Defensive Generation

Gabriele Farina GFARINA@MIT.EDU
Massachusetts Institute of Technology

Juan Carlos Perdomo J.PERDOMO.SILVA@NYU.EDU
New York University

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the problem of efficiently producing, in an online fashion, generative models of scalar, multiclass, and vector-valued outcomes that cannot be falsified on the basis of the observed data and a pre-specified collection of computational tests. Our contributions are twofold. First, we expand on connections between online high-dimensional multicalibration with respect to an RKHS and recent advances in expected variational inequality problems, enabling efficient algorithms for the former. We then apply this algorithmic machinery to the problem of outcome indistinguishability. Our procedure, Defensive Generation, is the first to efficiently produce online outcome indistinguishable generative models of non-Bernoulli outcomes that are unfalsifiable with respect to infinite classes of tests, including those that examine higher-order moments of the generated distributions. Furthermore, our method runs in near-linear time in the number of samples and achieves the optimal, vanishing $1/\sqrt{T}$ rate for generation error.

Keywords: Outcome Indistinguishability, Defensive Forecasting, Multicalibration

1. Introduction

Supervised learning frames learning in terms of loss minimization. Under this paradigm, an algorithm succeeds if it is able to find a prediction rule that has low excess risk over the underlying data distribution, with respect to some loss and hypothesis class. Drawing on a rich set of ideas from complexity theory and cryptography, outcome indistinguishability offers a different perspective (Dwork et al., 2021). A learner succeeds if it finds a generative model of outcomes that cannot be falsified on the basis of the observed data and a pre-specified collection of computational tests.

The motivation behind outcome indistinguishability is that if no computational process can tell the difference between the "true" data produced by Nature and data produced by the Learner's model, then this model effectively *is* the "real" data generating process. The validity of the Learner's model is staked on its ability to withstand falsification.

In this paper, we study the problem of designing computationally efficient algorithms that provably find outcome indistinguishable generative models for rich classes of outcomes. We study this question in the online setting where data is arbitrarily (possibly adversarially) chosen. Samples are not assumed to be drawn i.i.d. from any distribution, yet the Learner is tasked with finding a generative model that, for all intents and purposes, "looks like" it generated the observed sequence.

Our main contribution is an efficient online procedure that finds generative models of scalar, multiclass, or vector-valued outcomes that are provably outcome indistinguishable with respect to rich, infinite collections of tests that live in a vector-valued reproducing kernel Hilbert space. A generative model here is a function that given a set of features x produces a conditional distribution μ over outcomes $\mathcal{Y}$. Prior work in this area was either restricted to the case of finding generative

models of simple binary outcomes, or required explicit enumeration over the set of tests (which ruled out efficiently guaranteeing indistinguishability with respect to super polynomially-sized classes). Our algorithm not only expands the scope of settings where one can guarantee indistinguishability, it also achieves the optimal $\sqrt{T}$ regret bound for this problem (see Vovk (2007) for a lower bound).

On a technical level, our result comes from reducing online outcome indistinguishability to high-dimensional multicalibration and using recent advances on expected variational inequalities (EVIs) to efficiently solve these underlying calibration problems. A core part of our work hence relies on exploring the spiderweb of connections between indistinguishability, online calibration, research on learning in games, and nonlinear optimization.

On a conceptual level, we highlight how our work provides a different perspective on generative modeling, based on online learning as its foundation. Much of the recent theoretical literature assumes that there is a fixed distribution we wish to sample from, and relies on function approximation assumptions to guarantee that the learned model is close in statistical distance to the truth. Our work on online OI, on the other hand, does not guarantee closeness in statistical distance nor does it rely on unverifiable function approximation assumptions. Instead, it provides an end-to-end unconditional guarantee that the learned model is *computationally* indistinguishable from the "truth"; where truth is in quotation marks since data is arbitrary and there is no underlying distribution to speak of.

Lastly, we produce these indistinguishable models through defensive forecasting, an algorithmic methodology pioneered in Vovk et al. (2005). Rather than trying to make a good guess around what the true data will be, defensive forecasting views predictions as a game. A good prediction is not one which aims to mimic Nature, but rather one that ensures that the generative model looks good in hindsight (from the perspective of the set of tests) *no matter* the choice of Nature.

We provide an overview of our contributions in Section 1.1. The interested reader can skip ahead to Section 1.2 to see example guarantees of our Defensive Generation algorithm in domains like language generation, learning linear dynamical systems, and weather forecasting. A technical overview is given in Section 2.

1.1. Overview of Contributions

We design algorithms that work in the following online protocol.

At every time step t, Nature first chooses features x_t. After observing x_t, the Learner produces a mixture $\mathcal{D}_t$ over conditional distributions μ_t on the outcome space $\mathcal{Y}$. Then, Nature, knowing the Learner's choice of $\mathcal{D}_t$, reveals an outcome $y_t \in \mathcal{Y}$.

We will associate each μ_t in the support of $\mathcal{D}_t$ with a vector of statistics p_t, where $p_t = g(\mu_t)$ for some g. For concreteness, the reader can think of p_t as the first moment of the distribution $p_t = \mathbb{E}_{\widetilde{y}_t \sim \mu_t}[\widetilde{y}_t]$. These statistics are given as input to the distinguisher to enhance the indistinguishability guarantee. Our goal is to design algorithms that achieve the following desideratum.

Definition 1 *An algorithm $\mathcal{A}$ satisfies online outcome indistinguishability with respect to a class of distinguishers $\mathcal{F}$ if it produces a sequence of mixtures $\mathcal{D}_t$ over conditional distributions $\mu_t \in \Delta(\mathcal{Y})$ such that for any $f \in \mathcal{F}$,*

$$\mathrm{OIGap}_T(f) := \left| \sum_{t=1}^{T} \mathbb{E}_{\mu_t \sim \mathcal{D}_t}[f(x_t, p_t, y_t)] - \sum_{t=1}^{T} \mathbb{E}_{\widetilde{y}_t \sim \mu_t, \mu_t \sim \mathcal{D}_t}[f(x_t, p_t, \widetilde{y}_t)] \right|, \tag{1}$$

is $o(T)$ regardless of Nature's choices of (x_t, y_t). Here, $p_t = g(\mu_t)$ for some fixed g. We define $\mathrm{OIGap}_T \coloneqq \sup_{f \in \mathcal{F}} \mathrm{OIGap}_T(f)$.

On the left, we have the outputs of the distinguishers f on the real outcomes y_t. On the right, there is the expected value of the distinguisher over the outcomes $\widetilde{y}_t \sim \mu_t$ sampled by the Learner's model. Dividing both sides of the above equation by T, an algorithm guarantees online outcome indistinguishability if the value of every function in $\mathcal{F}$ is the same when: (1) averaged over the realized sequence of outcomes y_t, or (2) the simulated outcomes $\widetilde{y}_t$. No distinguisher that has access to the generated samples $\widetilde{y}_t$ and the vector of statistics p_t can spot a difference between the real data and simulated data:

$$\lim_{T\to\infty} \left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mu_t \sim \mathcal{D}_t}[f(x_t, p_t, y_t)] - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\widetilde{y}_t \sim \mu_t, \mu_t \sim \mathcal{D}_t}[f(x_t, p_t, \widetilde{y}_t)] \right| = 0.$$

As a thought experiment, if the outcomes y_t were truly random and drawn from μ_t, by a martingale argument, we should expect that for any f, $\mathrm{OIGap}_T(f) \approx \sqrt{T}$. Our main contribution is a new algorithm, Defensive Generation, that achieves exactly the same guarantee, online, and for broad classes of adversarially chosen outcomes y_t.

Theorem 2 (Informal) *The following are true regarding the Defensive Generation procedure. Furthermore, in each setting, it runs in time $\mathcal{O}(\mathrm{poly}(d) \cdot \log(t))$ at time t and has $\mathrm{OIGap}_T \leq \mathcal{O}(\sqrt{T})$.*

1. *Multiclass outcomes, $\mathcal{Y} = [d]$. The procedure outputs distributions μ_t that are online outcome indistinguishable with respect to any infinite collection of distinguishers $f(x, p, y)$ that form an RKHS, and which have full access to the entire conditional distribution μ_t (i.e., $p_t = \mu_t$).*

   As a specific example, this in particular implies indistinguishability with respect to all linear functions at a rate bounded by $d\sqrt{T}$. See Example 1.

2. *Scalar outcomes, $\mathcal{Y} = [-1, 1]$. Defensive Generation produces distributions μ_t that are online outcome indistinguishable with respect to all low-degree tests in any RKHS that only examine the first d moments of μ_t. That is, $p_t = (1, \mathbb{E}_{\mu_t}[\widetilde{y}_t], \mathbb{E}_{\mu_t}[\widetilde{y}_t^2], \ldots, \mathbb{E}_{\mu_t}[\widetilde{y}_t^d])$.*

   For Boolean x, this implies that one can produce distributions over scalar $\widetilde{y}$ whose first d conditional moments, $\mathbb{E}[\widetilde{y}^j | c(x) = 1]$, for $j = 1$ to d, match those of the observed sequence over all subsets of the hypercube computable by low-depth decision trees $c(x)$. See Example 2.

3. *High-dimensional outcomes, $\mathcal{Y} = \{y \in \mathbb{R}^d : \|y\| \leq 1\}$. It outputs μ_t that are online OI with respect to distinguishers in any RKHS that examine the mean and covariance of μ_t. That is,*

$$p_t = (\mathbb{E}_{\mu_t}[\widetilde{y}_t], \mathbb{E}_{\mu_t}[\widetilde{y}_t \widetilde{y}_t^\top]).$$

   This in particular implies indistinguishability with respect to the infinite set of distinguishers that can lie in the span of a fixed set of (nonlinear) features $\Phi(x, p)$. See Example 4.

   Moreover, if the distinguishers only examine the first moments, $p = \mathbb{E}_{\mu_t}[\widetilde{y}_t]$, then the Defensive Generation algorithm works for any compact convex set in $\mathbb{R}^d$, not just the unit ball.

   See Example 3.

To the best of our knowledge, this is the first algorithm that can efficiently guarantee online outcome indistinguishability with respect to infinite classes $\mathcal{F}$ beyond the case of Bernoulli outcomes. As mentioned previously, the main backbone of this result is a new near-linear-time algorithm for online multicalibration in high-dimensions. We state this problem formally:

Definition 3 *Assume that at every time step, Nature selects features $x_t \in \mathcal{X}$ arbitrarily and then the Learner produces a distribution $\mathcal{P}_t$ over forecasts p_t in a compact, convex set $\mathcal{Z} \subset \mathbb{R}^d$. Lastly, Nature reveals the target $z_t \in \mathcal{Z}$. An algorithm $\mathcal{A}$ guarantees online multicalibration with respect to a class of functions $\mathcal{H} \subseteq \{\mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d\}$ if*

$$\sup_{h \in \mathcal{H}} \left| \sum_{t=1}^{T} \mathbb{E}_{p_t \sim \mathcal{P}_t} \left[h(x_t, p_t)^\top (z_t - p_t) \right] \right| \leq o(T)$$

regardless of Nature's choices of $(x_t, z_t) \in \mathcal{X} \times \mathcal{Z}$ in this online protocol.

Unlike other versions of online (ℓ_1) calibration studied in the literature (Peng, 2025; Fishelson et al., 2025) where errors are summed up over a grid of predicted values $p_t = v$, one can achieve error ε for this multicalibration problem after $\mathcal{O}(\varepsilon^{-2})$ many rounds for rich classes of functions $\mathcal{H}$.

This is in contrast to the lower bound of $d^{\,\mathrm{poly}(1/\varepsilon)}$ many rounds for the ℓ_1 version. In particular, we design an algorithm that guarantees $\sqrt{T}$ regret for any vector-valued RKHS $\mathcal{H}$. Our construction relies on recent advances in solving expected variational inequality problems (Zhang et al., 2025), along with the high-dimensional defensive forecasting algorithm from Dwork et al. (2025). We state its guarantees below:

Theorem 4 (Informal) *Let $\mathcal{H} \subseteq \{\mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d\}$ be any vector-valued RKHS with corresponding matrix valued kernel Γ. Then, the high-dimensional defensive forecasting algorithm can be efficiently implemented to run in time $\mathcal{O}(\mathrm{poly}(d)\log(t))$ at time step t and guarantee that for any $h \in \mathcal{H}$,*

$$\left| \sum_{t=1}^{T} \mathbb{E}_{p_t \sim \mathcal{P}_t} \left[h(x_t, p_t)^\top (z_t - p_t) \right] \right| \leq \|h\|_{\mathcal{H}} \sqrt{\sum_{t=1}^{T} \mathbb{E}_{p_t \sim \mathcal{P}_t} (z_t - p_t)^\top \Gamma((x_t, p_t), (x_t, p_t))(z_t - p_t)}.$$

Here, $\|\cdot\|_{\mathcal{H}}$ denotes the norm of the functions in the RKHS $\mathcal{H}$.

As a final note, prior work (Gopalan et al., 2023; Dwork et al., 2025; Okoroafor et al., 2025) has established that models that are OI are also loss minimizing. OI in fact implies loss minimization not just for a single loss, but rather for many losses simultaneously (that is, omniprediction) (Gopalan et al., 2022). By efficiently guaranteeing OI in higher-dimensions, we hope to enable future work on faster algorithms for omniprediction in richer domains.

