

On the Optimal Boolean Function for Prediction under Quadratic Loss

Nir Weinberger, *Student Member, IEEE*, and Ofer Shayevitz, *Senior Member, IEEE*

Abstract

Suppose Y^n is obtained by observing a uniform Bernoulli random vector X^n through a binary symmetric channel. Courtade and Kumar asked how large the mutual information between Y^n and a Boolean function $b(X^n)$ could be, and conjectured that the maximum is attained by a dictator function. An equivalent formulation of this conjecture is that dictator minimizes the prediction cost in a sequential prediction of Y^n under *logarithmic loss*, given $b(X^n)$. In this paper, we study the question of minimizing the sequential prediction cost under a different (proper) loss function – the *quadratic loss*. In the noiseless case, we show that majority asymptotically minimizes this prediction cost among all Boolean functions. We further show that for weak noise, majority is better than dictator, and that for strong noise dictator outperforms majority. We conjecture that for quadratic loss, there is no single sequence of Boolean functions that is simultaneously (asymptotically) optimal at all noise levels.

Index Terms

Boolean functions, sequential prediction, logarithmic loss function, quadratic loss function, Pinsker's inequality.

I. INTRODUCTION AND PROBLEM STATEMENT

Let $X^n \in \{0, 1\}^n$ be a uniform Bernoulli random vector,¹ and let Y^n be the result of passing X^n through a memoryless binary symmetric channel (BSC) with crossover probability $\alpha \in [0, \frac{1}{2}]$. Recently, Courtade and Kumar conjectured the following:

Conjecture 1 ([1]). *For any Boolean function $b(X^n) : \{0, 1\}^n \rightarrow \{0, 1\}$*

$$I(b(X^n); Y^n) = H(Y^n) - H(Y^n | b(X^n)) \leq 1 - h_b(\alpha) \quad (1)$$

where $h_b(\alpha) := -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha)$ is the binary entropy function.²

The work of the first author was supported by the Gutwirth scholarship for Ph.D. students of the Technion, Israel Institute of Technology. The work of the second author was supported by an ERC grant no. 639573, and an ISF grant no. 1367/14. The material in this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, July 2016.

¹As customary, upper case letters will denote random variables/vectors, and their lower case counterparts will denote specific values that they take.

²Throughout, the logarithm $\log(t)$ is on base 2, while $\ln(t)$ is the natural logarithm.

Since the *dictator* function $\text{Dict}(x^n) := x_1$ (or any other coordinate) achieves this upper bound with equality, then loosely stated, Conjecture 1 claims that dictator is the most “informative” one-bit quantization of X^n in terms of reducing the entropy of Y^n . Despite considerable effort in several directions (e.g. [1], [2], [3], [4]), Conjecture 1 remains generally unsettled. Recently, it was shown in [5] that Conjecture 1 holds for very noisy channels, to wit for all $\alpha \geq \frac{1}{2} - \alpha^*$, for some absolute constant $\alpha^* > 0$.

From a different perspective, defining $Q_k := \mathbb{P}[Y_k = 1 | Y^{k-1}, \mathbf{b}(X^n)]$, and using the chain rule, we can write

$$\begin{aligned} H(Y^n | \mathbf{b}(X^n)) &= \sum_{k=1}^n H(Y_k | Y^{k-1}, \mathbf{b}(X^n)) \\ &= \sum_{k=1}^n \mathbb{E} [\ell_{\log}(Y_k, Q_k)] \end{aligned} \quad (2)$$

where $\ell_{\log}(b, q) := -\log[1 - q - b(1 - 2q)]$ is the *binary logarithmic loss* function.³ Thus, the most informative Boolean function $\mathbf{b}(x^n)$ can also be interpreted as the one that minimizes the (expected) *sequential prediction cost* incurred when predicting the sequence $\{Y_k\}$ from its past, under logarithmic loss, and given $\mathbf{b}(X^n)$. It is important to note that the logarithmic loss function is *proper*, i.e., corresponds to a *proper scoring rule* [6].⁴ This means that using the true conditional distribution Q_k as the predictor for Y_k is guaranteed to minimize the expected prediction cost at time k .

Given the above interpretation, it seems natural to ask the same question for other loss functions. Namely, what is the minimal sequential prediction cost of $\{Y_k\}$ incurred under a general loss function $\ell : \{0, 1\} \times [0, 1] \rightarrow \mathbb{R}_+$,

$$L(Y^n | \mathbf{b}(X^n)) := \sum_{k=1}^n \mathbb{E} [\ell(Y_k, Q_k)], \quad (3)$$

and what is the associated optimal Boolean function $\mathbf{b}(x^n)$? Specifically, it makes sense to consider proper loss functions, as for such functions the optimal prediction strategy is “honest”. The family of proper loss functions contains many members besides the logarithmic loss; in fact, the exact characterization of this family is well known [6]. In this work we focus on another prominent member of this family, the *quadratic loss function*. This loss function is simply the quadratic distance between the expected guess and the outcome. In the binary case, it is given by $\ell_{\text{quad}}(b, q) := (b - q)^2$. Following that, we can define the *sequential mean squared error* (SMSE) to be the (expected) sequential prediction cost of Y^n incurred under quadratic loss given $\mathbf{b}(X^n)$, namely

$$\begin{aligned} M(Y^n | \mathbf{b}(X^n)) &:= \sum_{k=1}^n \mathbb{E} [\ell_{\text{quad}}(Y_k, Q_k)] \\ &= \sum_{k=1}^n \mathbb{E} [Q_k(1 - Q_k)] \end{aligned}$$

³The first argument of $\ell_{\log}(b, q)$ represents the outcome of the next bit, and the second argument is the probability assignment for the bit being 1.

⁴Scoring rules are typically defined in the literature as a quantity to maximize, hence are the negative of cost functions.

$$:= \sum_{k=1}^n M(Y_k | Y^{k-1}, \mathbf{b}(X^n)). \quad (4)$$

In what follows, we show that for $\alpha = 0$ (noiseless channel) the SMSE is asymptotically minimized by the majority function.⁵ We further show that majority is better than dictator for small α . This might tempt one to conjecture that majority is always asymptotically optimal for SMSE. However, we show that dictator is in fact better than majority for α close to $\frac{1}{2}$. Intuitively, it would seem that dictator is in some sense the function “least affected” by noise, and hence while majority is better at weak noise, dictator “catches up” with it as the noise increases. This intuition sits well Conjecture 1, since for logarithmic loss all (balanced) functions are equally good at $\alpha = 0$. We conjecture that the optimal function under quadratic loss must be close to majority for $\alpha \approx 0$, and close to dictator for $\alpha \approx \frac{1}{2}$. The validity of this conjecture would imply in particular that, in contrast to the common belief in the logarithmic loss case, for quadratic loss there is no single sequence of Boolean functions that is simultaneously (asymptotically) optimal at all noise levels.

II. RESULTS

Let $W_H(x_k^m)$ be the Hamming weight of x_k^m . We denote the majority function by $\text{Maj}(x^n)$, which is equal to 1 whenever $W_H(x^n) > \frac{n}{2}$, and 0 whenever $W_H(x^n) < \frac{n}{2}$. When n is odd this definition is unambiguous, but when n is even, the values of $\text{Maj}(x^n)$ when $W_H(x^n) = \frac{n}{2}$ are not defined, and any arbitrary choice of assignment of values to $\text{Maj}(x^n)$ is proper for our needs.

In the noiseless case ($\alpha = 0$), the assertion in Conjecture 1 for the logarithmic loss is trivial, and equality is obtained for any *balanced* function ($\mathbb{P}[\mathbf{b}(X^n) = 1] = \frac{1}{2}$), and specifically, for the dictator function. By contrast, for quadratic loss, finding the optimal function seems far from trivial even for $\alpha = 0$. In the next theorem we provide a lower bound on the noiseless SMSE for any Boolean function, and show that the majority function asymptotically achieves it.

