Supplementary Materials for "Prediction with Finitely many Errors Almost Surely"

1 Omitted proofs in Example 1 and Example 2

We first introduce the following lemma, which is due to LeCam (1973).

Lemma 1 (Le Cam's Two Point Theorem)**.** *Let p_0, p_1 be two distribution over the same probability space $\mathcal{X}$. Then for any estimator $\Phi : \mathcal{X} \to \{0,1\}$ we have*

$$\max_{i\in\{0,1\}} \Pr_{X\sim p_i}(\Phi(X) \neq i) \geq \frac{1 - ||p_0 - p_1||_{TV}}{2},$$

where $||\cdot||_{TV}$ denotes for total variation.

Proof of the sufficiency part in Example 1:
Let $\mathcal{P}$ be the class of all random processes over $\{0,1\}^\infty$ such that for any $p \in \mathcal{P}$ there exists a parameter $b \in \{0,1\}$ and strictly monotonically increasing integer sequence $M_1, M_2, \cdots$ with $M_1 = 1$ satisfying for all $n \geq 1$, $X_{M_n} = X_{M_n+2} = \cdots = X_{M_{n+1}-1}$, and $X_{M_n}^{M_{n+1}-1}$ is *independent* of all *other* random variables in the process, and

$$p(X_{M_n} = b) = 1 - \frac{1}{(n+1)^2}.$$

Note that a process in $\mathcal{P}$ is purely determined by the parameter b and sequence $\{M_n\}_{n\geq 1}$. We will denote a process to be p_b if the parameter is b. Clearly, by Borel-Cantelli lemma $\mathcal{P}$ is $e.a.s.$-predictable under the loss ℓ in Example 1 by predicting X_{n-1} at each step n. We now show that there is no nesting $\{\mathcal{P}_i, i \geq 1\}$ of $\mathcal{P}$ such that $(\mathcal{P}_i, \ell)$ is η-predictable for all $\eta > 0$ and $i \geq 1$. Our approach is a proof by contradiction—should such a decomposition exist, we construct a distribution in $\mathcal{P}$ that is not in $\bigcup_{i\geq 1} \mathcal{P}_i$, a contradiction on our supposition that $\mathcal{P} = \bigcup_{i\geq 1} \mathcal{P}_i$.

Let R_n be a number such that the class $\mathcal{P}_n$ is η_n-predictable with a sample of size R_n, where η_n will be determined later. Wolog, we may assume R_n to be strictly increasing on n. Let p_0, p_1 be two distributions in $\mathcal{P}$ that are associated with the sequence $M_1 = 1$ and $(M_n = R_{n-1} + 1)_{n\geq 2}$ with parameter $b = 0$ and $b = 1$ respectively, i.e. p_0, p_1 share the same partition of independent blocks but with different parameter.

Let $||p_0^{R_n} - p_1^{R_n}||_{TV} = 1 - \epsilon_n$, where $p_i^{R_n}$ is the distribution of p_i on the first length-R_n binary strings. Observe that the probability of any length-R_n binary string under either p_0 or p_1 is purely a function of the number of blocks which are all-0 (or equivalently all-1), and therefore, so is ϵ_n, the total variation distance. Hence ϵ_n does not depend on the sequence R_n, or η_n, and in particular we can choose $\eta_n < \frac{\epsilon_n}{2}$.

By Lemma 1, we know that any prediction rule will make error with probability at least $\frac{\epsilon_n}{2}$ on either p_0 or p_1 at time step $R_n + 1$ if they both belong to $\mathcal{P}_n$. Since $\epsilon_n/2 > \eta_n$, we conclude that at least one of p_0 or p_1 cannot be in $\mathcal{P}_n$, and that the above conclusion works for all n.

But $\mathcal{P}_n \subset \mathcal{P}_m$ for all $m \geq n$. Let n be the smallest number that contains one of p_0 or p_1. $\mathcal{P}_n$ cannot contain both, per the argument above, it follows that $\mathcal{P}_m$ contains that distribution for all $m \geq n$. However, the argument above implies that the distribution missing from $\mathcal{P}_n$ cannot be in $\bigcup_{k\geq 1} \mathcal{P}_k$, contradicting the assumption that $\{\mathcal{P}_i, i \geq 1\}$ is a nesting of $\mathcal{P}$.