1.2. Example Applications

We provide in this section a few different examples of generative modeling problems and the associated guarantees of the Defensive Generation algorithm.

Rather than highlighting the strongest possible guarantees, we focus on highlighting concrete examples and the intuition behind the indistinguishability definitions. Some of these example instantiations may be of independent interest.

Token-Based Sequence Modeling. To start, we consider token-based generation of sequences, such as language modeling. In this setting, the goal is to produce, given the history of tokens x_t, a conditional distribution μ_t over the next token $y_t \in \mathcal{Y} = [d]$ where $\mathcal{Y}$ is a vocabulary of d tokens.

In more detail, at each round $t = 1, 2, \ldots,$ we let x_t be the entire history of tokens observed so far $x_t = (y_1, y_2, \ldots, y_{t-1})$. Having seen x_t, the learner produces a distribution $\mu_t \in \Delta([d]) := \{\mu \in \mathbb{R}^d_{\geq 0} : 1^\top \mu = 1\}$. That is, a generative model of the next token given the history, $\Pr_{\mu_t}[y_t = \cdot \mid x_t]$. Having produced μ_t, the next token $y_t \in [d]$ is revealed.

In this setting, we can produce, as a simple example, distributions μ_t that are unfalsifiable with respect to tests that are linear functions of some feature embedding $\Phi(x_t)$ of the history, such as a frozen Transformer or BERT representation of the context. We define

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle, \qquad \Gamma(x, x') = I_d \, k(x, x'), \qquad G = \sup_t k(x_t, x_t) = \sup_t \|\Phi(x_t)\|_2^2.$$

Associated distinguishers take the form

$$f_h(x, p, y) = h(x)^\top \vec{y}, \text{ where } h(x) = W^\top \Phi(x) \in \mathbb{R}^d \text{ and } \vec{y} = (1\{y = 1\}, \ldots, 1\{y = d\})$$

with norm $\|h\|_{\mathcal{H}} = \|W\|_F$. In this case, Defensive Generation guarantees that for any such h,

$$\frac{1}{T} \cdot \mathrm{OIGap}_T(f_h) = \left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\big[h(x_t)^\top (\vec{y}_t - \mu_t)\big] \right| \leq 4\|h\|_{\mathcal{H}} \sqrt{\frac{G}{T}}.$$

This means, no matter how the tokens y_t are generated, every linear function of the embedded history $\Phi(x_t)$ looks the same when evaluated 1) over the empirical distribution of tokens $\sum_{t=1}^T \vec{y}_t$ or 2) the conditional distributions $\sum_{t=1}^T \mu_t$. We note that in this case the distinguishers f are only a function of x_t and the outcomes y_t. They ignore the side-information p_t, which in this case is exactly equal to the full conditional distribution μ_t. Also note that the generation error bound only depends on the norms of the functions and not explicitly on the ambient dimension.

The guarantees presented here follow directly from Theorem 9.

Predicting Correlated Rain Across the United States. Next, we illustrate how Defensive Generation can be used to produce conditional distributions over high-dimensional, real-valued outcomes. In particular, consider the problem of predicting how much it will rain in all 50 US states every day. Here, $y_t \in \mathbb{R}^{50}$ is a vector where $y_{t,i} \in \mathbb{R}$ denotes the amount of rainfall in state i on day t. We normalize precipitation units so that $\|y_t\|_2 \leq 1$.

Let $x_t \in \mathbb{R}^n$ be a vector of features (e.g. atmospheric measurements, seasonality information, prior rainfall, etc.) with ℓ_2 norm uniformly bounded by B. At every time t, the algorithm takes in x_t and outputs a joint distribution μ_t over vectors in $\mathbb{R}^{50}$ (rainfall in all 50 states).[1] It also outputs an auxiliary statistic $p_t = (v_t, Q_t)$ where $\mathbb{E}_{\widetilde{y}_t \sim \mu_t}[\widetilde{y}_t] = v_t$ and $\mathbb{E}_{\widetilde{y}_t \sim \mu_t}[\widetilde{y}_t \widetilde{y}_t^\top] = Q_t$ describing the mean and covariance of the conditional distributions of rain.

If we use the linear kernel, the algorithm guarantees (as a special case) indistinguishability with respect to all functions of the form

$$f_\theta(x_t, p_t, y_t) = y_{t,i} \cdot \theta^\top x_t, \quad f_\beta(x_t, p_t, y_t) = y_{t,i} \cdot y_{t,j} \cdot \beta^\top x_t, \quad f_w(x_t, p_t, y_t) = v_t^\top w \cdot y_{t,i},$$

1. As per Algorithm 3, this is an atomic measure supported on 51 points.

where w, β, θ are vectors in $\mathbb{R}^d$, $p_t = (v_t, Q_t) = (\mathbb{E}_{\mu_t}[\widetilde{y}_t], \mathbb{E}_{\mu_t}[\widetilde{y}_t \widetilde{y}_t^\top])$, and i, j are arbitrary state indices in $\{1, \ldots, 50\}$. Expanding the definition of indistinguishability and plugging in the guarantees for Defensive Generation we get the following statements:

$$\frac{1}{T} \cdot \mathrm{OIGap}(f_\theta) = \left| \frac{1}{T} \sum_{t=1}^{T} \theta^\top x_t (y_{t,i} - \mathbb{E}_{\mu_t}[\widetilde{y}_{t,i}]) \right| \leq 4\|\theta\|_2 \sqrt{\frac{(B^2+2)}{T}}, \quad \text{(correct mean per state)}$$

$$\frac{1}{T} \cdot \mathrm{OIGap}(f_\beta) = \left| \frac{1}{T} \sum_{t=1}^{T} \beta^\top x_t (y_{t,i} \cdot y_{t,j} - \mathbb{E}_{\mu_t}[\widetilde{y}_{t,i} \widetilde{y}_{t,j}]) \right| \leq 4\|\beta\|_2 \sqrt{\frac{(B^2+2)}{T}},$$

(correct covariances)

$$\frac{1}{T} \cdot \mathrm{OIGap}(f_w) = \left| \frac{1}{T} \sum_{t=1}^{T} w^\top \mathbb{E}_{\mu_t}[\widetilde{y}_t](y_{t,i} - \mathbb{E}_{\mu_t}[\widetilde{y}_{t,i}]) \right| \leq 4\|w\|_2 \sqrt{\frac{(B^2+2)}{T}}.$$

(self-consistent means)

Unpacking this a bit further, from the first set of conditions, we get that for every state i the expected value of rain is uncorrelated with any linear function of the features x. That is, the generative model of rain has means that are conditionally correct as per this test.

The second equation shows that not only are the means per state correct, the conditional distribution also captures the pairwise correlations between any pair of states. If a storm hits Massachusetts, it likely also hits Rhode Island. This second set of tests guarantees that the joint distribution of rain passes all these pairwise checks. It also asserts that the variances of the per state rainfall distributions are correct (i, j can be the same).

The last set of conditions shows that the means are not only conditionally correct, they are self-consistent in the sense that the errors in the expected value of rain in state i, $\mathbb{E}_{\mu_t}[\widetilde{y}_{t,i}] - y_{t,i}$, are uncorrelated with any linear function of the means of the distributions themselves. This is a strengthening of the indistinguishability guarantee and is not implied by either of the first two conditions.[2] See Foster and Kakade (2006) for further discussion of this point.

The guarantees presented in this example follow from Theorem 12 (see also Appendix F). We note how this example highlights the distinction between the conditional distributions μ_t produced by the algorithm and the statistics p_t consumed by the distinguisher. While Defensive Generation produces a probability measure over points in the unit ball, the distinguishers only examine the first and second moments of the distribution.

Learning Linear Dynamical Systems. Next, we consider the task of finding a generative model that is indistinguishable with respect to data generated by a linear dynamical system. In particular, consider the system

$$z_{t+1} = Az_t + \zeta_t, \qquad y_t = Cz_t + \nu_t.$$

Here, $y_t \in \mathbb{R}^d$ is the observation, z_t is the hidden state, A and C are matrices, and ζ_t and ν_t are noise vectors which could be adversarial (not i.i.d.).

Assume that observations y_t are uniformly bounded, $\|y_t\| \leq B$. This is true whenever the initial hidden state x_0 is bounded, the noise terms (ν_t, ζ_t) are bounded, and the matrix A has spectral radius strictly less than 1 (A is strictly stable). To keep notation simple, we set $B = 1$ (assuming strict

2. The covariances in this example are also self-consistent, we omit the equation for the sake of concision.

stability one can always renormalize). Now fix a history length $\ell \geq 1$ and set the features x_t in the online protocol to be the history of observations y_j up to some lag ℓ:

$$x_t := (y_{t-1}, y_{t-2}, \ldots, y_{t-\ell}) \in \mathbb{R}^{d\ell}.$$

Note that $\|x_t\|_2^2 \leq \ell$. Given x_t, the Defensive Generation algorithm produces a probability measure μ_t over points in $\mathbb{R}^d$ along with statistics $p_t = (\mathbb{E}_{\mu_t}[\widetilde{y}_t], \mathbb{E}_{\mu_t}[\widetilde{y}_t \widetilde{y}_t^\top])$.

Again using the linear kernel $\Gamma((x,p),(x',p')) = I \cdot x^\top x'$, by Theorem 12, the Defensive Generation algorithm guarantees indistinguishability with respect to all distinguishers of the form

$$f(x_t, p_t, y_t) = \sum_{i=1}^{d} x_t^\top \alpha_i \cdot y_{t,i} + \sum_{1 \leq i \leq j \leq d} x_t^\top \beta_{ij} \cdot (y_{t,i} y_{t,j}),$$

where α_i and $\beta_{i,j}$ are all vectors in $\mathbb{R}^d$. In particular, for any such function f,[3]

$$\mathrm{OIGap}(f) \leq \left(\sum_{i=1}^{d} \|\alpha_i\|_2 + \sum_{1 \leq i \leq j \leq d} \|\beta_{i,j}\|_2 \right) \sqrt{4T\ell}. \tag{2}$$

This in particular means that the Defensive Generation algorithm produces probability distributions μ_t that "look like" they generated the observations y_t, at least from the perspective of any linear function of the truncated history of observations,

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^{T} \left[\sum_{i=1}^{d} x_t^\top \alpha_i \cdot y_{t,i} + \sum_{1 \leq i \leq j \leq d} x_t^\top \beta_{ij} \cdot (y_{t,i} y_{t,j}) \right] \\ \approx &\frac{1}{T} \sum_{t=1}^{T} \left[\sum_{i=1}^{d} x_t^\top \alpha_i \cdot \mathbb{E}_{\mu_t}[\widetilde{y}_{t,i}] + \sum_{1 \leq i \leq j \leq d} x_t^\top \beta_{ij} \cdot \mathbb{E}_{\mu_t}[\widetilde{y}_{t,i} \widetilde{y}_{t,j}] \right]. \end{aligned}$$

This indistinguishability guarantee of the probability measures μ_t holds for any truncation length ℓ and with no stochasticity assumptions on the noise. Furthermore, the hidden state z_t can be infinite-dimensional and the pair of matrices (A, C) need not satisfy any observability conditions. We only required that the observations y_t had bounded norm.

Moreover, unlike other approaches to learning dynamical systems that only produce point estimates $\mathbb{E}_{\mu_t}[\widetilde{y}_t]$ for y_t, our algorithm produces a full probability measure with calibrated estimates of what the covariances $\mathbb{E}_{\mu_t}[\widetilde{y}\widetilde{y}^\top]$ will be.

We note that our procedure does not recover the matrices (A, C). It only learns to produce distributions that are indistinguishable with respect to the true data generated by the system. Using arguments from the literature (e.g., Simchowitz (2021)), it is plausible one can show that by setting ℓ in the order of $\mathcal{O}(\log(T))$, the means of the distributions μ_t will be as good as those produced by a Kalman filter. However, we defer a detailed investigation to future work.

3. Compared to the distinguishers in the previous example, which looked at each mean $y_{t,i}$ individually, each f in this example has a much larger scale, since it looks at all of the terms simultaneously. This is reflected by the term in parentheses in (2), which increases when the number of dimensions d increases.

Generative Models of Educational Outcomes. As a final example, we mention how the techniques we develop in this paper can provide fine-grained indistinguishability guarantees for generative models of social outcomes, like student test scores. Our algorithm can generate conditional distributions over scalar-valued outcomes that are valid even after conditioning on rich subsets of the features and examining higher-order moments.

To illustrate this, assume that at every time step t we see students with Boolean features $x_t \in \mathcal{X} = \{\pm 1\}^n$ and we want to output a probability distribution μ_t over scalar outcomes $y_t \in [0,1]$ describing their test scores on an exam. Using the polynomial kernel, we can guarantee that, not just the mean, but any fixed number of moments of these distributions will be accurate over all subpopulations $c(x) \subseteq \{\pm 1\}^n$ computable by depth-r decision trees.

$$\frac{1}{T}\sum_{t:c(x_t)=1}^{T} y_t^j \approx \frac{1}{T}\sum_{t:c(x_t)=1}^{T} \mathbb{E}_{\mu_t}[\widetilde{y}_t^j] \quad \text{for all } j = 1, \dots 2d.$$

Here, $c(x)$ is any Boolean function computable by a decision tree of depth r. The Defensive Generation algorithm guarantees that if we look at the subset of points in $\{\pm 1\}^n$ where $c(x) = 1$, the empirical moments of students' test scores y_t will match those of the conditional distributions μ_t. This guarantee holds not just for any fixed tree c or moment j, but rather for all trees and moments $j \in [2d]$ *simultaneously*.