Theorem 2 (Noiseless case). *For any Boolean function $\mathbf{b}(X^n)$*

$$M(X^n | \mathbf{b}(X^n)) \geq \frac{n - 2 \ln 2}{4}, \quad (5)$$

and for majority

$$M(X^n | \text{Maj}(X^n)) \leq \frac{n - 2 \ln 2}{4} + o(1). \quad (6)$$

Clearly, for dictator

$$M(X^n | \text{Dict}(X^n)) = \frac{n - 1}{4} \quad (7)$$

which is strictly worse than the SMSE of the majority function. In fact, it is easy to see that dictator in fact maximizes the SMSE.

⁵In fact, for balanced functions, it is trivially maximized by the dictator.

	$M(X^n \text{Maj}(X^n))$	$\min_{b(\cdot)} M(X^n b(X^n))$	Excess SMSE of majority	Lower bound (5)
$n = 3$	0.4792	0.4792	0	0.4034
$n = 5$	0.9676	0.9686	0.0010	0.9034
$n = 7$	1.4552	1.4618	0.0066	1.4034
$n = 9$	1.9483	1.9569	0.0086	1.9034
$n = 11$	2.4435	2.4532	0.0097	2.4034

Table I
SMSE OF MAJORITY AND SMSE OF THE OPTIMAL FUNCTION, AND (5).

The minimal SMSE for moderate values of n , can be found efficiently. The idea is to trace, for each n , the optimal functions $\{b_w^{(n)}\}_{w \in \{0,1,\dots,2^n\}}$ under a weight constraint

$$b_w^{(n)} := \arg \min_{b(\cdot): |\{x^n: b(x^n)=1\}|=w} M(X^n | b(X^n)). \quad (8)$$

The optimal function $b^{(n)}$ is then given by optimizing over w , i.e.,

$$b^{(n)} := \arg \min_{w \in \{0,1,\dots,2^n\}} M(X^n | b_w^{(n)}(X^n)). \quad (9)$$

Now, assuming that $\{b_w^{(n)}\}$ were found for all input of size less than n , $b_w^{(n+1)}$ can be found by partitioning it into two functions of input size n - one pertaining to $x_1 = 0$ and the other to $x_1 = 1$. Indeed, observing (4) for any given function $b(\cdot)$, it can be noted that the SMSE of the first time point, i.e., $M(X_k | X^{k-1}, b(X^n))$, depends only on the weights $w_0 = |\{x_2^n : b(0, x_2^n) = 1\}|$ and $w_1 = |\{x_2^n : b(1, x_2^n) = 1\}|$. Further, for any given $(w_0, w_1) : w = w_0 + w_1$, the SMSE of all other time points, i.e. $\sum_{k=2}^n M(X_k | X^{k-1}, b(X^n))$, is minimized by setting

$$b(0, x_2^{n+1}) = b_{w_0}^{(n)}(x_2^{n+1}) \quad (10)$$

and

$$b(1, x_2^{n+1}) = b_{w_1}^{(n)}(x_2^{n+1}). \quad (11)$$

Hence, given $\{b_w^{(n)}\}$ for all n , we can find $b_w^{(n+1)}$ by simply going over all possible allocation of weights $(w_0, w_1) : w = w_0 + w_1$. The output of such an algorithm is shown in Table I for moderate input sizes. It can be seen that majority is optimal for $n = 3$, but not for $n = 5, 7, 9, 11$. However, Theorem 2 states that the difference tends to 0, as $n \rightarrow \infty$. For $n = 5$, the optimal function disagrees with majority on 4 inputs.

Next, we consider the noisy case $\alpha \in (0, \frac{1}{2}]$, and derive a simple lower bound on the noisy SMSE for any Boolean function. Then, we provide an upper bound and a lower bound for the SMSE of majority.⁶

⁶Eqs. (5) and (6) of Theorem 2 can be obtained as special cases of (12) and (13) of Theorem 3, by setting $\alpha = 0$, but since the proof of the noisy case is based on Theorem 2, we have separated the results on the noiseless and noisy cases to two different theorems.

Theorem 3 (Noisy case). *For any Boolean function $b(X^n)$*

$$M(Y^n|b(X^n)) \geq \frac{n - 2 \ln 2 \cdot (1 - 2\alpha)^2}{4}. \quad (12)$$

Furthermore, for majority

$$M(Y^n|\text{Maj}(X^n)) \leq \frac{n - 2 \ln 2 \cdot (1 - 2\alpha)^2 \cdot [1 - \mu(\alpha)]}{4} + o(1), \quad (13)$$

where

$$\mu(\alpha) := h_b \left(\frac{\arccos(1 - 2\alpha)}{\pi} \right), \quad (14)$$

and

$$M(Y^n|\text{Maj}(X^n)) \geq \frac{n - \frac{1}{2\pi\alpha(1-\alpha)}(1 - 2\alpha)^2}{4} - O\left((1 - 2\alpha)^4\right) + o(1). \quad (15)$$

Since a straightforward derivation shows that for the dictator function,

$$M(Y^n|\text{Dict}(X^n)) = \frac{n - (1 - 2\alpha)^2}{4}, \quad (16)$$

the above theorem implies that majority is asymptotically better than dictator for all $\alpha \in [0, \underline{\alpha}]$ where $\underline{\alpha} \approx 0.0057$, but that on the other hand, there exists $\bar{\alpha} < \frac{1}{2}$ such that dictator is better than majority for all $\alpha \in [\bar{\alpha}, \frac{1}{2})$.

Remark 4. To improve the SMSE, unbalanced majority functions $\text{Maj}_q(\cdot)$ may be proposed, which assign 1 to a set of $q \cdot 2^n$ vectors of maximal Hamming weight, $q \in (0, 1)$. In the noiseless case, such functions cannot asymptotically improve the SMSE, since the lower bound is achieved by ordinary majority functions ($q = \frac{1}{2}$). Furthermore, it can be shown that they offer no improvement even in the noisy case. Indeed, the noiseless SMSE of such functions is

$$M(X^n|\text{Maj}_q(X^n)) \leq \frac{n - 2 \ln 2 \cdot h_b(q)}{4} + o(1), \quad (17)$$

which is minimized for $q = \frac{1}{2}$. In addition, the effect of the noise of the SMSE is related to boundary size between vectors with $\text{Maj}_q(x^n) = 1$ and vectors with $\text{Maj}_q(x^n) = 0$. For any fixed $q \in (0, 1)$, the value of 1 will be assigned by $\text{Maj}_q(\cdot)$ to vectors of Hamming weight $\frac{n}{2} - O(n^{n/2+\rho}) \leq \frac{n}{2} \leq \frac{n}{2} + O(n^{n/2+\rho})$, which is asymptotically the same as for ordinary majority with $q = \frac{1}{2}$. So, the boundary size of $\text{Maj}_q(\cdot)$ is roughly as the boundary size of $\text{Maj}(\cdot)$, and the effect of the noise on the SMSE is asymptotically the same for all $q \in (0, 1)$. Since the noiseless SMSE for $q = \frac{1}{2}$ is minimal, this seems to be the optimal choice even in the presence of noise ($\alpha \in (0, \frac{1}{2})$).

The proofs of Theorems 2 and 3 appear in Sections III and IV, respectively, and will shortly outlined. Throughout the proofs, we will only consider positive sequences of n and so Landau notations should be interpreted with a positive sign. For example, if $a_n = \Theta(n)$ then a_n is a positive sequence, increasing approximately linearly. In addition, we will denote the *binary divergence* by $d_b(\alpha||\beta) := \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{(1-\alpha)}{(1-\beta)}$, and the support of a random vector X^n by $\mathcal{S}_{X^n} := \{x^n : \mathbb{P}(X^n = x^n) > 0\}$. For brevity, we ignore integer constraints throughout the paper, as they do no affect the results.

