{k}$, the expression simplifies to:

$$= \sum_{k=1}^m \binom{m}{k} \frac{(-1)^{m-k}}{m}$$

From the binomial theorem, we know that $\sum_{k=0}^m \binom{m}{k} (-1)^{m-k} = (1-1)^m = 0$. Therefore $\sum_{k=1}^m \binom{m}{k} (-1)^{m-k} = 0 - \binom{m}{0} (-1)^m = -(-1)^m = (-1)^{m-1}$. Substituting back into the LHS:

$$\text{LHS} = \sum_{U:i \in U} \frac{(-1)^{|U|-1}}{|U|} \prod_{\ell \in U} y_\ell = \text{RHS}. \quad \blacksquare$$

Lemma 1 For any two indices i and j , $q_i(x) - q_j(x) = (x_i - x_j) \cdot K_{i,j}(x)$ where

$$K_{i,j}(x) = \int_{t=0}^1 \left(\prod_{k \notin \{i,j\}} (1 - tx_k) \right) dt.$$

Further, $K_{i,j}(x) \geq 0$, so $x_i > x_j \iff q_i(x) > q_j(x)$.

Proof From the definition of $q_i(\cdot)$, $q_j(\cdot)$, we obtain

$$q_i(x) - q_j(x) = \sum_{S:i \in S} \frac{(-1)^{|S|-1}}{|S|} \prod_{k \in S} x_k - \sum_{S:j \in S} \frac{(-1)^{|S|-1}}{|S|} \prod_{k \in S} x_k$$

after eliminating sets that contain both i and j since they cancel out,

$$= \sum_{S:i \in S, j \notin S} \frac{(-1)^{|S|-1}}{|S|} \prod_{k \in S} x_k - \sum_{S:j \in S, i \notin S} \frac{(-1)^{|S|-1}}{|S|} \prod_{k \in S} x_k$$

splitting the set S in the first term as $S = U \cup \{i\}$ and then factoring out x_i (and doing similarly for the second term),

$$\begin{aligned}
 &= x_i \cdot \left(\sum_{U:\{i,j\} \notin U} \frac{(-1)^{|U|}}{|U|+1} \prod_{k \in U} x_k \right) - x_j \cdot \left(\sum_{U:\{i,j\} \notin U} \frac{(-1)^{|U|}}{|U|+1} \prod_{k \in U} x_k \right) \\
 &= (x_i - x_j) \cdot \left(\sum_{U:\{i,j\} \notin U} \frac{(-1)^{|U|}}{|U|+1} \prod_{k \in U} x_k \right) \\
 &= (x_i - x_j) \cdot K_{i,j}(x),
 \end{aligned}$$

where we define

$$K_{i,j}(x) = \sum_{U:\{i,j\} \notin U} \frac{(-1)^{|U|}}{|U|+1} \prod_{k \in U} x_k.$$

To complete the proof, we further simplify it. Using the identity $\frac{1}{k+1} = \int_0^1 t^k dt$ to get rid of the denominator term, we obtain

$$\begin{aligned}
 K_{i,j}(x) &= \sum_{U:\{i,j\} \notin U} (-1)^{|U|} \cdot \int_{t=0}^1 t^{|U|} dt \cdot \prod_{k \in U} x_k \\
 &= \int_{t=0}^1 \left(\sum_{U:\{i,j\} \notin U} (-1)^{|U|} \cdot t^{|U|} \cdot \prod_{k \in U} x_k \right) dt \\
 &= \int_{t=0}^1 \left(\sum_{U:\{i,j\} \notin U} \prod_{k \in U} (-tx_k) \right) dt
 \end{aligned}$$

simplifying the sum over all subsets via binomial expansion: $\prod_k (1 + a_k) = \sum_U \prod_{k \in U} a_k$.

$$= \int_{t=0}^1 \left(\prod_{k \notin \{i,j\}} (1 - tx_k) \right) dt.$$

But for any $t \in [0, 1]$, we have $1 - tx_k \geq 0$, so the integral is always non-negative as desired and we have the lemma. \blacksquare

Lemma 3 (Monotonicity) $F(x) = F(x_1, \dots, x_n)$ is a monotonically decreasing function of each of its variables $x_k \in [0, 1]$.