In this case, we let each distinguisher have access to the vector of moments of the distribution μ_t produced by the algorithm, $p_t = (1, \mathbb{E}_{\mu_t}[\widetilde{y}_t], \dots, \mathbb{E}_{\mu_t}[\widetilde{y}_t^{2d}])$. Defensive Generation is able to produce these distributions efficiently, as we show in more technical terms in Example 2.

1.3. Related Work

The notion of outcome indistinguishability was first defined in the batch case and for binary outcomes by Dwork et al. (2021). They proved that OI was equivalent to the influential notion of multicalibration (Hébert-Johnson et al., 2018) if the distinguishers f cannot examine the underlying computational circuits that produce the samples. Using ideas on *moment* multicalibration from Jung et al. (2021), these equivalences between OI and multicalibration (for the batch setting) were extended to the case of non-Bernoulli outcomes by Dwork et al. (2022).

Following (and prior to) work on OI and multicalibration in the batch setting, there's been significant interest in extending the analyses to the online setting, with the vast majority of work focused on the case of predicting (equivalently, generating) a binary outcome. Some of this work goes back to Foster and Vohra (1998); Sandroni et al. (2003); Vovk (2007); Foster and Kakade (2006). Following Hébert-Johnson et al. (2018) there was renewed interest in the problem (Gupta et al., 2022; Okoroafor et al., 2025). Our work builds on that of Dwork et al. (2025) who focused on outcome indistinguishable models of binary outcomes with respect to scalar valued RKHSs.

Our work is also closely related to work designing algorithms for high-dimensional multicalibration. In this vein, Noarov et al. (2025) design the first algorithms for this problem. They achieve $\sqrt{T\log(|\mathcal{F}|)}$ regret but require enumerating over the functions in a finite set $\mathcal{F}$. Our work also builds on the idea of guaranteeing indistinguishability for richer outcome spaces and non-linear distinguishers by converting OI to a multicalibration problem in a higher-dimensional space where non-linear functions become linear in an expanded basis (Lu et al., 2025; Gopalan et al., 2023; Gupta et al., 2022). We build on these results to design computationally efficient algorithms that can cope with distinguisher classes $\mathcal{F}$ that are infinitely large.

Lastly, our technical approach relies and expands on the philosophy of *defensive forecasting*, a methodology introduced by Vovk et al. (2005) for online prediction where forecasts are derived by correcting past mistakes. Perdomo and Recht (2025) provide an overview of this line of work and show how it can be used to design algorithms for online calibration, quantile regression, and loss minimization (amongst others). Our core results essentially extend the meta-algorithm discussed in Perdomo and Recht (2025) to work in high-dimensional settings.

We defer a discussion of expected variational inequalities to Section A.

2. Technical Overview

Our work builds on layers of ideas from different research areas. In this section, we provide a conceptual overview of how these relate and lead to our final result.

Recall that the goal is, given features x_t, produce a mixture $\mathcal{D}_t$ over (conditional) distributions μ_t over outcomes in a set $\mathcal{Y}$ that withstands falsification with respect to the true revealed outcome y_t for that x_t. That is, $f(x_t, p_t, y_t) \approx \mathbb{E}_{\widetilde{y}_t \sim \mu_t}[f(x_t, p_t, \widetilde{y}_t)]$. This goal is well-defined even if the distinguishers only examine the features and the sampled outcome $f(x_t, p_t, y_t) = f(x_t, y_t)$. Allowing f access to the "side information" in p_t only strengthens the indistinguishability guarantee.[4]

For certain classes of outcomes y_t, we can guarantee indistinguishability even if we provide distinguishers with a full description of the conditional distribution μ_t from which we sample $\widetilde{y}_t$. That is, we let $p_t = \mu_t$. In the binary case, this conditional distribution is just a single number $\Pr_{\mu_t}[\widetilde{y}_t = 1]$. And in the multiclass case where $\mathcal{Y} = [d]$, this is a point on the simplex $\mu_{t,j} = \Pr_{\mu}[\widetilde{y}_t = j]$ for $j \in [d]$. However, for continuous outcomes, one might in principle require infinitely many parameters to specify the full conditional distribution over $\widetilde{y} \in [-1, 1]$.

To address this computational issue, when dealing with real-valued outcomes y, we restrict ourselves to guaranteeing *oblivious* outcome indistinguishability as defined in Dwork et al. (2022). Here, we restrict the class of distinguishers f to those that can be written as a function of a finite-dimensional vector of statistics $s(y)$ living in set $\mathcal{Z} \subset \mathbb{R}^d$,

$$f(x, p, y) = g_f(s(y), x, p).$$

For instance, if the distinguisher f only examines the first two moments of the distribution then we can write $f(x, p, y) = g_f(y, y^2, x, p)$ where $s(y) = (y, y^2)$ is a vector of sufficient statistics.

In many important cases, these distinguishers are in fact linear in the sufficient statistics. That is, for every f, there exists a function $h_f(x, p) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ such that $f(x, p, y)$ is equal to $h_f(x, p)^\top s(y)$ for all (x, p). This "linearization" is always true for the case of discrete outcomes since the conditional distribution is a sufficient statistic for any f. If we can write $f(x, p, y) = h(x, p)^\top s(y)$, then outcome indistinguishability reduces to online high-dimensional, multicalibra-

4. Using the terminology from Dwork et al. (2021), we operate within the sample-access formulation of OI.

tion with respect to the collection of functions $\mathcal{H}$ that depend on $\mathcal{F}$. More precisely,

$$\begin{aligned}\mathrm{OIGap}_T &= \sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\mathbb{E}_{p_t}[f(x_t,p_t,y_t)] - \sum_{t=1}^{T}\mathbb{E}_{p_t,\widetilde{y}_t\sim\mu_t}[f(x_t,p_t,\widetilde{y}_t)]\right| \\ &= \sup_{h\in\mathcal{H}}\left|\sum_{t=1}^{T}\mathbb{E}_{p_t}[h(x_t,p_t)^\top s(y_t)] - \mathbb{E}_{p_t,\mu_t}[h(x_t,p_t)^\top s(\widetilde{y}_t)]\right| \\ &= \sup_{h\in\mathcal{H}}\left|\sum_{t=1}^{T}\mathbb{E}_{p_t}[h(x_t,p_t)^\top (z_t-p_t)]\right|,\end{aligned}$$

where in the last equality we let $z_t = s(y_t)$ and set $p_t = \mathbb{E}_{\mu_t}[s(\widetilde{y}_t)]$.

Therefore, as long as we can solve for a distribution μ_t over $\mathcal{Y}$ such that $\mathbb{E}_{\mu_t}[s(\widetilde{y}_t)] = p_t$, we can reduce online outcome indistinguishability to high-dimensional multicalibration. In particular, to produce the mixture $\mathcal{D}_t$ over conditional distributions μ_t, given x_t, we first produce a multicalibrated distribution $\mathcal{P}_t$ over forecasts p_t such that $p_t \approx z_t = s(y_t)$. Then, for every p_t in the support of $\mathcal{P}_t$, we solve for a measure μ_t. The distribution $\mathcal{D}_t$ is simply the pushforward of $\mathcal{P}_t$ under the backfitting operation.[5]

Having established this reduction, our central algorithmic contribution is a procedure that efficiently guarantees online high-dimensional multicalibration with respect to any vector-valued reproducing kernel Hilbert space $\mathcal{H}$. These are rich spaces that can express functions such as all polynomials and which can be learned efficiently. We provide a brief primer for those unfamiliar.

A vector-valued RKHS is a Hilbert space of functions $\mathcal{H} \subseteq \{\mathcal{X}\times\mathcal{Z}\to\mathbb{R}^d\}$ that is equal to the closure of the set

$$h(x,p) = \sum_{i=1}^{n}\Gamma((x_i,p_i),(x,p))\theta_i.$$

Here, $\Gamma((x,p),(x',p')) \in \mathbb{R}^{d\times d}$ is a matrix-valued kernel, θ_i are vectors, and (x_i,p_i) are elements in $\mathcal{X}\times\mathcal{Z}$. In particular, Γ is a symmetric function whose outputs are positive semidefinite matrices and which satisfies the reproducing property. For any $h\in\mathcal{H}$ and $\theta\in\mathbb{R}^d$,

$$h(x,p)^\top\theta = \langle h, \Phi(x,p)\theta\rangle_{\mathcal{H}}$$

where $\Phi(x,p) = \Gamma(\cdot,(x,p))$ is the evaluation functional for the RKHS. By virtue of being a Hilbert space, an RKHS has a unique inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and induced norm $\|h\|_{\mathcal{H}} = \sqrt{\langle h,h\rangle_{\mathcal{H}}}$ that serves an instance-dependent notion of complexity.

A simple case to keep in mind is when $\Phi(x,p)\in\mathbb{R}^{r\times d}$ is a finite-dimensional feature map and $\Gamma((x,p),(x',p')) = \Phi(x,p)^\top\Phi(x',p')$. In this case the RKHS contains the set of functions $h(x,p) = \Phi(x,p)^\top\theta$ for $\theta\in\mathbb{R}^r$, the inner product is equal to the standard inner product in $\mathbb{R}^r$ $\langle\theta,\theta'\rangle_{\mathcal{H}} = \sum_{i=1}^{r}\theta_i\theta_i'$, and the norm of the functions $h\in\mathcal{H}$ is equal to $\|h\|_{\mathcal{H}} = \|h\|_2$.

To guarantee indistinguishability with respect to vector-valued RKHS, we make use of ideas from *defensive forecasting*, an algorithmic template introduced by Vovk et al. (2005). Defensive forecasting is a game-theoretic strategy for prediction where forecasts are derived not by prognostication, but rather by correcting past mistakes. Our main advancement here is to draw on recent

5. Solving for μ_t is trivial in the discrete case where $p_t = \mu_t$. It is less obvious for the scalar case where $p_t = (\mathbb{E}_{\mu_t}[\widetilde{y}],\ldots,\mathbb{E}_{\mu_t}[\widetilde{y}^{\,d}])$, but we show it can still be done by leveraging results from semi-algebraic optimization.

breakthroughs on algorithms for solving *expected variational inequalities* (EVI) to be able to efficiently do defensive forecasting in high dimensions (Zhang et al., 2025).

Expected variational inequalities have appeared in different areas with different names, including "outgoing minimax problems" (Foster and Hart, 2021) or "accuracy certificates" (Nemirovski et al., 2010). Given a desired tolerance $\varepsilon > 0$, a convex, compact set $\mathcal{Z} \subset \mathbb{R}^d$, and an operator $S : \mathcal{Z} \to \mathbb{R}^d$, the goal of an EVI is to find a (finitely supported) distribution $\mathcal{D} \in \Delta(\mathcal{Z})$ such that

$$\mathbb{E}_{p\sim\mathcal{D}}[S(p)^\top (z - p)] \leq 0$$

for all $z \in \mathcal{Z}$.

As we discuss in Section A, these problems always admit a $\mathcal{O}(\mathrm{poly}(d, 1/\varepsilon))$ time algorithm via a rather clean reduction to regret minimization.[6] While Nemirovski et al. (2010) proposed algorithms with $\log(1/\varepsilon)$ complexity for the limited case in which $S(\cdot)$ is a monotone operator, it was not until recently that these problems (and some generalizations) were shown to be solvable in $\mathcal{O}(\mathrm{poly}(d)\log(1/\varepsilon))$ time for general operators $S(\cdot)$ (Zhang et al., 2025). This is in stark contrast with traditional variational inequality problems ("find $p \in \mathcal{Z}$ such that $S(p)^\top(z - p) \leq 0$ for all $z \in \mathcal{Z}$"), which are computationally intractable already when S is a linear function and $\mathcal{Z}$ is the product of two simplices in light of standard connections with the problem of computing Nash equilibria in bimatrix games (Chen et al., 2009; Daskalakis et al., 2009).

Acknowledgments

GF was supported in part by the National Science Foundation award CCF-2443068, the Office of Naval Research grant N000142512296, and an AI2050 Early Career Fellowship. We would like to thank Benjamin Recht and the anonymous COLT reviewers for their helpful comments.

References

Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 2012.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player nash equilibria. *Journal of the ACM (JACM)*, 56(3):1–57, 2009.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *ACM Symposium on Theory of Computing*, 2021.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Beyond bernoulli:generating random outcomes that cannot be distinguished from nature. In *International Conference on Algorithmic Learning Theory*, 2022.

Cynthia Dwork, Chris Hays, Nicole Immorlica, Juan C Perdomo, and Pranay Tankala. From fairness to infinity: Outcome-indistinguishable (omni) prediction in evolving graphs. *Conference on Learning Theory*, 2025.

6. For simplicity, throughout the discussion we suppress polynomial dependence on the diameters of $\mathcal{Z}$ and S.

Gabriele Farina. Turning defense into offense in $O(\log 1/\varepsilon)$ steps: Efficient constructive proof of the minimax theorem. *ACM SIGecom Exchanges*, 2026.

Gabriele Farina and Charilaos Pipis. Polynomial-time computation of exact Φ-equilibria in polyhedral games. *Advances in Neural Information Processing Systems*, 2024.