III. PROOF OF THE NOISELESS CASE THEOREM

In this section, we consider the noiseless case $\alpha = 0$, namely where $X^n = Y^n$ with probability 1, and prove Theorem 2. The outline of the proof is as follows. To prove the lower bound (5) on the SMSE, we use the binary Pinsker inequality to upper bound the quadratic loss using the binary divergence. To prove that majority asymptotically achieves this lower bound, we first note that since $\text{Maj}(X^n)$ is a balanced function, its value does not help predict X_1 at all, and similarly, the gain in SMSE from knowing $\text{Maj}(X^n)$ at the first few time points is negligible. In the same spirit, at the last time point, the value of $\text{Maj}(X^n)$ is only useful if $W_H(x^{n-1}) = \frac{n}{2}$ (assuming odd n), which occurs with negligible probability, and similarly, the gain at the last few time points due to value of $\text{Maj}(X^n)$ is also negligible. Hence, the gain in prediction cost from knowing $\text{Maj}(X^n)$ is mainly obtained in the “middle” time points. However, even at those time points, the gain is moderate and the probability of the next bit, given the past and $\text{Maj}(X^n)$ is still close to $\frac{1}{2}$, with high probability. So, as Pinsker’s inequality is tight around $\frac{1}{2}$, the quadratic loss function can be replaced with a function of the binary divergence. In turn, the binary divergence is related to the entropy, conditioned on $\text{Maj}(X^n)$. The entropy is simpler to handle, since conditioned on $\text{Maj}(X^n)$ the reduction in the entropy of X^n is 1 bit, and this leads directly to (6). It should be noted that while the above intuition is fairly simple, a careful analysis is required for the proof, since a constant deviation $\frac{2 \ln 2}{4}$ from $\frac{n}{4}$ is sought, which does not depend on n . We begin with proving the lower bound (5) using Pinsker’s inequality.

Proof of (5): Suppose that $\mathbb{P}[\text{b}(X^n) = 1] = q$, and let $P_k := \mathbb{P}[X_k = 1 | X^{k-1}, \text{b}(X^n) = 1]$. Conditioning on $\text{b}(X^n) = 1$, X^n is distributed uniformly over a set of size $q \cdot 2^n$ and thus

$$\begin{aligned}
 \mathbb{M}(X^n | \text{b}(X^n) = 1) &= \sum_{k=1}^n \mathbb{E}[P_k(1 - P_k)] \\
 &= \frac{n}{4} - \sum_{k=1}^n \mathbb{E}\left[\left(P_k - \frac{1}{2}\right)^2\right] \\
 &\stackrel{(a)}{\geq} \frac{n}{4} - \frac{2 \ln 2}{4} \sum_{k=1}^n \mathbb{E}[\text{d}_b(P_k || 1/2)] \\
 &= \frac{n}{4} - \frac{2 \ln 2}{4} \sum_{k=1}^n \mathbb{E}[1 - \text{h}_b(P_k)] \\
 &= \frac{n}{4} - \frac{2 \ln 2}{4} [n - H(X^n | \text{b}(X^n) = 1)] \\
 &= \frac{n}{4} + \frac{2 \ln 2 \log(q)}{4}
 \end{aligned} \tag{18}$$

where (a) is using a binary version of Pinsker’s inequality [7, p. 370, Eq. (11.139)]

$$\text{d}_b(\alpha || \beta) \geq \frac{4}{2 \ln 2} (\alpha - \beta)^2 \tag{19}$$

(where equality is achieved iff $\alpha = \beta$). Deriving a similar bound for the event $\text{b}(X^n) = 0$, we obtain (5) from

$$\mathbb{M}(X^n | \text{b}(X^n)) = q \cdot \mathbb{M}(X^n | \text{b}(X^n) = 1) + (1 - q) \cdot \mathbb{M}(X^n | \text{b}(X^n) = 0)$$

$$\begin{aligned}
&\geq \frac{n}{4} - \frac{2 \ln 2 \cdot h_b(q)}{4} \\
&\geq \frac{n}{4} - \frac{2 \ln 2}{4}.
\end{aligned} \tag{20}$$

■

Proving the asymptotic achievability of the lower bound (5) by the majority function is more intricate, and is based on the asymptotic achievability of equality in Pinsker's inequality (19). We will need several definitions and lemmas.

Definition 5. A vector $v^n \in \{0, 1\}^n$ is termed *t-majority vector* if $W_H(v^n) \geq tn$, where $t \in [0, 1]$ is referred to as the *threshold*. A random vector V^n will be termed *t-majority random vector* if it is uniformly distributed over all *t-majority* vectors of length n . Let $\zeta_n(t)$ be the minimal integer larger or equal to tn . A random vector V^n will be termed *pseudo t-majority random vector* if it is uniformly distributed over all *t-majority* vectors of length n , except possibly for some set \mathcal{D}_n , such that $W_H(v^n) = \zeta_n(t)$ for all $v^n \in \mathcal{D}_n$, and there exists $v^n \in \mathcal{S}_{V^n}$ such that $W_H(v^n) = \zeta_n(t)$. For brevity, we will sometime omit the parameter t when $t = \frac{1}{2}$.

The first lemma provides an approximation for the marginal distributions of a *t-majority* random vector.

Lemma 6. Let $\eta \in [0, \frac{1}{2})$ be given. Then, if V^n is a pseudo *t-majority* random vector,

$$\max \left[\frac{1}{2}, t \right] \leq \mathbb{P}[V_k = 1] \leq \max \left[\frac{1}{2}, t \right] + O_\eta \left(\frac{1}{n^{1/2-\eta}} \right) \tag{21}$$

for all $k \in [n]$.

Proof: See Appendix A. ■

Before we continue, we shortly comment on notation conventions. There is obviously a difference between a majority random vector of length k , and the first k coordinates of a majority random vector of length n , when $k < n$. Nonetheless, to avoid double indexing, we will assume that n is large enough but fixed, and the indices of V^n will denote the corresponding components, e.g. V_k^{k+m} are the components (V_k, \dots, V_{k+m}) of the majority random vector V^n .

The following lemma shows that if m_n increases slowly enough, then the entropy loss of 1 bit of a majority random vector V^n , compared to the entropy of a uniform binary i.i.d. random vector, is mainly due to the entropy of the middle part of the vector $V_{m_n+1}^{n-m_n}$. In other words, the conditional entropies of the beginning and end parts are close to their maximal values, given by their length.

Lemma 7. Let $\rho \in (0, \frac{1}{4})$ and $m_n = O(n^{1/4-\rho})$. Then, for a majority random vector V^n

$$H(V_{m_n+1}^{n-m_n}) \leq n - 1 - 2m_n + o(1). \tag{22}$$

Proof: See Appendix A. ■

The following corollary is a weakening of lemma 7.

Corollary 8. Let $\rho \in (0, \frac{1}{4})$ and $m_n = O(n^{1/4-\rho})$. Then, for a majority random vector V^n

$$H(V_1^{n-m_n}) \leq n - 1 - m_n + o(1). \quad (23)$$

Now, consider a time index k which is sufficiently far from the last index n . In the next lemma, we bound the probability that at time k , the number of ones in the vector is still significantly less than the minimal weight $\frac{k}{2}$ of vectors in the support of a majority random vector of length k .

Lemma 9. Let m_n be an increasing positive sequence, and let $\rho \in (0, 1)$ be given. Then, for all majority random vectors V^n with sufficiently large n ,

$$\mathbb{P} \left[W_H(V_1^k) \leq \frac{k-1}{2} - (n-k+1)^{1/2+\rho} \right] \leq 2^{-\Omega(m_n^{2\rho})}, \quad (24)$$

for all $k \in [n - m_n]$.

Proof: See Appendix A. ■

We are now ready to prove that majority functions are asymptotically optimal.