Proof We show that $\frac{\partial F}{\partial x_k} \leq 0$. By symmetry, we only need to check for x_n . Let $S_{n-1} = \sum_{k=1}^{n-1} x_k$ and $P_{n-1} = \prod_{k=1}^{n-1} (1 - x_k)$. Then

$$F = \frac{1 - P_{n-1}(1 - x_n)}{S_{n-1} + x_n}.$$

The partial derivative is:

$$\frac{\partial F}{\partial x_n} = \frac{P_{n-1}(S_{n-1} + x_n) - (1 - P_{n-1}(1 - x_n)) \cdot 1}{(S_{n-1} + x_n)^2} = \frac{P_{n-1}(S_{n-1} + 1) - 1}{(S_{n-1} + x_n)^2}.$$

The sign is determined by the numerator. The lemma is true if $P_{n-1}(S_{n-1} + 1) \leq 1$, which holds if

$$\left(1 + \sum_{k=1}^{n-1} x_k\right) \prod_{k=1}^{n-1} (1 - x_k) \leq 1.$$

This is a known inequality; we provide the proof by induction for completeness. The base case (i.e., $n = 2$) clearly holds since $(1 + x_1)(1 - x_1) = 1 - x_1^2 \leq 1$. Inductively, let $S_k = \sum_{i=1}^k x_i$, $P_k = \prod_{i=1}^k (1 - x_i)$, and by the induction hypothesis, $P_k(1 + S_k) \leq 1$. Then, for $n = k + 1$,

$$\begin{aligned} (1 + S_k + x_{k+1})P_k(1 - x_{k+1}) &= P_k(1 + S_k - S_k x_{k+1} - x_{k+1}^2) \\ &= P_k(1 + S_k) - P_k x_{k+1}(S_k + x_{k+1}) \\ &\leq 1 - P_k x_{k+1}(S_k + x_{k+1}) \leq 1, \end{aligned}$$

where the last step follows because P_k, x_{k+1}, S_k are all non-negative. ■

Lemma 14 For any $x_1, \dots, x_n \in [0, 1]$, we have $(1 - \prod_{k=1}^n (1 - x_k)) \geq \mu(x)$.

Proof Let $z_i = 1 - x_i$ and note that $z_i \in [0, 1]$. Hence, $z_i \geq \prod_{k=1}^n z_k$, for all $i \in [n]$. Averaging this inequality over all $i \in [n]$, we obtain:

$$\frac{1}{n} \sum_{i=1}^n z_i \geq \frac{1}{n} \sum_{i=1}^n \left(\prod_{k=1}^n z_k \right) = \prod_{k=1}^n z_k. \quad (6)$$

Now,

$$1 - \prod_{k=1}^n (1 - x_k) = 1 - \prod_{k=1}^n z_k \geq 1 - \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n (1 - z_i) = \frac{1}{n} \sum_{i=1}^n x_i = \mu(x),$$

where the inequality follows from rearranging (6). ■

B.2. Proofs from Section 6

Theorem 9 We have the following two lower bounds:

- $\text{value}(S_{\text{peel}}; D) \geq \text{value}(S_{\text{det}}; D) + \sum_{k=1}^n w_k \cdot \frac{n-k+1}{n-k} \cdot \sigma^2(x; k),$
- $\text{value}(S_{\text{peel}}; D) \geq \text{value}(S_{\text{det}}; D) + \sum_{k=1}^n w_k \cdot \left(1 - \prod_{j=k}^n (1 - x_j)\right) \cdot \frac{\sigma^2(x; k)}{\mu(x; k)},$

where $\mu(x; k) = \frac{\sum_{j=k}^n x_j}{n-k+1}$ and $\sigma^2(x; k) = \left(\frac{1}{n-k+1} \cdot \sum_{j=k}^n x_j^2\right) - \mu(x; k)^2$ are defined to be the mean and variance of the components of x with indices k or higher.