Maxwell Fishelson, Noah Golowich, Mehryar Mohri, and Jon Schneider. High-dimensional calibration from swap regret. In *Conference on Neural Information Processing Systems*, 2025.

Dean P Foster and Sergiu Hart. Forecast hedging and calibration. *Journal of Political Economy*, 2021.

Dean P Foster and Sham M Kakade. Calibration via regression. In *IEEE Information Theory Workshop*, 2006.

Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 1997.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 1998.

Drew Fudenberg and David K Levine. An easier way to calibrate. *Games and Economic Behavior*, 1999.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *Innovations in Theoretical Computer Science Conference*, 2022.

Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. *Innovations in Theoretical Computer Science*, 2023.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*. 1993.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. *Innovations in Theoretical Computer Science*, 2022.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, 2018.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, 2021.

Yin Tat Lee, Aaron Sidford, and Santosh S Vempala. Efficient convex optimization with membership oracles. In *Conference On Learning Theory*, 2018.

Ehud Lehrer. Any inspection is manipulable. *Econometrica*, 2001.

Jiuyao Lu, Aaron Roth, and Mirah Shi. Sample efficient omniprediction and downstream swap regret for non-linear losses. *Conference on Learning Theory*, 2025.

Arkadi Nemirovski, Shmuel Onn, and Uriel G Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 2010.

Jiawang Nie. *Moment and Polynomial Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2023.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. *International Conference on Machine Learning*, 2025.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Princewill Okoroafor, Robert Kleinberg, and Michael P Kim. Near-optimal algorithms for omniprediction. *arXiv preprint arXiv:2501.17205*, 2025.

Pablo A Parrilo. Polynomial games and sum of squares optimization. In *Proceedings of the 45th IEEE Conference on Decision and Control*. IEEE, 2006.

Binghui Peng. High dimensional online calibration in polynomial time. *Symposium on the Foundations of Computer Science*, 2025.

Juan Carlos Perdomo and Benjamin Recht. In defense of defensive forecasting. *arXiv preprint arXiv:2506.11848*, 2025.

Alvaro Sandroni, Rann Smorodinsky, and Rakesh V Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 2003.

Max Simchowitz. *Statistical complexity and regret in linear control*. PhD thesis, University of California, Berkeley, 2021.

Vladimir Vovk. Non-asymptotic calibration and resolution. *Theoretical Computer Science*, 2007.

Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. In *International Workshop on Artificial Intelligence and Statistics*, 2005.

Vladimir A Yakubovich. S-procedure in nonlinear control theory. *Vestnik Leninggradskogo Universiteta, Ser. Matematika*, 1971.

Brian Hu Zhang, Ioannis Anagnostides, Emanuel Tewolde, Ratip Emin Berker, Gabriele Farina, Vincent Conitzer, and Tuomas Sandholm. Expected variational inequalities. *International Conference on Machine Learning*, 2025.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, 2003.

Appendix A. Efficient High-Dimensional Multicalibration

In this section, we present an efficient algorithm for high-dimensional multicalibration with respect to general vector-valued reproducing kernel Hilbert spaces. Throughout the section, we assume that $\mathcal{Z} \subset \mathbb{R}^d$ is a given convex body for which an efficient (i.e., responding to queries in time polynomial in d and logarithmic in the desired precision) projection, linear optimization, or separation oracle is available. These oracles are all known to be efficiently reducible to each other in time polynomial in the dimension d under minimal regularity assumptions that we assume are satisfied (Grötschel et al., 1993; Lee et al., 2018). Furthermore, we will denote with

$$D \coloneqq \max_{z,z' \in \mathcal{Z}} \|z - z'\|_2 \quad \text{and} \quad G \coloneqq \sup \|\Gamma(\cdot,\cdot)\|_{\text{op}} \tag{3}$$

the diameter of $\mathcal{Z}$ and the maximum operator norm of the matrix kernel Γ, respectively. Finally, we assume that the matrix kernel Γ can be evaluated in $\text{poly}(d)$ time for any input, and likewise that the feature map Φ can be evaluated in $\text{poly}(d, r)$ time for any input (if r is finite).

Given T samples, our algorithm guarantees order $\sqrt{T}$ regret in polynomial-time. At the heart of the construction, our algorithm relies on recent algorithmic advances for solving expected variational inequality problems efficiently (Zhang et al., 2025). To appreciate why EVIs are connected to multicalibration, and what operators arise in these contexts, it is instructive to look at the general template of *defensive forecasting*.

Defensive forecasting was introduced by Vovk et al. (2005), and can be thought of as a particular instantiation of the framework of Blackwell approachability. Related ideas have appeared throughout the literature, including Foster and Vohra (1997); Sandroni et al. (2003); Lehrer (2001); Fudenberg and Levine (1999); Foster and Hart (2021), and others. We refer the reader to Perdomo and Recht (2025) for an overview of the defensive forecasting philosophy. At all times t, defensive forecasting picks distributions $\mathcal{P}_t$ so that the quantity

$$Z_T = \left\| \sum_{t=1}^{T} \mathbb{E}_{p_t \sim \mathcal{P}_t} \Phi(x_t, p_t)(z_t - p_t) \right\|_{\mathcal{H}}$$

is guaranteed to be sublinear (as a function of T) no matter how Nature selects z_t (after $\mathcal{P}_t$ has been picked) and x_t (before $\mathcal{P}_t$ has been picked). This is sufficient for ensuring sublinear multicalibration error, since by the reproducing property:

$$\left| \sum_{t=1}^{T} \mathbb{E}_{p_t \sim \mathcal{P}_t} h(x_t, p_t)^\top (z_t - p_t) \right| = \left| \langle h, \sum_{t=1}^{T} \mathbb{E}_{p_t \sim \mathcal{P}_t} \Phi(x_t, p_t)(z_t - p_t) \rangle_{\mathcal{H}} \right| \leq \|h\|_{\mathcal{H}} Z_T.$$

Although, strictly speaking, Z_T is *not* a Blackwell approachability game due to the presence of the (potentially adversarial) context x_t in the otherwise bilinear objective (as a function of $\mathcal{P}_t$ and z_t), the same idea behind Blackwell's approachability algorithm yields a sublinear guarantee. In particular, observe the recursive expansion

$$Z_t^2 = Z_{t-1}^2 + \left\| \mathbb{E}_{p_t \sim \mathcal{P}_t} \Phi(x_t, p_t)(z_t - p_t) \right\|_{\mathcal{H}}^2 + 2\mathbb{E}_{p_t \sim \mathcal{P}_t} \Bigg[(z_t - p_t)^\top \underbrace{\Phi(x_t, p_t)^\top \sum_{\tau=1}^{t-1} \mathbb{E}_{p_\tau \sim \mathcal{P}_\tau} \Phi(x_\tau, p_\tau)(z_\tau - p_\tau)}_{=: S_t(p_t)} \Bigg].$$

Algorithm 1 Efficient High-Dimensional Multicalibration via Expected VIs

Matrix-valued kernel function $\Gamma : (\mathcal{X} \times \mathcal{Z})^2 \to \mathbb{R}^{d\times d}$

For $t = 1, 2, \ldots, T$:

See the features x_t and define $S_t : \mathcal{Z} \to \mathbb{R}^d$ as

$$S_t(p) = \Phi(x_t, p)^\top \sum_{\tau=1}^{t-1} \mathbb{E}_{p_\tau\sim\mathcal{P}_\tau} \Phi(x_\tau, p_\tau)(z_\tau - p_\tau) = \sum_{\tau=1}^{t-1} \mathbb{E}_{p_\tau\sim\mathcal{P}_\tau} \Gamma((x_t, p), (x_\tau, p_\tau))(z_\tau - p_\tau)$$

Predict $p_t \sim \mathcal{P}_t$, where the distribution $\mathcal{P}_t \in \Delta(\mathcal{Z})$ satisfies the expected variational inequality

$$\mathbb{E}_{p\sim\mathcal{P}_t}\Big[S_t(p)^\top (z - p)\Big] \le \varepsilon_t \qquad \forall z \in \mathcal{Z} \tag{EVI}$$

for some small error ε_t; for example, $\varepsilon_t = 1$ or $\varepsilon_t = 1/\sqrt{t}$

It is then clear that, as long as $\mathcal{P}_t$ is selected so that

$$\mathbb{E}_{p_t\sim\mathcal{P}_t}\Big[S_t(p_t)^\top (z - p_t)\Big] \le 0 \qquad \forall z \in \mathcal{Z}, \tag{4}$$

the third term can be dropped, yielding the bound

$$Z_T^2 \le \sum_{t=1}^{T} \big\|\mathbb{E}_{p_t\sim\mathcal{P}_t} \Phi(x_t, p_t)(z_t - p_t)\big\|_{\mathcal{H}}^2 \quad \Longrightarrow \quad Z_T \le \sqrt{\sum_{t=1}^{T} \big\|\mathbb{E}_{p_t\sim\mathcal{P}_t} \Phi(x_t, p_t)(z_t - p_t)\big\|_{\mathcal{H}}^2}.$$

This guarantees a multicalibration error growing at a $\sqrt{T}$ rate, with constants depending on D, G, and the norm of the functions in the RKHS. We summarize the procedure in Algorithm 1, incorporating the possibility of error ε_t on the right of (4). From the above discussion, we immediately derive the following guarantee.

Proposition 5 *Let Γ be a matrix-valued kernel with associated RKHS $\mathcal{H} \subseteq \{\mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d\}$. If the predictions $p_t \sim \mathcal{P}_t$ are made according to Algorithm 1, then for all $T \ge 1$ and $h \in \mathcal{H}$:*

$$\left|\sum_{t=1}^{T} \mathbb{E}_{p_t\sim\mathcal{P}_t} h(x_t, p_t)^\top (z_t - p_t)\right| \le \|h\|_{\mathcal{H}} \sqrt{TD^2G + \sum_{t=1}^{T} \varepsilon_t}.$$

Hence, as long as $\sum_{t=1}^{T} \varepsilon_t = \mathcal{O}(T)$, Algorithm 1 guarantees order $\mathcal{O}(\sqrt{T})$ multicalibration error.

Efficiently Solving the EVI. The key (and effectively, only) step in Algorithm 1 is constructing a distribution $\mathcal{P}_t$ over $\mathcal{Z}$ that solves the expected variational inequality problem (EVI). As mentioned in Section 1.3, a solution of an EVI can always be computed efficiently for any general, even discontinuous, bounded VI operator (in our case, S_t), given oracle access—for example, a membership, linear optimization, or separation oracle—to the convex compact domain (in our case, $\mathcal{Z}$). In the rest of this section, we sketch two known general approaches for solving EVIs, with runtimes scaling polynomially and polylogarithmically in the desired inverse precision $1/\varepsilon_t$, respectively.

To start, a distribution $\mathcal{P}_t$ such that (EVI) holds can be found in $\mathrm{poly}(d, D, G, 1/\varepsilon_t)$ time by no-regret algorithms (see also Zhang et al. (2025); Farina (2026)). In particular, consider setting up any no-regret learning algorithm outputting at every time k a point $p_k \in \mathcal{Z}$, and receiving as utilities (losses) the functions $p \mapsto S_t(p_k)^\top p$.[7] Then, expanding the definition of what it means to have no-regret over the course of an arbitrary number K of iterations, one has

$$\frac{\mathrm{Regret}_K}{K} = \frac{1}{K}\sum_{k=1}^{K} S_t(p_k)^\top (z - p_k) = o(1) \qquad \forall z \in \mathcal{Z},$$

implying that the uniform distribution $\mathcal{P}_t$ over the set of iterates $\{p_1, \dots, p_K\}$ produced by the no-regret algorithm converges to an arbitrarily good solution of the expected variational inequality as the number of iterations K grows. By using known regret bounds for online projected gradient ascent, setting the learning rate appropriately, and accounting for the fact that the diameter of the image of S_t is of order tGD, we obtain the following rate (see Appendix C.1 for the derivation).

Proposition 6 *There exists a no-regret-based implementation of Algorithm 1 that achieves* average *multicalibration error* $|\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{p_t \sim \mathcal{P}_t} h(x_t, p_t)^\top (z_t - p_t)| \le \varepsilon$ *in time* $\mathcal{O}(\mathrm{poly}(d, D, G) \cdot \min\{r, \varepsilon^{-6}\} \cdot \varepsilon^{-6})$. *In particular, the runtime is never worse than* $\mathcal{O}(\mathrm{poly}(d, D, G) \cdot \varepsilon^{-12})$.

Significantly faster approaches solving EVIs (and even harder generalizations) at $\log(1/\varepsilon)$ rates have been very recently developed by leveraging a new constructive version of the minimax theorem (Farina and Pipis, 2024; Zhang et al., 2025; Farina, 2026). These methods are based on the ellipsoid method, and can solve EVIs to error ε_t with only order $\log(1/\varepsilon_t)$ evaluations of the operator S_t. We include a high-level, self-contained description of these more advanced algorithms in Appendix C.2. Here, we only state the final result; a proof is deferred to Appendix C.3.