Proof of (6): Let $\rho \in (0, 1/8)$ be given, and define $m_n := n^{1/4-\rho}$. Let us define V^n as the random vector distributed as X^n conditioned on $\text{Maj}(X^n) = 1$. Clearly, V^n is a majority random vector. For any given $k \in [n - m_n]$ let us define the events

$$\begin{aligned} \mathcal{A}_k &:= \left\{ W_H(V_1^k) \geq \frac{k-1}{2} - (n-k+1)^{1/2+\rho} \right\} \\ &= \left\{ W_H(V_1^k) \geq \frac{n}{2} - r_k + 1 \right\} \end{aligned} \quad (25)$$

where $r_k := \frac{(n-k+1)}{2} + (n-k+1)^{1/2+\rho}$. Now, letting $P_k := \mathbb{P}[V_k = 1 | V^{k-1}]$ we have

$$\begin{aligned} M(X^n | \text{Maj}(X^n) = 1) &= \sum_{k=1}^n \mathbb{E}[P_k(1 - P_k)] \\ &= \frac{n}{4} - \sum_{k=1}^n \mathbb{E} \left[\left(P_k - \frac{1}{2} \right)^2 \right] \\ &\leq \frac{n}{4} - \sum_{k=1}^{n-m_n} \mathbb{E} \left[\left(P_k - \frac{1}{2} \right)^2 \right] \\ &\leq \frac{n}{4} - \sum_{k=1}^{n-m_n} \sum_{v^{k-1} \in \mathcal{A}_{k-1}} \mathbb{P}[V^{k-1} = v^{k-1}] \mathbb{E} \left[\left(P_k - \frac{1}{2} \right)^2 | V^{k-1} = v^{k-1} \right]. \end{aligned} \quad (26)$$

Now, let $v^{k-1} \in \mathcal{A}_{k-1}$. Conditioning on $V^{k-1} = v^{k-1}$, we have that V_k^n is a t -majority random vector of length $n - k + 1 \geq m_n$, and its threshold t is less than

$$\begin{aligned} t &\leq \frac{r_k}{n - k + 1} = \frac{1}{2} + \frac{1}{(n - k + 1)^{1/2-\rho}} \\ &\leq \frac{1}{2} + \frac{1}{m_n^{1/2-\rho}}. \end{aligned} \quad (27)$$

So, assuming that n is large enough, Lemma 6 (with $\eta < \rho$) implies that conditioned on the event $V^{k-1} = v^{k-1}$ with $v^{k-1} \in \mathcal{A}_k$

$$\frac{1}{2} \leq P_k \leq \frac{1}{2} + \frac{1}{m_n^{1/2-\rho}} + \frac{1}{m_n^{1/2-\eta}} \leq \frac{1}{2} + O_\eta \left(\frac{1}{n^{1/8-\rho}} \right), \quad (28)$$

for all $k \in [n - m_n]$. Consequently, as Pinsker's inequality is tight around $\frac{1}{2}$,

$$\left(P_k - \frac{1}{2} \right)^2 \geq [1 - o(1)] \frac{\ln 2}{2} d_b(P_k || 1/2) \quad (29)$$

and so

$$\begin{aligned} M(X^n | \text{Maj}(X^n) = 1) &\leq \frac{n}{4} - \frac{2 \ln 2}{4} [1 - o(1)] \times \\ &\sum_{k=1}^{n-m_n} \sum_{v^{k-1} \in \mathcal{A}_{k-1}} \mathbb{P} [V^{k-1} = v^{k-1}] \mathbb{E} [d_b(P_k || 1/2) | V^{k-1} = v^{k-1}]. \end{aligned} \quad (30)$$

Denoting $\tau_k := \mathbb{P} [V^k \notin \mathcal{A}_k]$, we have

$$\begin{aligned} \mathbb{E} [d_b(P_k || 1/2)] &= \sum_{v^{k-1} \in \mathcal{A}_{k-1}} \mathbb{P} [V^{k-1} = v^{k-1}] \mathbb{E} [d_b(P_k || 1/2) | V^{k-1} = v^{k-1}] \\ &+ \sum_{v^{k-1} \notin \mathcal{A}_{k-1}} \mathbb{P} [V^{k-1} = v^{k-1}] \mathbb{E} [d_b(P_k || 1/2) | V^{k-1} = v^{k-1}] \\ &\leq \sum_{v^{k-1} \in \mathcal{A}_{k-1}} \mathbb{P} [V^{k-1} = v^{k-1}] \mathbb{E} [d_b(P_k || 1/2) | V^{k-1} = v^{k-1}] + \tau_{k-1}, \end{aligned} \quad (31)$$

because $d_b(P_k || 1/2) = 1 - h_b(P_k) \leq 1$. Hence,

$$\begin{aligned} M(X^n | \text{Maj}(X^n) = 1) &\leq \frac{n}{4} - \frac{2 \ln 2}{4} [1 - o(1)] \sum_{k=1}^{n-m_n} \{ \mathbb{E} [d_b(P_k || 1/2)] - \tau_{k-1} \} \\ &\stackrel{(a)}{\leq} \frac{n}{4} - \frac{2 \ln 2}{4} [1 - o(1)] [n - m_n - H(V_1^{n-m_n})] \\ &\quad + [1 - o(1)] \sum_{k=1}^{n-m_n} 2^{-cm_n^{2\rho}} \\ &\stackrel{(b)}{\leq} \frac{n}{4} - [1 - o(1)] \frac{2 \ln 2}{4} + o(1) + n 2^{-cm_n^{2\rho}} \\ &= \frac{n}{4} - \frac{2 \ln 2}{4} + o(1), \end{aligned} \quad (32)$$

where (a) is using the chain rule, $d_b(P_k || 1/2) = 1 - h_b(P_k)$, and since from Lemma 9, for some $c > 0$ we have $\tau_k \leq 2^{-cm_n^{2\rho}}$ for all $k \in [n - m_n]$, and (b) is using Corollary 8.

Finally, from symmetry, conditioning on $\text{Maj}(X^n) = 0$ we have

$$M(X^n | \text{Maj}(X^n) = 0) \leq \frac{n}{4} - \frac{2 \ln 2}{4} + o(1) \quad (33)$$

and so (6) is obtained by averaging over $\text{Maj}(X^n)$ (as in (20)). ■

IV. PROOF OF THE NOISY CASE THEOREM

In this section, we consider the noisy case, and prove Theorem 3. The outline of the proof is as follows. The lower bound of (12) is based on the result of the noiseless case (5), while taking into account that a noisy bit Y_k is to be predicted rather than X_k . To prove (13) we use the noiseless SMSE of majority (6), and quantify the loss in the SMSE conditioned on majority, due to the fact that noisy past bits Y^{k-1} are observed, rather than the noiseless X^{k-1} . As in the noiseless case, the “middle” time points contain most of the loss. In addition, we use a bound on $H(Y^n | \text{Maj}(X^n))$ based on the *stability* of majority. Finally, to prove (15) we use a different asymptotic lower bound on $H(\text{Maj}(X^n) | Y^n)$, which is based on the Gaussian approximation of a binomial random variable, resulting from the Berry-Essen central limit theorem. We then apply Pinsker’s inequality, as in the noiseless case, to bound the SMSE via that entropy.

To prove (12) begin with the next lemma, which states a bound on SMSE of a channel output in terms of the input’s SMSE, for any input distribution.

Lemma 10. *For $V \sim \text{Bern}(\beta)$, $Z \sim \text{Bern}(\alpha)$ independent of V , and $W = V + Z$ (modulo-2 sum),*

$$M(W) = \alpha(1 - \alpha) + (1 - 2\alpha)^2 \cdot M(V). \quad (34)$$

Proof: See Appendix A. ■

Lemma 11. *Let $V^n \in \{0, 1\}^n$ be a random vector, and W^n be the output of a BSC with crossover α fed by V^n , i.e. $W^n = V^n + Z^n$, where $Z^n \sim \text{Bern}(\alpha)$, independent of V^n . Then,*

$$M(W^n) \geq \alpha(1 - \alpha) \cdot n + (1 - 2\alpha)^2 \cdot M(V^n) \quad (35)$$

with equality if V^n is a memoryless random vector.

Proof: See Appendix A. ■

Using the above, we can prove (12).

Proof of (12): Consider any Boolean function $b(X^n)$ and suppose that $\mathbb{P}[b(X^n) = 1] = q$. Then,

$$\begin{aligned} M(Y^n | b(X^n)) &\stackrel{(a)}{\geq} \alpha(1 - \alpha) \cdot n + (1 - 2\alpha)^2 \cdot M(X^n | b(X^n)) \\ &= \alpha(1 - \alpha) \cdot n + q(1 - 2\alpha)^2 \cdot M(X^n | b(X^n) = 1) + (1 - q)(1 - 2\alpha)^2 \cdot M(X^n | b(X^n) = 0) \\ &\stackrel{(b)}{\geq} \alpha(1 - \alpha) \cdot n + (1 - 2\alpha)^2 \cdot \frac{(n - 2 \ln 2)}{4} \\ &\geq \frac{n - (1 - 2\alpha)^2 \cdot 2 \ln 2}{4}, \end{aligned} \quad (36)$$

where (a) follows from Lemma 11, and (b) follows from (5). ■

To prove (13), we analyze, in the next two lemmas, the SMSE of a majority random vector V^n , and show that the quadratic loss in the beginning and end of the vector is close to its maximal value of $\frac{1}{4}$ per bit.