Proof Since S_{peel} is defined to be a linear combination of n different S_{Bern} strategies, we can again use the multilinearity of S_{Bern} to easily compute its value. Lemma 5 gives us the following two bounds.

$$\begin{aligned} \text{value}(S_{\text{Bern}}^k; D) &\geq \mu(x; k) + \frac{n-k+1}{n-k} \cdot \sigma^2(x; k), \text{ and} \\ \text{value}(S_{\text{Bern}}^k; D) &\geq \mu(x; k) + (1 - \prod_{j=k}^n (1-x_j)) \cdot \frac{\sigma^2(x; k)}{\mu(x; k)}. \end{aligned}$$

We can now compute the value of S_{peel} as follows:

$$\text{value}(S_{\text{peel}}; D) = \sum_{k=1}^n w_k \cdot \text{value}(S_{\text{Bern}}^k; D)$$

Using just the first bound gives:

$$\begin{aligned} &\geq \sum_{k=1}^n w_k \left(\mu(x; k) + \frac{n-k+1}{n-k} \cdot \sigma^2(x; k) \right) \\ &= \sum_{k=1}^n p_k x_k + \sum_{k=1}^n w_k \cdot \frac{n-k+1}{n-k} \cdot \sigma^2(x; k) \\ &= \text{value}(S_{\text{det}}; D) + \sum_{k=1}^n w_k \cdot \frac{n-k+1}{n-k} \cdot \sigma^2(x; k). \end{aligned}$$

Similarly, using the second bound gives:

$$\text{value}(S_{\text{peel}}; D) \geq \text{value}(S_{\text{det}}; D) + \sum_{k=1}^n w_k (1 - \prod_{j=k}^n (1-x_j)) \frac{\sigma^2(x; k)}{\mu(x; k)}. \quad \blacksquare$$

Lemma 15 Consider the outcome $y = (0, \dots, 0)$. If $S \succeq S_{\text{peel}}$, then $S(y) = S_{\text{peel}}(y)$.

Proof Let $y = (0, \dots, 0)$. Suppose for contradiction that $S(y) \neq S_{\text{peel}}(y)$. Then since both $S(y)$ and $S_{\text{peel}}(y)$ are in Δ_n , $\exists j \in [n]$ such that $S(y)_j < S_{\text{peel}}(y)_j$. For convenience, let $\delta(j) = S_{\text{peel}}(y)_j - S(y)_j > 0$. We construct a product distribution D , parametrized by a small $\epsilon > 0$, as follows:

- $D_j = \text{Bern}(p)$; hence $x_j = p$.
- For each $k \neq j$, D_k is a point mass at 0; hence $x_k = 0, \forall k \neq j$.

Under this distribution D , the only outcomes are $y = (0, \dots, 0)$ or the outcome z that has $z_j = 1$ and all other $z_k = 0, \forall k \neq j$.

$$\begin{aligned} \text{value}(S_{\text{peel}}; D) - \text{value}(S; D) &= \Pr[y_j = 0] \cdot (x_j \delta(j)) + \Pr[y_j = 1] \cdot (x_j (S_{\text{peel}}(z)_j - S(z)_j)) \\ &\geq (1-p)p\delta(j) + p(-p) > 0, \end{aligned}$$

for small enough p since $\delta(j) > 0$, contradicting the dominance of S . ■

Theorem 10 *If $S \succeq S_{\text{peel}}$, then $S = S_{\text{peel}}$. Hence S_{peel} is admissible.*