Proposition 7 *There exists an ellipsoid-based implementation of Algorithm 1 that achieves* average *multicalibration error* $|\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{p_t \sim \mathcal{P}_t} h(x_t, p_t)^\top (z_t - p_t)| \le \varepsilon$ *in time* $\tilde{\mathcal{O}}(\mathrm{poly}(d, D, G) \cdot \min\{r, \varepsilon^{-2}\} \cdot \varepsilon^{-2})$. *In particular, the runtime is never worse than* $\tilde{\mathcal{O}}(\mathrm{poly}(d, D, G) \cdot \varepsilon^{-4})$, *and it is of order* $\tilde{\mathcal{O}}(\mathrm{poly}(d, D, G) \cdot \varepsilon^{-2})$ *if the number of features* r *is polynomial in* d, D, *and* G.

Appendix B. Online Outcome Indistinguishable Generative Models

In this section, we illustrate how to use the high-dimensional multicalibration algorithm from Section A to produce online outcome indistinguishable generative models for multiclass, scalar-valued, and vector-valued outcomes. We assume throughout that $\varepsilon_t = GD^2$ in Algorithm 2.

Algorithm Description. As discussed in Section 2, the Defensive Generation algorithm guarantees indistinguishability with respect to functions f that can be written as $f(x, p, y) = h(x, p)^\top s(y)$ where h belongs to vector-valued RKHS $\mathcal{H}$. The procedure takes as input the outcome and prediction sets $\mathcal{Y}$ and $\mathcal{Z} \subset \mathbb{R}^d$, as well as the function $s : \mathcal{Y} \to \mathcal{Z}$ that together define the generative modeling task. It also takes as input the matrix-valued kernel function $\Gamma : (\mathcal{X} \times \mathcal{Z})^2 \to \mathbb{R}^{d \times d}$ that implicitly defines the RKHS $\mathcal{H}$.

7. We use k to indicate the iteration of the no-regret algorithm, to avoid confusion with scalar kernel k used in Section B. An EVI (EVI) must be solved at every time t in Algorithm 1, and we are proposing an iterative algorithm indexed over iterations c to solve such an EVI at that time t.

Algorithm 2 The Defensive Generation Algorithm

Input: matrix-valued kernel Γ, outcome and prediction sets $\mathcal{Y}$ and $\mathcal{Z}$, function $s : \mathcal{Y} \to \mathbb{R}^d$
For $t = 1, \ldots, T$**:**
  Receive features x_t
  Produce distribution $\mathcal{P}_t$ over $p_t \in \mathcal{Z}$ via defensive forecasting (Algorithm 1) to error ε_t
  Output $\mathcal{D}_t$ equal to the pushforward of $\mathcal{P}_t$, where for each $p_t \in \mathsf{supp}(\mathcal{P}_t)$, μ_t solves

$$\mathbb{E}_{\widetilde{y}_t \sim \mu_t}[s(\widetilde{y}_t)] = p_t$$

  Record true outcome, $y_t \in \mathcal{Y}$ and statistic $s(y_t) = z_t$

Defensive Generation is just a simple wrapper around the online multicalibration routine (Algorithm 1). Given x_t, it appeals to the defensive forecasting procedure to find a distribution $\mathcal{P}_t$ over forecasts p_t that is multicalibrated with respect to the outcome $z_t = s(y_t)$. Then, for every p_t it "backfits" a measure μ_t. That is, it solves for a distribution μ_t such that $\mathbb{E}_{\mu_t}[s(\widetilde{y}_t)] = p_t$. The final mixture distribution $\mathcal{D}_t$ is simply the pushforward measure of $\mathcal{P}_t$ under this backfitting operation.

To specialize it to different outcome spaces $\mathcal{Y}$, we need to ensure that we can 1) implement efficient oracle access (e.g., separation) for the set $\mathcal{Z}$ as discussed in Section A and 2) be able to solve for μ_t. The rest of this section shows how to do both of these things for different choices of $\mathcal{Z}$ and $\mathcal{Y}$. Before discussing these adaptations, we state the end-to-end guarantee of the algorithm:

Theorem 8 *Fix an outcome space $\mathcal{Y}$, a function $s : \mathcal{Y} \to \mathcal{Z}$, and let $\mathcal{F}$ be a class of distinguishers f that can be written as $f(x, p, y) = h_f(x, p)^\top s(y)$ for a function h_f in vector-valued RKHS $\mathcal{H}$.*

Given access to a sampling oracle that on input z returns a probability measure μ such that $\mathbb{E}_{\widetilde{y} \sim \mu}[s(\widetilde{y})] = z$, and a separation oracle for the set $\mathcal{Z}$,[8] the Defensive Generation algorithm with $\varepsilon_t = D^2G$, where D and G are as in (3), guarantees online outcome indistinguishability with respect to $\mathcal{F}$. More precisely, given any $f \in \mathcal{F}$,

$$\mathrm{OIGap}_T(f) \leq \|h_f\|_{\mathcal{H}} \sqrt{2D^2GT}.$$

B.1. Generative Models for Multiclass Outcomes

In this subsection, we specialize this main theorem to the case where $\mathcal{Y}$ consists of d labels. In this setting, our results simplify substantially and we can guarantee indistinguishability with respect to distinguishers that examine the entire conditional distribution from which we are sampling. That is, we let $p = \mu$ where μ is a point on the simplex,

$$\mu \in \mathcal{Z} = \Delta([d]) = \{p : p_j \geq 0, \sum_{j=1}^{d} p_j = 1\} \text{ and } \widetilde{y} \sim \mu.$$

In more detail, if $\mathcal{Y} = \{1, \ldots, d\}$, then for any function $f(x, p, y)$, we can write

$$f(x, p, y) = \sum_{j=1}^{d} f(x, p, j) \cdot 1\{y = j\}.$$

8. Given z', a separation oracle for the set $\mathcal{Z}$ returns true if $z' \in \mathcal{Z}$. If $z' \notin \mathcal{Z}$, it returns a hyperplane (w, b) such that $w^\top z' > b$ but $w^\top z < b$ for all $z \in \mathcal{Z}$

Therefore, any distinguisher f can be expressed as a linear function of its values at d points, $f(x,p,y) = h_f(x,p)^\top s(y)$ where $h_f(x,p) = (f(x,p,1),\dots,f(x,p,d))$ and $s(y) = (1\{y = 1\},\dots,1\{y = d\}) \in \mathcal{Z}$ is a one-hot encoded version of the outcomes y.

Moreover, for this multiclass case, there is a simple formula we can use to construct function spaces that contain these functions h_f. In particular, assuming that there is a *scalar*-valued RKHS that contains each of the scalar-valued functions $f(x,p,i)$ for $i \in [d]$, we can always construct a *vector*-valued RKHS that contains the function $h_f(x,p)$ from above.

Fact 1 (Alvarez et al. (2012)) *Let $k((x,p),(x',p')) \in \mathbb{R}$ be a scalar-valued kernel with RKHS $\mathcal{H}_{\text{scalar}} \subset \{\mathcal{X} \times \mathcal{Z} \to \mathbb{R}\}$. Then, $\Gamma((x,p),(x',p')) = I_d \cdot k((x,p),(x',p')) \in \mathbb{R}^{d\times d}$ is a matrix-valued kernel whose corresponding vector-valued RKHS $\mathcal{H}_d$ consists of all functions of the form:*

$$h(x,p) = (h_1(x,p),\dots,h_d(x,p))^\top \text{ where } h_j(x,p) \in \mathcal{H}_{\text{scalar}} \text{ for all } j \in [d].$$

Furthermore, for any function $h \in \mathcal{H}_d$ its RKHS norm is equal to $\|h\|^2_{\mathcal{H}_d} = \sum_{j=1}^d \|h_j\|^2_{\mathcal{H}_{\text{scalar}}}$.

Here, there is no need to implement a sampling oracle, since the algorithm directly solves for the distribution $p = \mu$ as part of the multicalibration subroutine. Furthermore, it is simple to check whether any point is in the simplex $\mathcal{Z} = \{p : p_i \geq 0, \sum_{i=1}^d p_i = 1\}$, and one can easily implement a separation oracle. Tying these facts together with Theorem 8, we get the following corollary:[9]

Corollary 9 *Set $\mathcal{Y} = [d]$ and define $\mathcal{Z}$ to be the simplex in $\mathbb{R}^d$, $\mathcal{Z} = \{p : p_j \geq 0, \sum_{j=1}^d p_j = 1\}$. Let k be a kernel such that $\sup_{x,p} |k((x,p),(x,p))| \leq G$ with an associated scalar-valued RKHS $\mathcal{H}_{\text{scalar}}$, such that $f(x,p,j) = h_j(x,p) \in \mathcal{H}_{\text{scalar}}$ for all $j \in [d]$.*

Then, Defensive Generation with kernel $\Gamma = I_d \cdot k$ produces distributions μ_t over outcomes $\mathcal{Y}$ that are online outcome indistinguishable with respect to the set of functions $f(x,\mu,y)$ where $h_j(x,\mu) = f(x,\mu,j)$ is in $\mathcal{H}_{\text{scalar}}$ for every $j \in [d]$. In particular, for every such f,

$$\text{OIGap}_T(f) \leq 4\left(\sum_{j=1}^d \|h_j\|_{\mathcal{H}_{\text{scalar}}}\right)\sqrt{TG}.$$

We can further specialize this statement to guarantee indistinguishability with respect to common function classes as per this example:

Example 1 *Let $\mathcal{Y} = [d]$, $\mathcal{X} = \{x : \|x\|_2 \leq 1\} \subset \mathbb{R}^r$, and set $k((x,p),(x',p')) = \langle x,x'\rangle + \langle p,p'\rangle$. Running the Defensive Generation with $\Gamma((x,p),(x',p')) = I_d \cdot k((x,p),(x',p'))$ guarantees online outcome indistinguishability with respect to the set of functions,*

$$\mathcal{F} = \{f(x,p,y) : f(x,p,j) = \langle x,\theta_j\rangle + \langle p,\theta'_j\rangle, \quad (\theta_i,\theta'_i) \in \mathbb{R}^r \times \mathbb{R}^d \text{ for all } i \in [d]\}.$$

In particular, if we restrict $\|\theta_j\|_2 + \|\theta'_j\|_2 \leq B$ for every $j \in [d]$, then we get that $\text{OIGap}_T(f)$ is at most $4dB\sqrt{T}$ for any f. Furthermore, at time t the algorithm runs in time $\mathcal{O}(\text{poly}(d,B,r)\log(t))$.

9. We use μ_t and p_t interchangeably in Theorem 9 since they are the same for this multiclass setting.

B.2. Generative Models for Scalar-Valued Outcomes

As a second application, we show how Defensive Generation can be used to produce distributions μ_t over scalar outcomes y_t that are outcome indistinguishable from the perspective of any f that examines higher-order moments of the distribution, up to some fixed degree $2d$. In symbols, we focus on the setting in which $\mathcal{Y} = [-1, 1]$ and $s(y) = (1, y, y^2, \ldots, y^{2d})$ and $\mathcal{Z}$ is the (convex) set of moments:

$$\mathcal{Z} \coloneqq \left\{ \left(\int_{[-1,1]} y^0 \, \mathrm{d}\mu, \ldots, \int_{[-1,1]} y^{2d} \, \mathrm{d}\mu \right) : \mu \text{ is a Borel probability measure on } [-1, 1] \right\}. \quad (5)$$

To extend our results to this setting, we leverage techniques from semi-algebraic optimization. As we discuss in Section D, $\mathcal{Z}$ admits an efficient separation oracle, thus satisfying the preconditions of our construction in Section A for efficiently producing defensive forecasts $p_t \in \mathcal{Z}$. Furthermore, once a moment vector $p_t \in \mathcal{Z}$ is selected, it is possible to efficiently construct a discrete distribution μ_t on $[-1, 1]$ whose moments match those specified by p_t. By sampling a $\widetilde{y}_t \sim \mu_t$, we can complete the roundtrip and yield predictions that are indistinguishable with respect to tests that only examine a fixed number of moments. Tying these ideas together, we get the following corollary:

Corollary 10 *Set $\mathcal{Y} = [-1, 1]$ and define $\mathcal{Z} \subset \mathbb{R}^{2d+1}$ to be the set of moments as in Equation (5). Let Γ be a matrix-valued kernel with associated vector-valued RKHS $\mathcal{H} \subseteq \{\mathcal{X} \times \mathcal{Z} \to \mathbb{R}^{2d+1}\}$ and let $\mathcal{F}$ be the set of functions,*

$$\mathcal{F} \coloneqq \{f(x, p, y) : f(x, p, y) = h(x, p)^\top s(y) \text{ where } s(y) = (1, y, \ldots, y^{2d})\}.$$

The Defensive Generation algorithm with the sampling and separation oracles from Section D produces distributions μ_t and statistics $p_t^\top = (1, \mathbb{E}_{\mu_t}[\widetilde{y}_t], \ldots, \mathbb{E}_{\mu_t}[\widetilde{y}_t^{2d}])$ such that for every $f \in \mathcal{F}$,

$$\mathrm{OIGap}_T(f) \leq \|h\|_{\mathcal{H}} \sqrt{\sum_{t=1}^{T} \mathbb{E}_{p_t}(p_t - s(y_t))^\top \Gamma((x_t, p_t), (x_t, p_t))(p_t - s(y_t)) + GD^2T}.$$

Using this we can generate distributions over scalar outcomes that fool all low degree tests. That is, we can use the Defensive Generation algorithm to produce distributions μ_t over scalar outcomes in $[-1, 1]$ such that all low degree moments are conditionally valid over all subsets of the Boolean hypercube computable by low-depth decision trees.