Lemma 12. Let $m_n = O(n^{1-\rho})$ for some $\rho \in (0, 1)$. Then, for a majority random vector V^n

$$\mathbb{M}(V_1^{m_n}) = \sum_{k=1}^{m_n} \mathbb{M}(V_k|V_1^{k-1}) \geq \frac{m_n}{4} - o(1). \quad (37)$$

Proof: See Appendix A. ■

Lemma 13. Let $\rho \in (0, \frac{1}{8})$ and $m_n = O(n^{1/4-\rho})$. Then, for a majority random vector V^n

$$\sum_{k=n-m_n+1}^n \mathbb{M}(V_k|V_1^{k-1}) \geq \frac{m_n}{4} - o(1). \quad (38)$$

Proof: See Appendix A. ■

We also need the following bound on the conditional entropy of the output, given a value of the majority of the input.

Lemma 14. Let $\mu(\cdot)$ be as defined in (14). Then,

$$H(Y^n | \text{Maj}(X^n) = 1) \leq n - 1 + \mu(\alpha) + o(1). \quad (39)$$

Proof: See Appendix A. ■

We can now prove (13).

Proof of (13): In (36), it may be observed that due to (6), inequality (b) is in fact an asymptotic equality, up to an $o(1)$ term. So, it remains to bound the loss in the inequality (a) of (36), which we denote by Φ . Let us also denote $m_n = n^{1/4-\rho}$ for some given $\rho \in (0, \frac{1}{4})$. Then, due to symmetry of the majority function, we may condition on the event $\text{Maj}(X^n) = 1$, and the loss of inequality (a) of (36) is

$$\begin{aligned} \Phi &:= \mathbb{M}(Y^n | \text{Maj}(X^n) = 1) - \alpha(1 - \alpha) \cdot n - (1 - 2\alpha)^2 \cdot \mathbb{M}(X^n | \text{Maj}(X^n) = 1) \\ &= \sum_{k=1}^n \mathbb{M}(Y_k | Y^{k-1}, \text{Maj}(X^n) = 1) - \alpha(1 - \alpha) \cdot n - (1 - 2\alpha)^2 \cdot \sum_{k=1}^n \mathbb{M}(X_k | X^{k-1}, \text{Maj}(X^n) = 1) \\ &\stackrel{(a)}{=} (1 - 2\alpha)^2 \cdot \left\{ \sum_{k=1}^n \mathbb{M}(X_k | Y^{k-1}, \text{Maj}(X^n) = 1) - \mathbb{M}(X_k | X^{k-1}, \text{Maj}(X^n) = 1) \right\}, \end{aligned} \quad (40)$$

where (a) is using a derivation similar to (79).

First, using Lemma 12

$$\begin{aligned} &\sum_{k=1}^{m_n} \mathbb{M}(X_k | Y^{k-1}, \text{Maj}(X^n) = 1) - \mathbb{M}(X_k | X^{k-1}, \text{Maj}(X^n) = 1) \\ &\leq \frac{m_n}{4} - \sum_{k=1}^{m_n} \mathbb{M}(X_k | X^{k-1}, \text{Maj}(X^n) = 1) \\ &\leq o(1), \end{aligned} \quad (41)$$

and similarly, using Lemma 13

$$\begin{aligned}
& \sum_{k=m_n+1}^n \mathbb{M}(X_k|Y^{k-1}, \text{Maj}(X^n) = 1) - \mathbb{M}(X_k|X^{k-1}, \text{Maj}(X^n) = 1) \\
& \leq \frac{m_n}{4} - \sum_{k=m_n+1}^n \mathbb{M}(X_k|X^{k-1}, \text{Maj}(X^n) = 1) \\
& \leq o(1).
\end{aligned} \tag{42}$$

Then, from (5) of Theorem 2, and the symmetry of conditioning $\text{Maj}(X^n) = 0$ and $\text{Maj}(X^n) = 1$, we have

$$\sum_{k=1}^n \mathbb{M}(X_k|X^{k-1}, \text{Maj}(X^n) = 1) \geq \frac{n - 2 \ln(2)}{4}, \tag{43}$$

and

$$\begin{aligned}
& \sum_{k=m_n+1}^{n-m_n} \mathbb{M}(X_k|X^{k-1}, \text{Maj}(X^n) = 1) \\
& = \sum_{k=1}^n \mathbb{M}(X_k|X^{k-1}, \text{Maj}(X^n) = 1) - \sum_{k=1}^{m_n} \mathbb{M}(X_k|X^{k-1}, \text{Maj}(X^n) = 1) \\
& \quad - \sum_{k=n-m_n+1}^n \mathbb{M}(X_k|X^{k-1}, \text{Maj}(X^n) = 1) \\
& \geq \sum_{k=1}^n \mathbb{M}(X_k|X^{k-1}, \text{Maj}(X^n) = 1) - \frac{m_n}{4} - \frac{m_n}{4} \\
& \geq \frac{n - 2m_n - 2 \ln(2)}{4}.
\end{aligned} \tag{44}$$

So it remains to upper bound the first term in the sum of (40), viz.

$$\sum_{k=m_n+1}^{n-m_n} \mathbb{M}(X_k|Y^{k-1}, \text{Maj}(X^n) = 1). \tag{45}$$

We follow the outline of the proof of (6) from Theorem 2. Let us denote the random variables $P_k(X^{k-1}) := \mathbb{P}(X_k = 1|X^{k-1}, \text{Maj}(X^n) = 1)$, and $R_k(Y^{k-1}) := \mathbb{P}(X_k = 1|Y^{k-1}, \text{Maj}(X^n) = 1)$, where their arguments will be sometimes omitted for brevity. In what follows, we will prove the existence of sets $\mathcal{B}_k \subset \{0, 1\}^k$ such that $v_k := \mathbb{P}[Y^k \notin \mathcal{B}_k] \leq 2^{-\frac{c}{2} m_n^{2\rho}}$ for some $c > 0$ and for all $k \in \{m_n + 1, \dots, n - m_n\}$, and

$$\frac{1}{2} \leq R_k(y^{k-1}) \leq \frac{1}{2} + O_\eta \left(\frac{1}{n^{1/s-\rho}} \right) \tag{46}$$

for all $y^{k-1} \in \mathcal{B}_{k-1}$. For $y^{k-1} \in \mathcal{B}_{k-1}$ Pinsker's inequality is tight and so

$$\left(R_k(y^{k-1}) - \frac{1}{2} \right)^2 \geq [1 - o(1)] \frac{\ln 2}{2} d_b(R_k(y^{k-1}) || 1/2). \tag{47}$$