Proof Suppose for the sake of contradiction that $S \succeq S_{\text{peel}}$ but $S \neq S_{\text{peel}}$. Let $\#_0(y) = |\{i : y_i = 0\}|$ be the number of zero coordinates in a vector y , let $\#_1(y) = |\{i : y_i = 1\}|$ be the number of one coordinates in a vector y , and let $i_{\max}(y) = \max\{i | y_i = 1\}$ be the index of last one coordinate if any and zero otherwise. Let y^* be an outcome such that $S(y^*) \neq S_{\text{peel}}(y^*)$ that lexicographically maximizes $(\#_0(y), \#_1(y), i_{\max}(y))$. Thus for any outcome y with $\#_0(y) > \#_0(y^*)$, we can assume that $S(y) = S_{\text{peel}}(y)$. Similarly, for any outcome y with $\#_0(y) = \#_0(y^*)$ and $\#_1(y) = \#_1(y^*)$ but $i_{\max}(y) > i_{\max}(y^*)$, we have $S(y) = S_{\text{peel}}(y)$.

Since $S(y), S_{\text{peel}}(y) \in \Delta_n$ and since $S(y^*) \neq S_{\text{peel}}(y^*)$, let $j \in [n]$ be an index such that $S(y^*)_j < S_{\text{peel}}(y^*)_j$. For convenience, let $\delta(y) := S_{\text{peel}}(y) - S(y)$; thus $\delta(y^*)_j > 0$. Let $Z = \{i | y_i^* = 0\}$ be the set of zero indices in y^* . We consider two cases depending on whether y_j^* is zero or not.

Case 1: $[j \notin Z]$. We construct a product distribution D , parametrized by small ϵ , over the outcomes as follows:

- D_j is a point mass at y_j^* ; hence $x_j = y_j^* > 0$.
- For each $i \in Z$, D_i is a point mass at 0; hence $x_i = 0, \forall i \in Z$
- For every other index k , $D_k = y_k^* \cdot \text{Bern}(\epsilon)$; hence $x_k = \epsilon \cdot y_k^*, \forall k \in [n] \setminus Z \setminus \{j\}$.

Note that under this distribution D , the only outcomes are either y^* or y with $\#_0(y) > \#_0(y^*)$ (where we know $S(y) = S_{\text{peel}}(y)$ by maximality). By choosing ϵ small enough, we can ensure that item j has the largest mean, i.e., $x_j > x_i, \forall i \neq j$. And for y^* , we know that S_{peel} puts more mass on item j , and so $\text{value}(S_{\text{peel}}; D) > \text{value}(S; D)$, contradicting the dominance of S .

Case 2: $[j \in Z]$ We construct a product distribution D , parametrized by small $p \in (0, 1)$, over the outcomes as follows:

- $D_j = \text{Bern}(p)$; hence $x_j = p$.
- For each $i \in Z \setminus \{j\}$, D_i is a point mass at 0; hence $x_i = 0, \forall i \in Z \setminus \{j\}$.
- For every other index k , $D_k = y_k^* \cdot \text{Bern}(q)$ for $q = p^2$; hence $x_k = q \cdot y_k^*, \forall k \in [n] \setminus Z$.

We first argue that $y_k^* < 1, \forall k > j$. Indeed, otherwise by definition we'll have that $S_{\text{peel}}(y^*)_j = 0$, which contradicts the choice of j . Now, let us consider the outcomes y in the support of distribution D . First, suppose $y_k = 0$ for any $k \notin Z$. Then either (i) we have $\#_0(y) > \#_0(y^*)$ (if $y_j = 0$), or (ii) we have $\#_0(y) \geq \#_0(y^*)$ and $\#_1(y) > \#_1(y^*)$ (if $y_j = 1$ and $y_k^* < 1$), or (iii) we have $\#_0(y) \geq \#_0(y^*)$ and $\#_1(y) = \#_1(y^*)$ and $i_{\max}(y) > i_{\max}(y^*)$ (if $y_j = 1$ and $y_k^* = 1$) since $y_k^* = 1 \implies k < j$. In either of the three cases, by the choice of y^* , we have $S(y) = S_{\text{peel}}(y)$. Hence, the only class of outcomes y from D to consider is when $y_k = y_k^*, \forall k \notin Z$.