Example 2 *Let $\mathcal{X} = \{\pm 1\}^n$ be the Boolean hypercube and consider the polynomial kernel,*

$$k_{\mathrm{poly}}(x, x') = (1 + x^\top x')^r.$$

It is a well-known result that the RKHS $\mathcal{H}_{\mathrm{poly}}$ for this kernel contains all Boolean functions $c : \{\pm 1\}^n \to \{0, 1\}$ computable by decision trees of depth at most r.[10]

10. See Proposition 4.12 in Dwork et al. (2025) for a proof. As per O'Donnell (2014), a decision tree is a representation of a boolean function as a rooted binary tree in which the internal nodes are labeled by coordinates $[i] \in [n]$, the outgoing edges are labeled by -1 and 1 and the leaves have real numbers corresponding to the outputs.

Moreover, by Fact 1, the matrix-valued kernel $\Gamma((x,p),(x',p')) = I_{2d+1} \cdot k_{\text{poly}}(x,x')$ *contains all functions,* $h(x) = (c_0(x),\dots,c_{2d}(x))$ *where each* c_i *is a Boolean function in* $\mathcal{H}_{\text{poly}}$*. Consequently, if we consider the space of functions* $\mathcal{F}$ *that can be written as,*

$$f(x,p,y) = h(x)^\top s(y) = \sum_{j=1}^{2d+1} c_j(x) y^j, \tag{6}$$

this space contains all tests of the form $f(x,p,y) = 1\{c(x)=1\}\cdot y^j$*. Therefore, the Defensive Generation algorithm guarantees that for any* $j = 1,\dots,2d+1$ *and all low-degree boolean functions* c*,*

$$\lim_{T\to\infty} \left| \frac{1}{T}\sum_{t=1}^{T} 1\{c(x_t)=1\} y_t^j - \frac{1}{T}\sum_{t=1}^{T} 1\{c(x_t)=1\} \mathbb{E}_{\widetilde{y}_t\sim\mu_t,\mu_t\sim\mathcal{D}_t}[\widetilde{y}_t^j] \right| = 0. \tag{7}$$

That is, the first $2d$ *moments of the conditional distributions we produce (online) match those of the revealed outcomes. Furthermore, this is true conditionally over all the subsequences* $\{x_t : c(x_t) = 1, c$ *is a decision tree of depth* $r\}$*. More formally, for all functions* f *of the form in Equation (6),*

$$\text{OIGap}_T(f) \le \mathcal{O}(\text{poly}(d)\cdot n^r \cdot \sqrt{T}).$$

While Dwork et al. (2025) achieved the guarantee in Equation (7) for the case where $j=1$, our algorithm produces distributions which (simultaneously) satisfy the guarantee for all $j = 1,\dots,2d$. This result is also closely related to the guarantees in Gupta et al. (2022). Their results hold for higher order moments, but only with respect to a finite class of distinguishers.

We note that one interesting consequence of producing distributions that have these high order guarantees of validity is that one can use them to derive high-probability prediction intervals via Chebyshev's inequality. We refer the reader to Gupta et al. (2022) for further details.

B.3. Generative Models for High-Dimensional Outcomes

As a final application, we consider guaranteeing outcome indistinguishability when outcomes are high-dimensional, $\mathcal{Y} \subset \mathbb{R}^d$.

Indistinguishability of conditional expectations To start, we show that we can always guarantee online outcome indistinguishability when the distinguishers f are linear functions of the first moment of y, $f(x,p,y) = h(x,p)^\top y$, and $s(y) = y$. That is, they guarantee that the distributions produced by the learner have first moments that are conditionally correct as measured by the functions h.

For this setting, we let $\mathcal{Z} = \mathcal{Y} \subset \mathbb{R}^d$ which is assumed to be any compact, convex set with diameter bounded by D with an efficient separation oracle (e.g. $\mathcal{Y} = [-1,1]^d$). Since $s(y)$ is just the identity function, the problem of finding a distribution μ_t such that $\mathbb{E}_{\widetilde{y}\sim\mu_t}[s(\widetilde{y}_t)] = \mathbb{E}_{\widetilde{y}_t\sim\mu_t}[\widetilde{y}_t] = p_t$ is easy, just return a point mass distribution at p_t. Here, we have the following corollary:

Corollary 11 *Let* Γ *be a matrix valued kernel with operator norm uniformly bounded by* G *and with vector-valued RKHS* $\mathcal{H} \subseteq \{\mathcal{X}\times\mathcal{Z}\to\mathbb{R}^d\}$ *where* $\mathcal{Z}$ *has diameter bounded by* D*. Then, the Defensive Generation algorithm with kernel* Γ *guarantees online OI with respect to the set of functions* $f(x,p,y) = h(x,p)^\top y$ *where* $h(x,p) \in \mathcal{H}$*. For any* $f \in \mathcal{F}$*,* $\text{OIGap}_T(f) \le \|h\|_{\mathcal{H}}\sqrt{2TD^2G}$.

As a concrete instantiation of this result, we can guarantee indistinguishability with respect to the set of functions: $\mathcal{F} = \{f(x,p,y) = h(x,p)^\top y \text{ where } h(x,p) = \Phi(x,p)\theta \text{ for } \theta \in \mathbb{R}^r\}$.

Here, Φ is a finite-dimensional feature map $\Phi(x,p)^\top \in \mathbb{R}^{d\times r}$ and $h(x,p)$ is any function that lies in the span of these features. For instance, if we let,

$$\Phi(x,p)^\top = \big[g_1(x,p) \mid \cdots \mid g_r(x,p)\big] \in \mathbb{R}^{d\times r} \tag{8}$$

where $\{g_1,\ldots,g_r\}$ is any arbitrary collection of r functions then we get that $\mathcal{H} = \operatorname{span}(\{g_1,\ldots,g_r\})$. This is just one choice of an explicit feature map, but one could of course consider others.

Example 3 *Let $\mathcal{Y} = \mathcal{Z}$ be a compact subset of $\mathbb{R}^d$ with diameter D and let $\Phi : \mathcal{X}\times\mathcal{Z} \to \mathbb{R}^{r\times d}$ be a feature map. The Defensive Generation algorithm with matrix valued kernel $\Gamma((x,p),(x',p')) = \Phi(x,p)^\top\Phi(x',p')$ guarantees online outcome indistinguishability with respect to the set of functions $f(x,p,y) = h(x,p)^\top y$ where $h(x,p) = \Phi(x,p)\theta$. In particular, for every such h,*

$$\mathrm{OIGap}_T(f) \le \|\theta\|_2\sqrt{2D^2T\cdot\sup_{x,p}\|\Phi(x,p)\|_{\mathrm{op}}^2}.$$

Therefore, if we set Φ as in (8) for functions $g_1,\ldots,g_r$, where $\max_{1\le j\le r}\|g_j(x,p)\|_2 \le B$, then for any $f(x,p,y) = g_j(x,p)^\top y$, we have $\mathrm{OIGap}_T(f) \le \sqrt{2rB^2D^2T}$.

Indistinguishability with respect to higher-order moments In general, the defensive generation approach described so far can guarantee efficient online outcome indistinguishability with respect to d-order moments, when the set of such moments admits an efficient separation oracle. This condition is equivalent to the existence of efficient algorithms for the minimization of polynomials of degree up to d on $\mathcal{Y}$.

As an illustration, we show that on the high-dimensional ball $\mathcal{Y} = \{y \in \mathbb{R}^d : \|y\|_2 \le 1\}$, we can guarantee indistinguishability with respect to functions f that examine the first two moments of the distributions μ_t,

$$f(x,p,y) = h(x,p)^\top s(y), \text{ and } s(y) = (y_1,\ldots,y_d\ ,\ y_1^2\ ,y_1y_2\ ,\ldots,y_iy_j) \text{ for } i,j\in[d], i\le j. \tag{9}$$

In other words, they guarantee that the distributions μ_t produced online have high-dimensional means, $\mathbb{E}_{\mu_t}[\widetilde{y}_t]$, and covariances, $\mathbb{E}_{\mu_t}[\widetilde{y}_t\widetilde{y}_t^\top] \in \mathbb{R}^{d\times d}$, that are conditionally correct as measured by h. In this setting, the set of moments $\mathcal{Z}$ is

$$\mathcal{Z} = \{(v,Q) : \exists \text{ a probability measure } \mu \text{ over } \|y\|_2 \le 1 \text{ such that } \mathbb{E}_\mu[y] = v, \mathbb{E}_\mu[yy^\top] = Q\} \tag{10}$$

Since quadratic functions on the unit ball can always be optimized efficiently (by performing a singular value decomposition), a separation oracle for $\mathcal{Z}$ can always be constructed efficiently. In fact, in this case we do not even need to resort to complicated separation oracle constructions: as it turns out, the set $\mathcal{Z}$ admits a simple characterization. Indeed, an application of the S-Lemma (Yakubovich, 1971) lets us write

$$\mathcal{Z} = \{(v,Q) : Q \succeq 0, Q \succeq vv^\top, \mathrm{Tr}[Q] \le 1\}, \tag{11}$$

where $\mathrm{Tr}[Q]$ denotes the trace of Q and $A \succeq B$ denotes that $A - B$ is positive semidefinite. (A more constructive proof of (11) via some intermediate results that will be used later in this section

Algorithm 3 Backfitting a probability measure μ.

// We use $\delta(y)$ to denote a point mass on a point y.

Input: PSD matrix Q and vector $v \in \mathbb{R}^d$ such that $Q \succeq vv^\top$

Compute the SVD of $Q - vv^\top = \sum_{i=1}^d \sigma_i u_i u_i^\top$

For $i = 1 \dots d$,

  Solve for the 2 real-valued roots $t_i^+ > 0$ and $t_i^- < 0$ that satisfy $\|v + t \cdot u_i\| = 1$

  Define

$$y_i^+ = v + t_i^+ u_i,\ y_i^- = v + t_i^- u_i \quad \text{and} \quad \lambda_i^+ = \frac{\sigma_i}{t_i^+(t_i^+ - t_i^-)}, \lambda_i^- = \frac{\sigma_i}{t_i^-(t_i^- - t_i^+)}$$

Return $\mu = \lambda_0 \delta(v) + \sum_{i=1}^d [\lambda_i^+ \delta(y_i^+) + \lambda_i^- \delta(y_i^-)]$ where $\lambda_0 = 1 - \sum_{i=1}^d (\lambda_i^+ + \lambda_i^-)$

is available in Appendix E.) From this rewriting, it is straightforward to efficiently check whether any given proposed moments (v', Q') belong to $\mathcal{Z}$, or return a separating hyperplane (violated constraint) otherwise.

Moreover, given $(v, Q) \in \mathcal{Z}$, we can efficiently find an atomic probability measure μ over the unit ball such that $\mathbb{E}_{y\sim\mu}[y] = v$ and $\mathbb{E}_\mu[yy^\top] = Q$. It involves nothing more complicated than taking the SVD of a matrix. We present the procedure in Algorithm 3 and the proof of its correctness in Appendix E.

Given that we can implement a separation oracle and a way to solve for a probability measure μ whose moments match a target vector of moments in $\mathcal{Z}$, we have the following corollary:

Corollary 12 *Let Γ be a matrix-valued kernel with operator norm uniformly bounded by G and with vector-valued RKHS $\mathcal{H} \subseteq \{\mathcal{X} \times \mathcal{Z} \to \mathbb{R}^{d+d(d+1)/2}\}$ where $\mathcal{Z}$ is as in (10) and $s(y) \in \mathbb{R}^{d+d(d+1)/2}$. Then, $\mathcal{Z}$ has diameter at most $\sqrt{8}$ and the Defensive Generation algorithm with kernel Γ guarantees online OI with respect to the set of functions $f(x, p, y)$ defined in (9). For any $f \in \mathcal{F}$,*

$$\mathrm{OIGap}_T(f) \le 4\|h\|_{\mathcal{H}} \sqrt{TG}.$$

As a concrete instantiation of this result, we can guarantee indistinguishability with respect to the set of functions in Equation (9) where $s(y) \in \mathbb{R}^m$ for $m = d + d(d+1)/2$ is a vector containing all the variables in the first and second moments of y (e.g. y_i, y_i^2 and also the mixed pairs $y_i y_j$ for $y = (y_1, \dots y_d)$). In this example, Φ is a finite-dimensional feature map $\Phi(x, p) \in \mathbb{R}^{m\times r}$ and $h(x, p)$ is any function that lies in the span of these features. For instance, if we let,

$$\Phi(x, p)^\top = \big[g_1(x, p) \mid \cdots \mid g_r(x, p)\big] \in \mathbb{R}^{m\times r}$$

where $g_1, \dots, g_r$ is any arbitrary set of r functions then we get that $\mathcal{H} = \mathrm{span}(\{g_1, \dots, g_r\})$. This is just one choice of an explicit feature map, but one could of course consider others.