Hence,

$$\begin{aligned}
& \sum_{k=m_n+1}^{n-m_n} \mathbb{M}(X_k|Y^{k-1}, \text{Maj}(X^n) = 1) \\
&= \sum_{k=m_n+1}^{n-m_n} \mathbb{E}[R_k(1 - R_k)] \\
&= \frac{n - 2m_n}{4} - \sum_{k=m_n+1}^{n-m_n} \mathbb{E}\left[\left(R_k - \frac{1}{2}\right)^2\right] \\
&\leq \frac{n - 2m_n}{4} - \sum_{k=m_n+1}^{n-m_n} \sum_{y_1^{k-1} \in \mathcal{B}_{k-1}} \mathbb{P}[Y^{k-1} = y_1^{k-1}] \mathbb{E}\left[\left(R_k - \frac{1}{2}\right)^2 | Y^{k-1} = y_1^{k-1}\right] \\
&\leq \frac{n - 2m_n}{4} - \frac{2 \ln(2)}{4} [1 - o(1)] \sum_{k=m_n+1}^{n-m_n} \sum_{y_1^{k-1} \in \mathcal{B}_{k-1}} \mathbb{P}[Y^{k-1} = y_1^{k-1}] \mathbb{E}[\mathbf{d}_b(R_k || 1/2) | Y^{k-1} = y_1^{k-1}] \\
&\stackrel{(a)}{\leq} \frac{n - 2m_n}{4} - \frac{2 \ln(2)}{4} [1 - o(1)] \sum_{k=m_n+1}^{n-m_n} \{\mathbb{E}[\mathbf{d}_b(R_k || 1/2)] - v_k\} \\
&\stackrel{(b)}{\leq} \frac{n - 2m_n}{4} - \frac{2 \ln(2)}{4} [1 - o(1)] \sum_{k=m_n+1}^{n-m_n} \mathbb{E}[\mathbf{d}_b(R_k || 1/2)] + o(1) \\
&= \frac{n - 2m_n}{4} - \frac{2 \ln(2)}{4} [1 - o(1)] \left[n - 2m_n - \sum_{k=m_n+1}^{n-m_n} H(X_k | Y^{k-1}, \text{Maj}(X^n) = 1) \right] + o(1) \\
&\stackrel{(c)}{\leq} \frac{n - 2m_n}{4} - \frac{2 \ln(2)}{4} [1 - o(1)] \left[n - 2m_n - \sum_{k=m_n+1}^{n-m_n} H(Y_k | Y^{k-1}, \text{Maj}(X^n) = 1) \right] + o(1) \\
&= \frac{n - 2m_n}{4} - \frac{2 \ln(2)}{4} [1 - o(1)] [n - 2m_n - H(Y_{m_n+1}^{n-m_n} | Y^{m_n}, \text{Maj}(X^n) = 1)] + o(1) \\
&\stackrel{(d)}{\leq} \frac{n - 2m_n}{4} - \frac{2 \ln(2)}{4} [1 - o(1)] [n - H(Y^n | \text{Maj}(X^n) = 1)] + o(1) \\
&\stackrel{(e)}{\leq} \frac{n - 2m_n}{4} - \frac{2 \ln(2)}{4} [1 + \mu(\alpha)] + o(1), \tag{48}
\end{aligned}$$

(a) is since, just as in (31),

$$\mathbb{E}[\mathbf{d}_b(R_k || 1/2)] \leq \sum_{y_1^{k-1} \in \mathcal{B}_{k-1}} \mathbb{P}[Y^{k-1} = y_1^{k-1}] \mathbb{E}[\mathbf{d}_b(R_k || 1/2) | Y^{k-1} = y_1^{k-1}] + v_k, \tag{49}$$

(b) is since $v_k \leq 2^{-\frac{\epsilon}{2} m_n^{2\rho}}$, (c) is using

$$\begin{aligned}
H(Y_k | Y^{k-1}, \text{Maj}(X^n) = 1) &= H(X_k + Z_k | Y^{k-1}, \text{Maj}(X^n) = 1) \\
&\geq H(X_k + Z_k | Y^{k-1}, Z_k, \text{Maj}(X^n) = 1) \\
&= H(X_k | Y^{k-1}, Z_k, \text{Maj}(X^n) = 1) \\
&= H(X_k | Y^{k-1}, \text{Maj}(X^n) = 1), \tag{50}
\end{aligned}$$

where the last equality is since Z_k is independent of (X_k, Y^{k-1}) . Transition (d) in (48) follows from

$$\begin{aligned}
H(Y_{n-m_m+1}^n | Y_1^{n-m_n}, \text{Maj}(X^n) = 1) &\stackrel{(i)}{\geq} H(Y_{n-m_m+1}^n | X_1^{n-m_n}, \text{Maj}(X^n) = 1) \\
&= H(X_{n-m_m+1}^n + Z_{n-m_m+1}^n | X_1^{n-m_n}, \text{Maj}(X^n) = 1) \\
&\geq H(X_{n-m_m+1}^n + Z_{n-m_m+1}^n | X_1^{n-m_n}, Z_{n-m_m+1}^n, \text{Maj}(X^n) = 1) \\
&= H(X_{n-m_m+1}^n | X_1^{n-m_n}, \text{Maj}(X^n) = 1) \\
&\stackrel{(ii)}{\geq} m_n - o(1),
\end{aligned} \tag{51}$$

where here (i) follows from the data processing theorem and the fact that $Y_1^{n-m_n} - X_1^{n-m_n} - Y_{n-m_m+1}^n$, and (ii) follows from (76) (proof of Lemma 7), and using a similar bound to $H(Y_1^{m_n} | Y_{m_n+1}^n, \text{Maj}(X^n) = 1)$. Transition (e) in (48) follows from Lemma 14. To conclude, combining (40), (41), (42), (44) and (48) implies that

$$\Phi \leq (1 - 2\alpha)^2 \cdot \frac{2 \ln 2}{4} \mu(\alpha) + o(1), \tag{52}$$

which, together with (36) implies (13).

To complete the proof, it remains to assert the existence of the sets \mathcal{B}_k . To this end, recall that in the proof of (6) in Section III, we have defined the sets

$$\mathcal{A}_k := \left\{ W_H(V_1^k) \geq \frac{k-1}{2} - (n-k+1)^{1/2+\rho} \right\} \tag{53}$$

(cf. (25)) and showed that $\frac{1}{2} \leq P_k(x^{k-1}) \leq \frac{1}{2} + O(1/n^{1/8-\rho})$ for all $x^{k-1} \in \mathcal{A}_{k-1}$. In addition, Lemma 9 implied that there that there exists $c > 0$ such that $\mathbb{P}[X^k \notin \mathcal{A}_k] \leq 2^{-cm_n^2}$ for all $k \in \{m_n+1, \dots, n-m_n\}$. Now, note that

$$\begin{aligned}
R_k(Y^{k-1}) &= \mathbb{P}(X_k = 1 | Y^{k-1}, \text{Maj}(X^n) = 1) \\
&= \sum_{x^{k-1}} \mathbb{P}(X^{k-1} = x^{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1) \cdot \mathbb{P}(X_k = 1 | X^{k-1} = x^{k-1}, Y^{k-1}, \text{Maj}(X^n) = 1) \\
&= \sum_{x^{k-1}} \mathbb{P}(X^{k-1} = x^{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1) \cdot P_k(x^{k-1}),
\end{aligned} \tag{54}$$

so $R_k(Y^{k-1})$ is just an averaging of $P_k(x^{k-1})$. Since $P_k(x^{k-1}) \geq \frac{1}{2}$ for all x^{k-1} , this immediately implies $R_k(y^{k-1}) \geq \frac{1}{2}$. On the other hand

$$\begin{aligned}
R_k(Y^{k-1}) &= \sum_{x^{k-1} \in \mathcal{A}_{k-1}} \mathbb{P}(X^{k-1} = x^{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1) \cdot P_k(x^{k-1}) \\
&\quad + \sum_{x^{k-1} \notin \mathcal{A}_{k-1}} \mathbb{P}(X^{k-1} = x^{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1) \cdot P_k(x^{k-1}) \\
&\leq \frac{1}{2} + O\left(\frac{1}{n^{1/8-\rho}}\right) + \mathbb{P}(X^{k-1} \notin \mathcal{A}_{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1),
\end{aligned} \tag{55}$$

where we have bounded the first term using $P_k(x^{k-1}) \leq \frac{1}{2} + O(1/n^{1/s-\rho})$ for all $x^{k-1} \in \mathcal{A}_{k-1}$, and we have bounded the second term simply by using $P_k(x^{k-1}) \leq 1$. Let us inspect the random variable $\mathbb{P}[X^{k-1} \notin \mathcal{A}_{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1]$. We know that its expected value satisfies

$$\mathbb{E} \left[\mathbb{P} \left(X^{k-1} \notin \mathcal{A}_{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1 \right) \right] = \mathbb{P} \left(X^{k-1} \notin \mathcal{A}_{k-1} | \text{Maj}(X^n) = 1 \right) \leq 2^{-cm_n^{2\rho}}. \quad (56)$$