There are only two outcomes in this class: either when $y_j = 0$ or $y_j = 1$. When $y_j = 0$, the outcome is exactly y^* ; call the outcome z^* when $y_j = 1$. Let $d = |[n] \setminus Z|$.

$$\begin{aligned} & \text{value}(S_{\text{peel}}; D) - \text{value}(S; D) \\ &= \Pr[y = y^*] \cdot \left(x_j \delta_j(y^*) + \sum_{k \notin Z} x_k \delta_k(y^*) \right) + \Pr[y = z^*] \cdot \left(x_j \delta_j(z^*) + \sum_{k \notin Z} x_k \delta_k(z^*) \right) \\ &= (1-p)q^d \cdot \left(p \delta_j(y^*) + \sum_{k \notin Z} q y_k^* \delta_k(y^*) \right) + p q^d \cdot \left(p \delta_j(z^*) + \sum_{k \notin Z} q y_k^* \delta_k(z^*) \right) \end{aligned}$$

We want to show that the above expression is positive for some appropriate value of p and q . We can loosely bound the term in the second parenthesis above by $-p$ as in the worst case S_{peel} gets zero value and S gets value p .

$$\geq (1-p)q^d \cdot \left(p\delta_j(y^*) + \sum_{k \notin Z} qy_k^* \delta_k(y^*) \right) - p^2q^d$$

Similarly $\sum_{k \notin Z} qy_k^* \delta_k(y^*) \geq -q$ since again in the worst case S_{peel} gets zero value but S gets the best item in Z which has value at most q .

$$\begin{aligned} &\geq (1-p)q^d \cdot (p\delta_j(y^*) - q) - p^2q^d = (1-p)pq^d\delta_j(y^*) - (1-p)q^{d+1} - p^2q^d \\ &= (1-p)p^{2d+1}\delta_j(y^*) - (1-p)p^{2d+2} - p^{2d+2} > 0, \end{aligned}$$

since $\delta_j(y^*) > 0$ and for p small enough, contradicting the dominance of S . \blacksquare

B.3. Proofs from Section 7

Theorem 16 *Suppose that $f : X \times Z \rightarrow [0, 1]$ is an arbitrary function. Let D_Z be an arbitrary distribution supported on Z and define $F(x) = \mathbb{E}_{z \sim D_Z}[f(x, z)]$. Suppose that N is even, and let z_1, \dots, z_N be an i.i.d. sample from $(D_Z)^N$ and let x_1, \dots, x_N be arbitrary random variables values in X subject to x_n being independent of z_n, \dots, z_N (i.e., x is adapted to the filtration associated with z_i). For $i \leq N/2$, define $y_i = 1 - f(x_i, z_{i+N/2})$. Then let $w_1, \dots, w_{N/2}$ be the result of applying our strategy (5) with uniform benchmark to $y_1, \dots, y_{N/2}$. Define $\hat{w}_i = w_i/2$ for $i \leq N/2$ and $\hat{w}_i = \frac{1}{N}$ for $i > N/2$. Select \hat{x} at random from x_1, \dots, x_N with $P(\hat{x} = x_i) = w_i$. Then we have for any $x_\star \in X$:*

$$\begin{aligned} \mathbb{E}[F(\hat{x}) - F(x_\star)] &\leq \frac{\sum_{i=1}^N \mathbb{E}[f(x_i, z_i) - f(x_\star, z_i)]}{N} \\ &\quad - \mathbb{E} \left[\left(1 - \prod_{i=1}^{N/2} F(x_i) \right) \frac{\sum_{i=1}^{N/2} (1 - F(x_i))^2 - \left(\sum_{i=1}^{N/2} 1 - F(x_i) \right)^2}{2 \left(\sum_{i=1}^{N/2} 1 - F(x_i) \right)} \right] \\ &\leq \frac{\sum_{i=1}^N \mathbb{E}[f(x_i, z_i) - f(x_\star, z_i)]}{N}. \end{aligned}$$