Example 4 *Let $\mathcal{Z}$ (see Equation (10)) be the set of first and second moments of probability distributions over the unit ball and let $\Phi : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^{r\times m}$ be a feature map.*

The Defensive Generation algorithm with kernel $\Gamma((x, p), (x', p')) = \Phi(x, p)^\top \Phi(x', p')$ guarantees online outcome indistinguishability with respect to the set of functions $f(x, p, y) = h(x, p)^\top s(y)$

where $h(x,p) = \Phi(x,p)\theta$. *In particular, for every such* h,

$$\mathrm{OIGap}_T(f) \le 4\|\theta\|_2 \sqrt{T \cdot \sup_{x,p} \|\Phi(x,p)\|_{\mathrm{op}}^2}.$$

Therefore, if we set Φ *as in* (8) *for functions* $g_1, \dots, g_r$, *where* $\max_{1\le j\le r} \|g_j(x,p)\|_2 \le B$, *then for any* $f(x,p,y) = g_j(x,p)^\top s(y)$,

$$\mathrm{OIGap}_T(f) \le 4\sqrt{rB^2T}.$$

We note that while we presented the results in this last section for the case where $\mathcal{Y}$ is the unit ball, they apply more generally to the case where $\mathcal{Y}$ is any ellipsoid

$$\mathcal{Y} = \{x : (x-b)^\top A(x-b) \le 1, A \succ 0, b \in \mathbb{R}^d\},$$

by simply changing basis $x \mapsto A^{-1/2}x + b$.

Appendix C. Expected Variational Inequalities

C.1. Proof of Proposition 6

From the definition of S_t, the norm of the utility gradient is at most $G' := \max_z \|S_t(z)\|_2 \le tGD$. Hence, by setting the learning rate $\eta = D/(G\sqrt{T})$, and using the known analysis of the online projected gradient ascent algorithm (Zinkevich, 2003), we can write

$$\frac{\mathrm{Regret}_K}{K} \le 2G'D\frac{\sqrt{K}}{K} \le 2D^2G\frac{t}{\sqrt{K}}.$$

This shows that it is enough to set $K = t^2$ to obtain an error of $2D^2G$ in the solution of (EVI) and thus, by Theorem 5, a multicalibration error bounded by $2\|h\|_{\mathcal{H}}\sqrt{D^2GT}$ against any $h \in \mathcal{H}$.

To achieve a given average multicalibration error ε, it is then necessary to run Algorithm 1 for $T = D^2G/\varepsilon^2$. The only step left to complete the proof is then to estimate the complexity of each iteration of the no-regret-based EVI solution algorithm.

Each iteration involves evaluating S_t at the current p_k, taking a gradient step, and projecting onto $\mathcal{Z}$. The latter two operations require $\mathrm{poly}(d)$ time. Thus, at every time t we need to perform an amount of work that is of order $t^2(\mathrm{poly}(d) + \mathsf{Eval}(S_t))$, where $\mathsf{Eval}(S_t)$ denotes the cost of evaluating S_t. We distinguish two possibilities for evaluating S_t.

- In general, S_t is defined as a sum of t expectations. Each expectation is over the distribution $\mathcal{P}_\tau$ that solved (EVI) at time τ. As discussed above, the support at time τ is $K_\tau = \tau^2$. So, given our assumption that each evaluation of Γ takes $\mathrm{poly}(d)$ time, we are left with a $\mathcal{O}(t^3\mathrm{poly}(d))$ time bound for each evaluation of S_t.

  Putting all the pieces together, we are left with a $\mathcal{O}(t^5\mathrm{poly}(d))$ time per iteration of Algorithm 1, leading to a $\mathcal{O}(T^6\mathrm{poly}(d))$ runtime to complete T iterations. Plugging $T = D^2G/\varepsilon^2$ yields a bound of $\mathrm{poly}(d, D, G) \cdot \varepsilon^{-12}$.

- When $r = \mathcal{O}(\mathrm{poly}(d, D, G))$, the cost of evaluating $S_t(p)$ can be amortized, by accumulating over time the vector quantity

$$m_t \coloneqq \sum_{\tau=1}^{t-1} \mathbb{E}_{p_\tau \sim \mathcal{P}_\tau} \Phi(x_\tau, p_\tau)(z_\tau - p_\tau).$$

  Indeed, by definition of S_t, we have $S_t(p) = \Phi(x_t, p)^\top m_t$, leading to a cost of $\mathcal{O}(rd)$ for each evaluation of S_t. To solve the EVI, we then require $t^2\mathrm{poly}(d)$ time.

  After every iteration t of Algorithm 1, the quantity m_t can be updated by summing the new expectation arising from $\mathcal{P}_t$. Since the distribution has support $K = t^2$, such an update has a $\mathcal{O}(t^2\mathrm{poly}(d))$ cost.

  In total, each iteration of Algorithm 1 requires $t^2\mathrm{poly}(d, r)$ time, for a total of $T^3\mathrm{poly}(d)r$ time. Plugging $T = D^2G/\varepsilon^2$ yields a bound of $\mathcal{O}(\mathrm{poly}(d, D, G) \cdot r \cdot \varepsilon^{-6})$ time to reach ε average multicalibration error.

Taking the minimum between the two cases yields the statement.

C.2. Intuition Behind the Construction of Zhang et al. (2025)

As mentioned in Section A, significantly faster approaches for solving EVIs (and even harder generalizations) at $\log(1/\varepsilon)$ rates have been very recently developed by leveraging a new constructive version of the minimax theorem (Farina and Pipis, 2024; Zhang et al., 2025; Farina, 2026).

We now give a high-level intuition for these methods; for more details, we refer the reader to the paper of Zhang et al. (2025). These faster algorithms can be understood as operating in two phases:

- First, they produce an extremely sparse set $\mathcal{S} = \{p_1, \dots, p_K\} \subset \mathcal{Z}$ of size

$$K = \mathcal{O}(\mathrm{poly}(d) \log(G, D, 1/\varepsilon)).$$

- Then, they search for a distribution over the K points in $\mathcal{S}$ that guarantees value ε for all $z \in \mathcal{Z}$. This discrete distribution $(\lambda_1, \dots, \lambda_K)$ solving the EVI problem can be computed by solving the convex optimization problem

$$\arg\min_{\lambda \in \Delta^K} \max_{z \in \mathcal{Z}} \sum_{k=1}^{K} \lambda_k S(p_k)^\top (z - p_k).$$

  over Δ^K. This can be solved using standard convex optimization techniques in time polynomial in K, the diameter of $\mathcal{Z}$ and S (Grötschel et al., 1993).

Conceptually, to produce the sparse support, the algorithm makes use of the ellipsoid algorithm to certify the emptiness of the set $\Omega \coloneqq \{z \in \mathcal{Z} : S(p)^\top (z - p) > 0 \ \ \forall p \in \mathcal{Z}\}$. The emptiness of Ω is direct, since for any $z \in \mathcal{Z}$, the constraint indexed by $p = z$ is trivially violated. By running the ellipsoid over Ω using $p = z$ as the separation oracle, the ellipsoid method is therefore able to produce a trace of violated constraints indexed by the ellipsoid centers $p_1 = z_1, p_2 = z_2, \dots, p_K = z_K$ that *sparsely certifies* the emptiness of Ω. The Farkas lemma then implies that a distribution over the constraints must be an EVI solution, justifying the soundness of the second step.

C.3. Proof of Proposition 7

The ellipsoid-based algorithm for EVIs is able to solve the EVI problem with only

$$\mathcal{O}(\text{poly}(d)\log(D, G, \log 1/\varepsilon))$$

calls to the operator S_t, producing a distribution with support size bounded by the same quantity. Following the discussion in Section C.1, we distinguish two cases for the cost of evaluating S_t.

- In general, S_t is defined as a sum of t expectations. Each expectation is over the distribution $\mathcal{P}_\tau$ that solved (EVI) at time τ. Since the support at time τ is $K_\tau = \tilde{\mathcal{O}}(\log(t/\varepsilon))$, given our assumption that each evaluation of Γ takes $\text{poly}(d)$ time, we are left with a $\mathcal{O}(t \cdot \text{poly}(d))$ time bound for each evaluation of S_t.

  Putting all the pieces together, we are left with a $\mathcal{O}(t\log^2(t/\varepsilon))$ time per iteration of Algorithm 1, leading to a $\tilde{\mathcal{O}}(T^2\text{poly}(d))$ runtime to complete T iterations. Plugging $T = D^2G/\varepsilon^2$ yields a bound of $\tilde{\mathcal{O}}(\text{poly}(d, D, G) \cdot \varepsilon^{-4})$.

- When $r = \mathcal{O}(\text{poly}(d, D, G))$, the cost of evaluating $S_t(p)$ can be amortized as described in Section C.1, leading to a $\mathcal{O}(\text{poly}(d)r)$ evaluation time for S_t. Hence, we can solve the EVI in time $\tilde{\mathcal{O}}(\text{poly}(d)r)$

  After every iteration t of Algorithm 1, the quantity S_t can be updated by summing the new expectation arising from $\mathcal{P}_t$. Since the distribution has support $K = \tilde{\mathcal{O}}(1)$, such an update has a $\tilde{\mathcal{O}}(\text{poly}(d))$ cost.

  In total, each iteration of Algorithm 1 requires $\tilde{\mathcal{O}}(\text{poly}(d)r)$ time, for a total of $\mathcal{O}(T\text{poly}(d)r)$ time. Plugging $T = D^2G/\varepsilon^2$ yields a bound of $\mathcal{O}(\text{poly}(d, D, G) \cdot r \cdot \varepsilon^{-2})$ time to reach ε average multicalibration error.

Taking the minimum between the two cases yields the statement.

Appendix D. More details on the Cone of the first d Moments in $\mathbb{R}$

In the interest of keeping the exposition as self-contained as possible, in this section we sketch classical results regarding the algebraic and computational properties of the cone of moments. For more details, we refer the reader to Chapters 3.3 and 3.4 of the book on moment and polynomial optimization by Nie (2023), or the paper of Parrilo (2006).

To lighten notation, in this section we use the notation $\langle \cdot, \cdot \rangle$ to denote the standard dot product in $\mathbb{R}^{2d+1}$.

Efficient Separation Oracle for the Moment Set An efficient separation oracle for $\mathcal{Z}$ can be constructed by leveraging a duality result between moments and nonnegative polynomials, and invoking classical Positivstellensatz results to characterize the latter via the positive semidefinite (PSD) cone. To start, it is well understood that the dual set $\mathcal{Z}^*$ of the $\mathcal{Z}$ from (5), defined as

$$\mathcal{Z}^* \coloneqq \{a = (a_0, a_1, \ldots, a_{2d}) \in \mathbb{R}^{2d+1} : \langle a, z \rangle \geq 0 \quad \forall z \in \mathcal{Z}\}$$

corresponds to the set of polynomials that are nonnegative everywhere on the interval $[-1, 1]$:

$$\mathcal{Z}^* = \left\{ (a_0, a_1, \ldots, a_{2d}) \in \mathbb{R}^{2d+1} : \sum_{i=0}^{2d} a_i y^i \geq 0 \quad \forall y \in [-1, 1] \right\}.$$

To see this, observe that for any $a \in \mathbb{R}^{2d+1}$,

$$\langle a, z\rangle \geq 0 \ \forall z \quad \iff \quad \sum_{i=0}^{2d} \left(a_i \int_{[-1,1]} y^i \, \mathrm{d}\mu \right) \geq 0 \ \forall \mu \quad \iff \quad \int_{[-1,1]} \left(\sum_{i=0}^{2d} a_i y^i \right) \mathrm{d}\mu \geq 0 \ \forall \mu.$$

Since, in particular, we are free to take μ to be a Dirac delta centered at any $y \in [-1,1]$, it is immediate to see that

$$\int_{[-1,1]} \left(\sum_{i=0}^{2d} a_i y^i \right) \mathrm{d}\mu \geq 0 \ \ \forall \mu \quad \iff \quad \sum_{i=0}^{2d} a_i y^i \geq 0 \ \ \forall y \in [-1,1],$$

completing the proof. Both $\mathcal{Z}$ and $\mathcal{Z}^*$ are convex and closed sets.

The dual set $\mathcal{Z}^*$ plays a key role in constructing an efficient separation oracle for $\mathcal{Z}$. Indeed, one can prove that $\mathcal{Z}^*$ defines a *cutting plane characterization* of $\mathcal{Z}$, in the sense that

$$\mathcal{Z} = \{z \in \mathbb{R}^{2d+1} : z_0 = 1 \wedge \langle z, a\rangle \geq 0 \quad \forall a \in \mathcal{Z}^*\}. \tag{12}$$

The proof of the above result is a standard application of separation for closed convex sets. Equation (12) implies that a separation oracle for $\mathcal{Z}$ can be constructed directly from a linear optimization oracle for $\mathcal{Z}^*$. Specifically, to check whether a given point w belongs to $\mathcal{Z}$, one can do the following:

- if $w_0 \neq 1$, then clearly $w \notin \mathcal{Z}$, and the vector $(1, 0, \ldots, 0)$ provides a separating direction; else
- We solve the convex optimization problem $\min_{a \in \mathcal{Z}^*} \langle w, a\rangle$. If the optimal value of the problem is non-negative, then $w \in \mathcal{Z}$; else, the minimizer a^* provides a separating direction, as $\langle a^*, w\rangle < 0$ by assumption, and yet $\langle a^*, z\rangle \geq 0$ for all $z \in \mathcal{Z}$ by Equation (12).