So, for any given $\eta > 0$ Markov's inequality implies that

$$\mathbb{P} \left[\mathbb{P} \left(X^{k-1} \notin \mathcal{A}_{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1 \right) \geq 2^{\eta m_n^{2\rho}} 2^{-cm_n^{2\rho}} \right] \leq 2^{-\eta m_n^{2\rho}}. \quad (57)$$

Choosing, e.g., $\eta = \frac{\epsilon}{2}$ we get that there exists a set \mathcal{B}_k whose probability is larger than $1 - 2^{-\frac{\epsilon}{2} m_n^{2\rho}}$ such that

$$\mathbb{P} \left(X^{k-1} \notin \mathcal{A}_{k-1} | Y^{k-1}, \text{Maj}(X^n) = 1 \right) \leq 2^{-\frac{\epsilon}{2} m_n^{2\rho}} \quad (58)$$

for all $y^{k-1} \in \mathcal{B}_k$. For this set, we have

$$R_k(Y^{k-1}) \leq \frac{1}{2} + O \left(\frac{1}{n^{1/s-\rho}} \right) + 2^{-\frac{\epsilon}{2} m_n^{2\rho}} = \frac{1}{2} + O \left(\frac{1}{n^{1/s-\rho}} \right), \quad (59)$$

as required. ■

To prove (15) we first need the following approximation to the entropy of majority functions.

Lemma 15 ([8]). *We have*

$$H(\text{Maj}(X^n) | Y^n) = \mathbb{E} \left\{ \mathbf{h}_b \left[Q \left(\frac{|G(1-2\alpha)|}{\sqrt{4\alpha(1-\alpha)}} \right) \right] \right\} + o(1) \quad (60)$$

where $G \sim \mathcal{N}(0, 1)$ is a standard Gaussian random variable, and $Q(\cdot)$ is the Q -function (the tail probability of the standard normal distribution).

Proof: See Appendix A. ■

Remark 16. If we replace Lemma 14 in the proof of (13) with Lemma 15, we can get a sharper bound than (13), yet less explicit.

In the next lemma, we evaluate $H(\text{Maj}(X^n) | Y^n)$ for $\alpha \approx \frac{1}{2}$.

Lemma 17. *We have*

$$H(\text{Maj}(X^n) | Y^n) \geq 1 - \frac{1}{\pi \cdot \ln 2} \left(\frac{(1-2\alpha)^2}{4\alpha(1-\alpha)} \right) - O((1-2\alpha)^4) + o(1). \quad (61)$$

Proof: See Appendix A. ■

We can now prove the lower bound on the SMSE of majority functions (15).

Proof of (15): Using Lemma 17 and a derivation similar to (90), for some $c > 0$, and all α sufficiently close

to $\frac{1}{2}$

$$\begin{aligned} H(Y^n | \text{Maj}(X^n)) &= n - 1 + H(\text{Maj}(X^n) | Y^n) \\ &\geq n - \frac{1}{\pi \cdot \ln 2} \left(\frac{(1 - 2\alpha)^2}{4\alpha(1 - \alpha)} \right) - c(1 - 2\alpha)^4 + o(1). \end{aligned} \quad (62)$$

Hence, as in the proof of (5) in Section III

$$\begin{aligned} M(Y^n | \text{Maj}(X^n)) &\geq \frac{n}{4} - \frac{\ln 2}{2} [n - H(Y^n | \text{Maj}(X^n))] \\ &\geq \frac{n}{4} - \frac{1}{2\pi\alpha(1 - \alpha)} \left(\frac{(1 - 2\alpha)^2}{4} \right) - c(1 - 2\alpha)^4 + o(1) \end{aligned} \quad (63)$$

for all sufficiently large n . ■

Remark 18. For the sake of proving (15), we only needed the second-order approximation, given by Lemma 17. However, we note that the expression on the left-hand side of (60) can be evaluated numerically to an arbitrary precision, e.g., via a power series expansion of the analytic function $h_b[Q(t)]$.

V. DISCUSSION AND OPEN PROBLEMS

The question addressed by Conjecture 1 can be equivalently cast as an optimal sequential prediction problem, seeking the Boolean function $b(X^n)$ that minimizes the cost in sequentially predicting the channel output sequence Y^n , under logarithmic loss. Adopting this point of view, it is natural to consider the same sequential prediction problem under other proper loss functions. In this paper, we have focused on the quadratic loss function. We began by considering the noiseless case $Y^n = X^n$, which is trivial under logarithmic loss but quite subtle under quadratic loss, and showed that majority asymptotically achieves the minimal prediction cost among all Boolean functions. For the case of noisy observations, we derived bounds on the cost achievable by general Boolean functions, as well as specifically by majority. Using these bounds, we showed that majority is better than dictator for weak noise, but that dictator catches up and outperforms majority for strong noise. This should be contrasted with Conjecture 1, which surmises that dictator minimizes the sequential prediction cost under logarithmic loss, simultaneously at all noise levels. Thus, viewed through the lens of sequential prediction, the validity of Conjecture 1 appears to possibly hinge on the unique property of logarithmic loss, namely the fact that in the noiseless case all (balanced) Boolean functions result in the exact same prediction cost.

The discussion above leads us to conjecture that under quadratic loss, there is no single sequence of functions $\{b_n(X^n)\}$ that asymptotically minimizes the prediction cost simultaneously at all noise levels. Moreover, it seems plausible that the optimal function must be close to majority for weak noise, and close to dictator for high noise. While it appears that characterizing the optimal function at a given noise level may be difficult, it would be interesting to understand its structural properties, e.g., whether it is monotone, balanced, odd, etc. For logarithmic loss, it is known that the optimal function is monotone [1]. This fact can be easily established by first switching any non-monotone coordinate with the last coordinate (losing nothing due to the entropy chain rule), and then "shifting"

[9] the last coordinate (which can only decrease the cost, as there are no subsequent coordinates). However, monotonicity seems more difficult to establish under quadratic loss, even in the noiseless case; for example, the switching/shifting technique above fails due to the lack of a chain rule under quadratic loss. Finally, it would be interesting to extend this study to non-Boolean functions as well as to other proper loss functions. For example, our results readily indicate that majority is asymptotically optimal in the noiseless case for any loss function that behaves similarly to quadratic loss around $\frac{1}{2}$ (e.g., logarithmic loss). What is the family of proper loss functions for which majority is asymptotically optimal?

ACKNOWLEDGMENT

We are grateful to Or Ordentlich for asking the question in the noiseless case that led to this research. We would also like to thank Or Ordentlich and Omri Weinstein for helpful discussions, and Uri Hadar for pointing out reference [6].

APPENDIX A

MISCELLANEOUS PROOFS

A. Noiseless case

Proof of Lemma 6: First assume that V^n is a t -majority random vector (and not a pseudo t -majority random vector). From symmetry of t -majority random vector, $\mathbb{P}(V_k = 1) = \mathbb{P}(V_1 = 1)$ for all $k \in [n]$, and so it remains to prove the statement for $k = 1$. Let us begin with the case $t \leq \frac{1}{2}$. For $t = 0$ we clearly have $P_k = \frac{1}{2}$. For $t = \frac{1}{2}$, the number M_1 of $\frac{1}{2}$ -majority vectors such that $v_1 = 1$ (M_0 for $v_1 = 0$, respectively) is

$$M_1 = \sum_{m=\frac{n}{2}-1}^{n-1} \binom{n-1}{m}, \quad (64)$$

and

$$M_0 = \sum_{m=\frac{n}{2}}^{n-1} \binom{n-1}{m}, \quad (65)$$

where the index m in the summation above counts the number of allowed ones in the vector v_2^n . So, as $M_1 > M_0$,

$$\mathbb{P}(V_k = 1) = \frac{M_1}{M_0 + M_1} \geq \frac{1}{2}. \quad (66)$$

Moreover, for all n sufficiently large,

$$\begin{aligned} \frac{\binom{\frac{n-1}{2}}{\frac{n}{2}-1}}{\sum_{m=\frac{n}{2}-1}^{n-1} \binom{n-1}{m}} &\leq \frac{\binom{\frac{n-1}{2}}{\frac{n}{2}-1}}{2^{n-1}} \\ &\stackrel{(a)}{\leq} \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{n}} \cdot 2^{(n-1)\left[\mathbf{h}_b\left(\frac{1}{2} - \frac{1}{2(n-1)}\right) - 1\right]} \\ &\leq \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{n}}, \end{aligned} \quad (67)$$