The following lemmas are used in Example 2:

Lemma 2. *Let $\mathcal{B}_k$ be the set of all Bernoulli processes with parameters in*

$$\mathcal{S}_k = \{r_1, \cdots, r_k\} \cup \left([0,1] \backslash \bigcup_{i=1}^{\infty} B(r_i, \frac{1}{k2^i}) \right)$$

where $B(r_i, \frac{1}{k2^i})$ is the open balls centered at r_i with radius $\frac{1}{k2^i}$, $r_1, r_2, \cdots$ is an arbitrary enumeration of rational numbers in $[0,1]$. Then $\mathcal{B}_k$ is η-predictable with irrationality loss for any $k \geq 1$ and $\eta > 0$.

Proof. We show that for any $k \in \mathbb{N}$ and $\eta > 0$, there exists b_η such that $\mathcal{B}_k$ is η-predictable with sample size b_η. Let $X_1, \cdots, X_n$ be an *i.i.d.* sample from some $p \in \mathcal{B}_k$ with $\mathbb{E}[X_i] = \mu$ and $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$. We have $\mathrm{Var}[X_i] \leq 1$. Chebyshev's inequality then shows that

$$p\left(|\bar{X} - \mu| \geq \epsilon\right) \leq \frac{1}{n\epsilon^2}.$$

Fix $\epsilon = \frac{1}{k2^{k+1}}$. Let b_η be a number large enough so that $\frac{1}{b_\eta \epsilon^2} < \eta$. Therefore for $n > b_\eta$, $p\left(|\bar{X} - \mu| \geq \epsilon\right)$ is less than η. Thus, we can conclude that $\mathcal{B}_k$ is η-predictable by simply predicting the irrationality of element in $\mathcal{S}_k$ that is closest to $\bar{X}$ at step b_η, retaining the prediction perpetually thereafter. □

Lemma 3. *Let $\mathcal{B}$ be a class of Bernoulli processes with parameters in $\mathcal{S}$, if $\mathcal{B}$ is η-predictable w.r.t. the irrationality loss for some $0 < \eta < \frac{1}{2}$, then*

$$\inf\{|x - r| : x, r \in \mathcal{S} \text{ and } r \in \mathbb{Q},\ x \in [0,1] \backslash \mathbb{Q}\} > 0. \tag{1}$$

Proof. By definition of η-predictability, there exists a number N_η and prediction rule Φ_η such that Φ_η makes no errors after step N_η with probability at least $1 - \eta$ for all $p \in \mathcal{B}$. Suppose, otherwise, that the infimum in equation (1) is 0. We now select two sources p_0, p_1 from $\mathcal{B}$ with parameters b_0, b_1 respectively, such that b_0 is rational and b_1 is irrational and $|b_0 - b_1| < \frac{1-2\eta}{2N_\eta}$.

We now have $||p_0^{N_\eta} - p_1^{N_\eta}||_{TV} < 1 - 2\eta$, where $p_i^{N_\eta}$ is the distribution of p_i restricted to the first N_η samples—using the fact that $||p^N - q^N||_{TV} \leq N||p - q||_{TV}$ for any distributions p, q with N-fold *i.i.d.* distributions p^N, q^N. Now, by Lemma 1, any prediction rule (in particular Φ_η) will make an error at time step $N_\eta + 1$ with probability $> \eta$ for either p_0 or p_1. This contradicts η predictability of Φ_η on $\mathcal{B}$. □

Lemma 4. *Let $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \cdots \subset \mathcal{S}_k \subset \cdots \subset [0,1]$ be countably many sets, such that*

$$\forall k,\ \inf\{|x - r| : x, r \in \mathcal{S}_k \text{ and } r \in \mathbb{Q},\ x \in [0,1] \backslash \mathbb{Q}\} > 0. \tag{2}$$

If $\bigcup_{k \in \mathbb{N}} \mathcal{S}_k$ contains all rational numbers in $[0,1]$, then the irrational numbers in $\mathcal{S}_k$ are nowhere dense in $[0,1]$ for all k.