Proof By definition, we have that y_i is independent from y_j given $x_1, \dots, x_{N/2}$ for all i, j . Moreover, $E[y_i] = 1 - F(x_i)$. Therefore by Lemma 5, we have (conditioned on $x_1, \dots, x_{N/2}$)

$$\begin{aligned} \sum_{i=1}^{N/2} w_i(1 - F(x_i)) &\geq \frac{2}{N} \sum_{i=1}^{N/2} 1 - F(x_i) \\ &\quad + \left(1 - \prod_{i=1}^{N/2} F(x_i) \right) \frac{\sum_{i=1}^{N/2} (1 - F(x_i))^2 - \left(\sum_{i=1}^{N/2} 1 - F(x_i) \right)^2}{\left(\sum_{i=1}^{N/2} 1 - F(x_i) \right)} \end{aligned}$$

Therefore, taking expectation with respect to also $z_1, \dots, z_{N/2}$:

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^{N/2} w_i F(x_i) \right] &\leq \mathbb{E} \left[\frac{2}{N} \sum_{i=1}^N F(x_i) \right] \\ &= \mathbb{E} \left[\left(1 - \prod_{i=1}^{N/2} F(x_i) \right) \frac{\sum_{i=1}^{N/2} (1 - F(x_i))^2 - \left(\sum_{i=1}^{N/2} 1 - F(x_i) \right)^2}{2 \left(\sum_{i=1}^{N/2} 1 - F(x_i) \right)} \right]. \blacksquare \end{aligned}$$

B.4. Proofs from Section 8

Theorem 11 *Let $N = 2$ and let $S : [0, 1]^N \rightarrow \Delta_2$, and suppose S is not equal to $(1/2, 1/2)$ for all inputs. Then there is a $x_1 \in [0, 1]$ and a function $x_2 : \{0, 1\} \rightarrow [0, 1]$ such that with random variables $y_1 \sim \text{Bern}(x_1)$ and $y_2 \sim \text{Bern}(x_2(y_1))$, we have:*

$$\mathbb{E}[\langle S(y_1, y_2), x \rangle] < \mathbb{E}[\langle (1/2, 1/2), x \rangle].$$

Proof Write $S(i, j) = (w_{ij}, 1 - w_{ij})$ and $x_2(i) = x_2^i$ for $i, j \in \{0, 1\}$. Then, consider the quantity $\mathbb{E}[\langle S(y_1, y_2), x \rangle] - \mathbb{E}[\langle (1/2, 1/2), x \rangle] = \mathbb{E}[\langle S(y_1, y_2) - (1/2, 1/2), x \rangle]$. A little calculation shows:

$$\begin{aligned} &\mathbb{E}[\langle S(y_1, y_2) - (1/2, 1/2), x \rangle] \\ &= (x_1 - x_2^0)(w_{00} - \frac{1}{2})(1 - x_1)(1 - x_2^0) + (x_1 - x_2^0)(w_{01} - \frac{1}{2})(1 - x_1)x_2^0 \\ &\quad + (x_1 - x_2^1)(w_{10} - \frac{1}{2})x_1(1 - x_2^1) + (x_1 - x_2^1)(w_{11} - \frac{1}{2})x_1x_2^1 \\ &= (x_1 - x_2^0)(1 - x_1) \left((w_{00} - \frac{1}{2})(1 - x_2^0) + (w_{01} - \frac{1}{2})x_2^0 \right) \\ &\quad + (x_1 - x_2^1)x_1 \left((w_{10} - \frac{1}{2})(1 - x_2^1) + (w_{11} - \frac{1}{2})x_2^1 \right) \\ &= (x_1 - x_2^0)(1 - x_1) \cdot g(x_2^0) + (x_1 - x_2^1)x_1 \cdot h(x_2^1), \end{aligned}$$

where

$$g(z) = (w_{00} - \frac{1}{2})(1 - z) + (w_{01} - \frac{1}{2})z \text{ and } h(z) = (w_{10} - \frac{1}{2})(1 - z) + (w_{11} - \frac{1}{2})z.$$