To complete the construction, it remains to show that the cone $\mathcal{Z}^*$ of nonnegative polynomials on $[-1,1]$ admits an efficient linear optimization oracle. As mentioned above, this follows from important results in semi-algebraic optimization. In particular, it is a celebrated result that in dimension one, polynomial nonnegativity is intimately connected with the notion of sum-of-squares (SOS) polynomials. A polynomial q of degree at most $2d$ is said to be SOS, denoted $q \in \Sigma_{2d}$, if it can be written in the form

$$q(y) = \begin{pmatrix} 1 \\ y \\ \vdots \\ y^d \end{pmatrix}^\top Q \begin{pmatrix} 1 \\ y \\ \vdots \\ y^d \end{pmatrix}, \text{ where } Q = \begin{pmatrix} q_{00} & q_{01} & \cdots & q_{0d} \\ q_{01} & q_{11} & \cdots & q_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ q_{0d} & q_{1d} & \cdots & q_{dd} \end{pmatrix} \succeq 0_{d+1} \text{ is a PSD matrix.} \tag{13}$$

An application of the Positivstellensatz result for univariate polynomials states that a polynomial $p(y) = a_0 + \cdots + a_{2d} y^{2d}$ is nonnegative on $[-1,1]$ if and only if it can be decomposed in the form

$$p(y) = r(y) + (1 - y^2) s(y), \quad \text{for } r \in \Sigma_{2d}, \ \ s \in \Sigma_{2d-2}.$$

The coefficients of the polynomial on the right-hand side depend linearly on the PSD matrices R and S underlying r and s. Letting $H : \mathbb{R}^{(d+1)\times(d+1)} \times \mathbb{R}^{d\times d} \to \mathbb{R}^{2d+1}$ denote the mapping from (R, S) to the coefficients of the right-hand side polynomial, we therefore conclude that

$$\mathcal{Z}^* = \left\{ a \in \mathbb{R}^{2d+1} : a = H(R,S), \ \ R \succeq 0_{d+1}, \ \ S \succeq 0_d \right\}.$$

Hence $\mathcal{Z}^*$ is a convex conic domain defined as the intersection between the linear constraint $a = H(R, S)$, and the product of two PSD cones. Minimization of a linear objective over $\mathcal{Z}^*$, as used by the separation oracle for $\mathcal{Z}$, can therefore be solved using standard semidefinite programming techniques.

Solving for μ_t. Having discussed how to implement a separation oracle for the moment cone, we now describe how given a target vector of moments $z_t \in \mathcal{Z}$, one can produce a discrete distribution μ over $[-1, 1]$ with at most $2d + 1$ atoms in the support, whose moments match z,

$$\mathbb{E}_{\mu_t}[s(\widetilde{y}_t)]^\top = (1, \mathbb{E}_{\mu_t}[\widetilde{y}_t], \ldots, \mathbb{E}_{\mu_t}[\widetilde{y}_t^{2d}]) = z_t.$$

To do so, we leverage once again the duality between the moment set $\mathcal{Z}$ and the set $\mathcal{Z}^*$, as captured by (12). In light of the connection, the statement that $z \in \mathcal{Z}$ is witnessed by the fact that $z_0 = 1$ and $\langle z, a\rangle \geq 0$ for all $a \in \mathcal{Z}^*$. In particular, let

$$a^* \in \operatorname*{arg\,min}_{a \in \mathcal{Z}^*:\ 1^\top a \geq 1} \langle a, z\rangle,$$

which can be computed efficiently by semidefinite programming using the ideas we discussed.[11] Since $(a^*)^\top z \geq 0$, we can assume without loss of generality that the optimal solution satisfies $1^\top a^* = 1$. Since a^* belongs in $\mathcal{Z}^*$, it induces a nonzero polynomial

$$p^*(y) \coloneqq a_0^* + a_1^* y + \cdots + a_{2d}^* y^{2d}, \qquad p^*(y) \geq 0 \quad \forall y \in [-1, 1].$$

Let $\{\gamma_1, \ldots, \gamma_m\}$ be the distinct zeros of p^* (of course, $0 \leq m \leq 2d$, since p^* is not identically zero).

We now show that the set $\{\gamma_0 \coloneqq 1, \gamma_1, \ldots, \gamma_m\}$ defines a valid discrete support for a distribution μ that matches the moments z. By the first-order optimality conditions, it must then be the case that the gradient of the objective is in the normal cone of the feasible set at a^*. Since all constraints are binding by construction, this condition is exactly the existence of convex combination coefficients $\lambda_0, \ldots, \lambda_m$ such that

$$z = \sum_{j=0}^{m} \lambda_j \begin{pmatrix} 1 \\ \gamma_j^1 \\ \vdots \\ \gamma_j^{2d} \end{pmatrix}. \tag{14}$$

In other words, there exists a discrete distribution supported on $\{1, \gamma_1, \ldots, \gamma_m\}$ whose moments match z. The probability masses λ_j can be computed directly by solving for the coefficients λ_j in (14), which can be done efficiently via a linear program.

Appendix E. The Set of First and Second Moments on the Euclidean Ball

In this section, we provide similar results as in the previous appendix section but for the case where the outcomes y live in the unit ball in $\mathbb{R}^d$. We start by recalling the definition of the moment set:

$$\mathcal{Z} = \{(v, Q) : \exists \text{ a probability measure } \mu \text{ over } \|y\|_2 \leq 1 \text{ such that } \mathbb{E}_\mu[y] = v, \mathbb{E}_\mu[yy^\top] = Q\} \tag{15}$$

11. We remark that the feasible set is not empty, since all $a \in \mathcal{Z}^*$ must be such that $a^\top w \geq 0$ for all $w \in \mathcal{Z}$, and $1 \in \mathcal{Z}$ since it is the moment vector of the point mass distribution with support $\{1\}$.

Since $\|y\|_2 \leq 1$ and $\text{Tr}[Q] = \mathbb{E}_\mu \|y\|_2^2$, it must be the case that $\text{Tr}[Q] \leq 1$ and that $Q = \mathbb{E}_\mu[yy^\top] \succeq 0$. Furthermore, for $(v, Q) = (\mathbb{E}_\mu[y], \mathbb{E}_\mu[yy^\top])$, it is also true that:

$$\mathbb{E}_\mu[(y - \mathbb{E}_\mu[y])(y - \mathbb{E}_\mu[y])^\top] \succeq 0 \iff Q - vv^\top = \mathbb{E}_\mu[yy^\top] - \mathbb{E}_\mu[y]\mathbb{E}_\mu[y]^\top \succeq 0.$$

Using these probability facts, stipulating necessary conditions on $(v, Q) \in \mathcal{Z}$, we conclude that $\mathcal{Z} \subseteq \mathcal{M}$ where,

$$\mathcal{M} = \{(v, Q) : Q \succeq vv^\top, Q \succeq 0, \text{Tr}[Q] \leq 1\}.$$

Next, we show that given any $(v, Q) \in \mathcal{M}$ we can construct a probability measure μ over points y in the unit ball using Algorithm 3 such that $\mathbb{E}_\mu[y] = v$ and $\mathbb{E}_\mu[yy^\top] = Q$. This both solves the problem of finding a measure μ with appropriate moments and shows that $\mathcal{M} \subseteq \mathcal{Z}$. Since we had already argued that $\mathcal{Z} \subseteq \mathcal{M}$, this establishes that $\mathcal{Z} = \mathcal{M}$ and solves the problem of characterizing the set $\mathcal{Z}$ in terms of PSD constraints.

We establish the correctness of Algorithm 3 in the following proposition. Similar results and constructions have appeared throughout the literature, we include the proof here simply for the sake of having a self-contained exposition.

Proposition 13 *Let (v, Q) be any element in $\mathcal{M}$. Algorithm 3 returns an atomic measure μ supported on $2d + 1$ points such that $\mathbb{E}_\mu[y] = v$ and $\mathbb{E}_\mu[yy^\top] = Q$.*

Proof Consider the SVD of $\Sigma = Q - vv^\top = \sum_{i=1}^d \sigma_i u_i u_i^\top$. We can assume without loss of generality that the u_i form an orthonormal basis for $\mathbb{R}^d$ and that $\sigma_i \geq 0$ for all i. Also for $i \in [d]$, the equation,

$$\|v + tu_i\|_2^2 = 1 \iff t^2 + 2tv^\top u_i - (1 - \|v\|^2) = 0,$$

always has either 2 solutions t_i^+ and t_i^- since $\|v\|^2 = \text{Tr}[vv^\top] \leq \text{Tr}[Q] \leq 1$ or 1 solution (which is $t = 0$) if $\|v\| = 1$. If there are 2 solutions, by the quadratic formula $t_i^+ t_i^- = -(1 - \|v\|_2^2) \leq 0$. Hence we can let $t_i^+ > 0$ and $t_i^- < 0$. We will deal with the case where there are 2 solutions that are non-zero ($\|v\| < 1$) and address the other case later. We define

$$y_i^+ = v + t_i^+ u_i,\ y_i^- = v + t_i^- u_i \quad \text{and} \quad \lambda_i^+ = \frac{\sigma_i}{t_i^+(t_i^+ - t_i^-)} > 0, \lambda_i^- = \frac{\sigma_i}{t_i^-(t_i^- - t_i^+)} > 0.$$

We can check that $\Lambda_i = \lambda_i^+ + \lambda_i^- = \sigma_i/(1 - \|v\|_2^2)$. Recall that the measure is,

$$\mu = \lambda_0 \delta(v) + \sum_{i=1}^d [\lambda_i^+ \delta(y_i^+) + \lambda_i^- \delta(y_i^-)],$$

where $\lambda_0 = 1 - \sum_{i=1}^d \Lambda_i$. By construction, the total weight amongst all of the $2d + 1$ points in the measure is:

$$\lambda_0 + \sum_{i=1}^d \Lambda_i = \lambda_0 + \frac{\sum_{i=1}^d \sigma_i}{1 - \|v\|^2} = 1.$$

Note that by definition of (v, Q), $1 - \|v\|_2^2 \geq \text{Tr}[Q] - \|v\|_2^2 = \text{Tr}[Q - vv^\top] = \sum_{i=1}^d \sigma_i \geq 0$ hence the fraction is well-defined. Here we note that if $\|v\| = 1$ (the other case) we are effectively just adding weight on the v so there is no issue.

It remains to show that μ has the right first and second moments. A direct calculation shows that for every i, $\lambda_i^+ t_i^+ + \lambda_i^- t_i^- = 0$. Using this, and plugging in the definitions of y_i^+, y_i^-, we get that

$$\mathbb{E}_\mu[y] = \lambda_0 v + \sum_{i=1}^{d}(\lambda_i^+ y_i^+ + \lambda_i^- y_i^-) = \lambda_0 v + \sum_{i=1}^{d}(\lambda_i^+ + \lambda_i^-)v + \sum_{i=1}^{d}(\lambda_i^+ t_i^+ + \lambda_i^- t_i^-)u_i = v$$

To verify the second moment, we use the fact that $\lambda_i^+(t_i^+)^2 + \lambda_i^-(t_i^-)^2 = \sigma_i$ and plug in again the definition of μ to get that:

$$\mathbb{E}_\mu[(y-v)(y-v)^\top] = \sum_{i=1}^{d}(\lambda_i^+(t_i^+)^2 + \lambda_i^-(t_i^-)^2)u_i u_i^\top = \sum_{i=1}^{d}\sigma_i u_i u_i^\top = Q - vv^\top.$$

■

Appendix F. Further Details on the Examples in Section 1.2

Predicting Correlated Rain Across the United States. The particular guarantees stated there are essentially a restatement of Theorem 12. In particular, let $k((x,p),(x',p')) = x^\top x' + p^\top p'$ be the linear, scalar-valued kernel. Then, $\Gamma = I \cdot k$ is a matrix-valued kernel that contains all the functions,

$$h(x,p) = Ax + Cp$$

where $\|h\|_{\mathcal{H}}^2 = \|A\|_F^2 + \|C\|_F^2$ and $\|\cdot\|_F$ denotes the Frobenius norm. Note that

$$\|\Gamma((x,p),(x,p))\|_{\text{op}} = \|I(\|x\|_2^2 + \|p\|_2^2)\|_{\text{op}} \tag{16}$$

By assumption, $\|x\|_2^2 \le B^2$. Also, since $p = (\mathbb{E}_{\mu_t}[\widetilde{y}_t], \mathbb{E}_{\mu_t}[\widetilde{y}_t\widetilde{y}_t^\top])$

$$\|p\|_2^2 = \|\mathbb{E}_{\mu_t}[\widetilde{y}_t]\|_2^2 + \|\mathbb{E}_{\mu_t}[\widetilde{y}_t\widetilde{y}_t^\top]\|_F^2$$

The first term is at most 1 since $\widetilde{y}_t$ is on the unit ball. By Jensen's, $\|\mathbb{E}_{\mu_t}[\widetilde{y}_t\widetilde{y}_t^\top]\|_F^2 \le \mathbb{E}_{\mu_t}\|[\widetilde{y}_t\widetilde{y}_t^\top]\|_F^2 \le 1$ and $\|\Gamma((x,p),(x,p))\|_{\text{op}} \le B^2 + 2$.

Defensive Generation guarantees indistinguishability with respect to all functions of the form $f(x,p,y) = h(x,p)^\top s(y)$ where, as in Equation (9), $s(y)$ is a vector containing $\widetilde{y}_t$ and $\widetilde{y}_t\widetilde{y}_t^\top$.

The specific examples in that section correspond to cases where A and C are zero everywhere except for one row corresponding to a specific entry in $s(y)$. In this case, the Frobenius norm of the matrices just becomes the ℓ_2 norm of that row. The rest of the statement follows from Theorem 12.