Proof. Suppose otherwise, the set of irrational numbers $\mathcal{I}_k$ in $\mathcal{S}_k$ is not nowhere dense. By definition, there exists an interval $[a, b] \subset \mathrm{col}(\mathcal{I}_k)$, where col denotes for closure. Since the rational numbers in $[0,1]$ are dense, there exists some rational number $r \in [a, b]$, and therefore $r \in \mathrm{col}(\mathcal{I}_k)$. Since $r \in \bigcup_{k \in \mathbb{N}^+} \mathcal{S}_k$, there exist some $k' \geq k$ such that $r \in \mathcal{S}_{k'}$. However, we also have $\mathcal{S}_k \subset \mathcal{S}_{k'}$. Which implies that r is the limit point of irrational numbers in $\mathcal{S}_{k'}$, contradicting the assumption (2) . □

2 Proof of Theorem 4

Suppose $(\mathcal{P}, \ell)$ is e.a.s.-learnable. Then for each i, we let Φ_i and τ_i be the predictor and stopping rule pair respectively that learns with confidence $1/2^i$. By definition, we have that the probability Φ_i makes an error after τ_i stops (i.e. $\tau_i = 1$) is $\leq \frac{1}{2^i}$. Let Φ_0 be an arbitrary predictor.

Now, there are countably many stopping rules (one for each natural number $i \geq 0$) and each such rule stops at a finite time with probability 1, we conclude that with probability 1 all of them would have stopped simultaneously at some finite time by a union bound.

We initialize $t = 1$ (t will stand for the stage). As we see more of the sample, at any stage t, we predict using the prediction rule Φ_{t-1}, till τ_t halts (i.e. $\tau_t = 1$). At that point, we move to stage $t + 1$. For $t \geq 2$, the probability of making an error in stage t is $\leq 2^{-t}$. Invoking the Borel-Cantelli lemma, we conclude that we make an error in finitely many stages almost surely, and the Theorem follows.

3 Proof of Corollary 1

We show that the open conjecture posed in (Dembo and Peres, 1994) is true if we assume some uniform bounds on the density. Let H_0, H_1 be disjoint sets of distributions over $\mathbb{R}^d$. The problem is to predict whether an underlying distribution $p \in H_0 \cup H_1$ is in H_0 or H_1, using *i.i.d.* samples from p. Therefore the class $\mathcal{H}$ will be the *i.i.d.* processes with marginals in $\mathcal{H}_0 \cup \mathcal{H}_1$, our prediction on seeing an sample X^n of size n is $Y_n(X_n) \in \{0, 1\}$, and the loss is $\ell(p, X_1^n, Y_n) = 1\{p \notin \mathcal{H}_{Y_n}\}$.

For any distributions p_1, p_2 over $\mathbb{R}^d$, we consider the Kolmogorov-Smirnov distance (abbreviate as KS-distance)

$$|p_1 - p_2|_\infty = \sup_{\mathbf{x} \in \mathbb{R}^d} |F_{p_1}(\mathbf{x}) - F_{p_2}(\mathbf{x})|,$$

where F_{p_i} is the CDF of p_i. The following lemma is well known in the literature, see e.g. (Athreya and Lahiri, 2006, Theorem 9.1.4).

Lemma 5 (Polya's Theorem)**.** *Let $p_1, p_2, \cdots$ and p be distributions over $\mathbb{R}^d$ with continuous CDF. Then $\lim_{n\to\infty} |p_n - p|_\infty = 0$ iff p_n weakly converges (converges in distribution) to p.*

For any $\mathbf{x} \in \mathbb{R}^d$, we denote $\mathbb{I}_{\mathbf{x}}$ be the indicator function of set $\prod_{i=1}^d (-\infty, \mathbf{x}_i]$, where $\mathbf{x}_i$ is the i'th coordinate of $\mathbf{x}$. Let $X_1, X_2, \cdots, X_n$ be *i.i.d.* distributions over $\mathbb{R}^d$, denote the CDF of the empirical distribution as follows

$$\forall \mathbf{x} \in \mathbb{R}^d,\ F_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\mathbf{x}}(X_i).$$

We have the following lemma, which is know as Dvoretzky-Kiefer-Wolfowitz Inequality, see e.g. Massart (1990); Kiefer and Wolfowitz (1958).