Our goal is to identify a setting for x_1, x_2^0, x_2^1 such that the above expression is negative.

By assumption, at least one of the w_{ij} is different from $1/2$. Without loss of generality, let us assume it is either w_{00} or w_{01} ; thus $g(x_2^0) \neq 0$ (the argument is completely symmetric if instead it is one of the other two). Now, set $x_2^1 = x_1$; thus $(x_1 - x_2^1)x_1 \cdot h(x_2^1) = 0$. Let $x_2^0 = z$ be such that $g(z) \neq 0$. Then it suffices to exhibit x_1 such that $(x_1 - z)(1 - x_1) \cdot g(z)$ is negative.

We do a case analysis on the sign of $g(z)$. Pick any x_1 such that

$$x_1 \in \begin{cases} (0, z) & \text{if } g(z) > 0, \\ (z, 1) & \text{if } g(z) < 0. \end{cases}$$

With this choice, it is easy to see that $(x_1 - z)(1 - x_1) \cdot g(z) < 0$. ■

Appendix C. Applications

We now describe a few ML applications where adaptive weighted averaging arises naturally. Recall that our results show that our adaptive algorithms provably outperform any (non-adaptive) baseline, in expectation. The bounds show that our method is especially effective in the setting where the underlying means are sufficiently different from one another.

Online-to-batch conversion (Iterate averaging). Our original motivation for studying the problem is iterate averaging. In stochastic convex optimization, an online learning algorithm generates a sequence $\theta_1, \dots, \theta_T$ of iterates. The standard approach to convert these online iterates into a final solution is to output the uniform average $\bar{\theta}$. (E.g., see, (Bubeck, 2015)). However, later iterates are often superior to earlier ones, suggesting that uniform averaging is suboptimal (in practice; in the worst case, it is well known that uniform averaging obtains the min-max optimal results).

This setting fits into our framework with $n = T$: the x_i values correspond to the *utility* (negative loss) of the iterate θ_t . To obtain an estimate of x_t , we can use a small set/mini-batch of examples for each t , to obtain y_t . Using these values, we can use our results to obtain a new weighted average.

Re-weighting an ensemble of models under distribution shifts. Model ensembles are ubiquitous in modern ML applications. To solve a task, we train a collection M_1, \dots, M_n of models, with the intuition that (i) different M_i may perform better on different “regions” of the input space, and (ii) for uncertain inputs, a weighted average prediction can provide a more reliable output. For the training distribution \mathcal{D} , the training process ensures that the average of the model outputs yields the best prediction for an input $x \sim \mathcal{D}$.

Now suppose we have a shift in the distribution, obtaining \mathcal{D}' (some regions of the space are now more/less important than before). It is natural to ask if there is a better weighting of the models. Once again, this falls into our framework, where the x_i values correspond to the negated loss of M_i on \mathcal{D}' . We can estimate x_i as before, using a small number of samples from \mathcal{D}' to obtain a better method for combining the models.

Aggregation in federated learning. Our framework can also be applied when aggregating updates that have varying impacts on a model. For example, in the federated learning setting (McMahan et al., 2017), we have n clients each providing a candidate update to a model. The loss reduction (to a global objective of interest) provided by each update is now the x_i . The server can obtain an estimate y_i of this quantity using a minibatch as before; our framework then gives new weights for aggregation that can improve upon the simple average.