Lemma 6 (Dvoretzky-Kiefer-Wolfowitz Inequality)**.** *Let $X_1, X_2, \cdots, X_n$ be* i.i.d. *samples of distributions p over $\mathbb{R}^d$, $F_n(\boldsymbol{x})$ is the CDF of the empirical distribution. Then there exist an constant C_d depends only on d such that*

$$p\left(|F_n(\boldsymbol{x}) - F_p(\boldsymbol{x})|_\infty \geq \epsilon\right) \leq C_d \exp(-n\epsilon^2).$$

Note that the tail bound given in Lemma 6 only depends on the sample size and is independent of the underlying distribution. We now provide the following theorem, which provide an alternative proof of Theorem 2(i) in Dembo and Peres (1994). We first introduce the following notion, which is equivalent to the F_σ-separability introduced in (Dembo and Peres, 1994), but with a more operational interpretation.

Definition 1. *Let A, B be two disjoint sets in a metric space with metric d. We say A, B are F_σ-separable if there exist nesting $A_1 \subset A_2 \subset \cdots A$ and $B_1 \subset B_2 \subset \cdots B$ such that*

1. *$\bigcup_{n\geq 1} A_n = A$ and $\bigcup_{n\geq 1} B_n = B$.*
2. *For all $n \geq 1$, we have*
$$\inf\{d(x, y) : x \in A_n,\ y \in B_n\} > 0.$$

Theorem 1. *Let $H_0, H_1 \subset \mathcal{M}_1(\mathbb{R}^d)$ be collections of distributions over $\mathbb{R}^d$ that is F_σ-separable under KS-distance. Then $(\mathcal{H}, \ell)$ is e.a.s.-predictable.*

Proof. By definition of F_σ separability, we have nesting $\{A_n\}$, $\{B_n\}$ of H_0, H_1 respectively, such that

$$\forall n \in \mathbb{N},\ \epsilon_n \stackrel{\text{def}}{=} \inf\{|p_0 - p_1|_\infty : p_0 \in A_n, p_1 \in B_n\} > 0.$$

We only need to show that $A_n \cup B_n$ is η-predictable for all η by Theorem 2 (in the main paper). By Lemma 6, we can simultaneously make $|F_n - F_p|_\infty \leq \epsilon_n/4$ with confidence $1-\eta$ for all $p \in A_n \cup B_n$ by choosing the sample size large enough. By triangle inequality of KS-distance, one can classify the distributions in $A_n \cup B_n$ successfully with probability at least $1 - \eta$. □

A collection $\mathcal{H}$ of distributions over $\mathbb{R}^d$ with density functions is said to be *uniformly bounded* if for all $\epsilon > 0$, there exist a number M_ϵ such that

$$\forall p \in \mathcal{H},\ p(f_p(\mathbf{x}) \geq M_\epsilon) \leq \epsilon,$$

where f_p is the density function of p. We have the following theorem

Theorem 2. *Let H_0, H_1 be collections of distributions that are absolutely continuous w.r.t. Lebesgue measure on $\mathbb{R}^d$, and $H_0 \cup H_1$ is uniformly bounded. Then $(\mathcal{H}, \ell)$ is e.a.s.-predictable, only if H_0, H_1 is F_σ-separable with KS-distance.*

Proof. By Theorem 2 (in the main paper), there exist nesting $\{A_n\}$, $\{B_n\}$ of H_0, H_1 respectively, such that $A_n \cup B_n$ is $\frac{1}{8}$-predictable. We show that there is no limit point of A_n in B_n or vice versa. Suppose otherwise, there exist $p_1, p_2, \cdots \subset A_n$ and $p \in B_n$, such that $|p_n - p|_\infty \to 0$. Let Φ be an arbitrary predictor that achieves $\frac{1}{8}$-predictability of $A_n \cup B_n$ with sample size n. Let Φ^n be the prediction function at step n. By Lemma 5, p_n weakly converges to p. Therefore, there exists a compact set S, such that $p(S) \geq \frac{7}{8}$ and $p_n(S) \geq \frac{7}{8}$ for all $n \in \mathbb{N}$. By uniform boundedness of $H_0 \cup H_1$, there exists a number M, such that

$$\forall p \in H_0 \cup H_1,\ p(f_p(x) \geq M) \leq \frac{1}{16}. \tag{3}$$

By Lusin's theorem, there exists a continuous function g and a set $E \subset S$, such that $\sup_{\mathbf{x} \in E} |\Phi^n(\mathbf{x}) - g(\mathbf{x})| \leq \frac{1}{4}$ and $m(S \backslash E) \leq \frac{1}{16M}$, where $m(\cdot)$ is Lebesgue measure. Let $\Omega = \{\mathbf{x} : g(\mathbf{x}) > \frac{1}{3}\}$, we have Ω is open and $\{\mathbf{x} \in E : \Phi^n(\mathbf{x}) = 1\} \subset \Omega$. By (3) and because $m(S \backslash E) \leq \frac{1}{8M}$, we have $p(S \backslash E) \leq \frac{1}{8}$ and $p_n(S \backslash E) \leq \frac{1}{8}$ for all $n \in \mathbb{N}$. By $\frac{1}{8}$-predictability, we have $p(\Omega \cap E) \geq \frac{7}{8} - \frac{1}{4} = \frac{5}{8}$. By weak convergence, we have $\liminf p_n(\Omega) \geq p(\Omega)$, since Ω is open. There exist some p_n such that $p_n(\Omega) \geq \frac{1}{2}$ since $p(\Omega) \geq \frac{5}{8}$, which implies $p_n(\Omega \cap E) \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4} > \frac{1}{8}$. Contradicting the $\frac{1}{8}$-predictability. □

We have the following corollary.

Corollary 1 (Corollary 1 in the main paper)**.** *Let $G : \mathbb{R}^+ \to \mathbb{R}^+$ be a strictly monotone increasing function such that $\lim_{x \to \infty} G(x) \to \infty$, H_0, H_1 be collections of distributions over $\mathbb{R}^d$ that is continuous w.r.t. Lebesgue measure. Suppose for all $p \in H_0 \cup H_1$, we have $\mathbb{E}_{X \sim p}[G(f_p(X))] < \infty$. Then $(\mathcal{H}, \ell)$ is e.a.s.-predictable iff H_0, H_1 are F_σ-separable with KS-distance.*

Proof. By breaking $H_0 \cup H_1$ into countably many subcollections, one may assume $\forall p \in H_0 \cup H_1$, $\mathbb{E}_{\mathbf{x} \sim p}[G(f_p(\mathbf{x}))] \leq M$ for some constant M. We only need to show that $H_0 \cup H_1$ is uniformly bounded. For any $p \in H_0 \cup H_1$, define random variable $Y_p = G(f_p(\mathbf{x}))$. We have by Markov inequality $p(Y_p \geq T) \leq \frac{M}{T}$. Note that the probability is independent of p. By letting $T = \frac{M}{\epsilon}$, one can make the probability less than ϵ. Since G is monotone increasing and goes to infinity, it is invertible on $\mathbb{R}^+$. We now have $p(f_p(\mathbf{x}) \geq G^{-1}(M/\epsilon)) \leq \epsilon$ for all $p \in H_0 \cup H_1$ and $\epsilon > 0$. □

Remark 1. *Note that, the condition of Theorem 2(ii) of Dembo and Peres (1994) is equivalent to take $G(x) = x^{q-1}$ for some $q > 1$.*

References

Krishna B Athreya and Soumendra N Lahiri. *Measure theory and probability theory*. Springer Science & Business Media, 2006.

Amir Dembo and Yuval Peres. A topological criterion for hypothesis testing. *Annals of Statistics*, pages 106–117, 1994.

J Kiefer and J Wolfowitz. On the deviations of the empiric distribution function of vector chance variables. *Transactions of the American Mathematical Society*, 87(1):173–186, 1958.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1): 38–53, 1973.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.