where (a) is using Lemma 19. So

$$\begin{aligned}
\mathbb{P}(V_k = 1) &= \frac{M_1}{M_0 + M_1} \\
&= \frac{\sum_{m=\frac{n}{2}-1}^{n-1} \binom{n-1}{m}}{\sum_{m=\frac{n}{2}}^{n-1} \binom{n-1}{m} + \sum_{m=\frac{n}{2}-1}^{n-1} \binom{n-1}{m}} \\
&= \frac{\sum_{m=\frac{n}{2}-1}^{n-1} \binom{n-1}{m}}{2 \cdot \sum_{m=\frac{n}{2}-1}^{n-1} \binom{n-1}{m} - \binom{n-1}{\frac{n}{2}-1}} \\
&= \frac{1}{2 - \frac{\binom{n-1}{\frac{n}{2}-1}}{\sum_{m=\frac{n}{2}-1}^{n-1} \binom{n-1}{m}}} \\
&\leq \frac{1}{2} + \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{n}},
\end{aligned} \tag{68}$$

where in the last inequality we have used $\frac{1}{2-s} \leq \frac{1}{2} + s$, valid for small s . Now, since P_k is monotonic in t , then clearly

$$P_k \leq \frac{1}{2} + \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{n}}, \tag{69}$$

for all $0 \leq t \leq \frac{1}{2}$.

Now for the case $t \geq \frac{1}{2}$. Using symmetry, the probability that $V_k = 1$ is equal to the total number of ones in the support of V^n , divided by the total number of zeros and ones in the support of V^n . So,

$$\mathbb{P}(V_k = 1) = \frac{\sum_{m=tn}^n \binom{n}{m} \cdot m}{\sum_{m=tn}^n \binom{n}{m} \cdot n} \geq \frac{\sum_{m=tn}^n \binom{n}{m} \cdot tn}{\sum_{m=tn}^n \binom{n}{m} \cdot n} \geq t. \tag{70}$$

On the other hand, denoting $l_n := n^{1/2+\eta}$, for all n sufficiently large,

$$\begin{aligned}
\mathbb{P}(V_k = 1) &= \frac{\sum_{m=tn}^n \binom{n}{m} \cdot \frac{m}{n}}{\sum_{m=tn}^n \binom{n}{m}} \\
&= \frac{\sum_{m=tn}^{tn+l_n} \binom{n}{m} \cdot \frac{m}{n}}{\sum_{m=tn}^n \binom{n}{m}} + \frac{\sum_{m=tn+l_n+1}^n \binom{n}{m} \cdot \frac{m}{n}}{\sum_{m=tn}^n \binom{n}{m}} \\
&\leq \frac{\sum_{m=tn}^{tn+l_n} \binom{n}{m} \cdot (t + \frac{l_n}{n})}{\sum_{m=tn}^{tn+l_n} \binom{n}{m}} + \frac{\sum_{m=tn+l_n+1}^n \binom{n}{m}}{\sum_{m=tn}^n \binom{n}{m}} \\
&= t + \frac{n^\eta}{\sqrt{n}} + \frac{\sum_{m=tn+l_n+1}^n \binom{n}{m}}{\sum_{m=tn}^n \binom{n}{m}} \\
&\leq t + O_\eta \left(\frac{n^\eta}{\sqrt{n}} \right).
\end{aligned} \tag{71}$$

The last inequality follows from

$$\begin{aligned}
\frac{\sum_{m=tn+l_n+1}^n \binom{n}{m}}{\sum_{m=tn}^n \binom{n}{m}} &\stackrel{(a)}{=} \frac{\sum_{m=tn}^n \binom{n}{m+l_n+1}}{\sum_{m=tn}^n \binom{n}{m}} \\
&\stackrel{(b)}{\leq} \max_{tn \leq m \leq n} \frac{\binom{n}{m+l_n+1}}{\binom{n}{m}}
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(c)}{\leq} \max_{tn \leq m \leq n-l_n-1} \frac{\sqrt{8n \frac{m}{n} (1 - \frac{m}{n})}}{\sqrt{\pi n \cdot \frac{m+l_n+1}{n} (1 - \frac{m+l_n+1}{n})}} \cdot \frac{2^{nh_b(\frac{m+l_n+1}{n})}}{2^{nh_b(\frac{m}{n})}} \\
&= [1 + o(1)] \sqrt{\frac{8}{\pi}} \max_{tn \leq m \leq n-l_n-1} 2^n [\mathbf{h}_b(\frac{m}{n} + \frac{n^{\eta/2}}{\sqrt{n}}) - \mathbf{h}_b(\frac{m}{n})] \\
&\stackrel{(d)}{\leq} [1 + o(1)] \sqrt{\frac{8}{\pi}} \max_{\frac{n}{2} \leq m \leq n-l_n-1} 2^n [\mathbf{h}_b(\frac{m}{n} + \frac{n^{\eta}}{\sqrt{n}}) - \mathbf{h}_b(\frac{m}{n})] \\
&\stackrel{(e)}{\leq} \sqrt{\frac{8}{\pi}} \cdot 2^n [\mathbf{h}_b(\frac{1}{2} + \frac{n^{\eta}}{\sqrt{n}}) - \mathbf{h}_b(\frac{1}{2})] \\
&\stackrel{(f)}{\leq} \sqrt{\frac{8}{\pi}} \cdot 2^{-\frac{2}{\ln 2} n^{\eta}}, \tag{72}
\end{aligned}$$

where (a) is using the convention $\binom{n}{m} = 0$ for $m > n$, (b) is using Lemma 20, (c) is using Lemma 19, (d) is as $t \geq \frac{1}{2}$, (e) is because the maximum is obtained at the minimal value of the feasible set, due the concavity of $\mathbf{h}_b(\cdot)$, and (f) is using the inequality $\mathbf{h}_b(\frac{1}{2} + s) \leq 1 - \frac{2}{\ln 2} s^2$.

Finally, the marginal probability of 1 for a pseudo t -majority random vector is only larger than for ordinary t -majority random vector, and smaller than the same marginal probability of a $(t + \frac{1}{n})$ -majority random vector. So, the asymptotic upper bound does not change for pseudo t -majority random vectors. ■

Proof of Lemma 7: From the chain rule for entropies and as conditioning reduces entropy

$$\begin{aligned}
n - 1 &= H(V_1^n) \\
&= H(V_{m_n+1}^{n-m_n}) + H(V_1^{m_n} | V_{m_n+1}^{n-m_n}) + H(V_{n-m_n+1}^n | V_1^{n-m_n}) \\
&\geq H(V_{m_n+1}^{n-m_n}) + H(V_1^{m_n} | V_{m_n}^n) + H(V_{n-m_n+1}^n | V_1^{n-m_n}). \tag{73}
\end{aligned}$$

Now, for any vector v^{n-m_n} such that $W_H(v_1^{n-m_n}) \geq \frac{n}{2} + 1$, it is assured that $v^n \in \mathcal{S}_{V^n}$, no matter what its suffix $v_{n-m_n+1}^n$ is. Thus, conditioning on this event, the suffix is distributed uniformly over $\{0, 1\}^{m_n}$. This implies that

$$H(V_{n-m_n+1}^n | V_1^{n-m_n}) \geq \mathbb{P} \left[W_H(V_1^{n-m_n}) \geq \frac{n}{2} + 1 \right] \cdot m_n. \tag{74}$$

Now, for all sufficiently large n

$$\begin{aligned}
\mathbb{P} \left[W_H(V_1^{n-m_n}) \geq \frac{n}{2} + 1 \right] &= \frac{\sum_{k=\frac{n}{2}+1}^{n-m_n} \binom{n-m_n}{k} \cdot 2^{m_n}}{2^{n-1}} \\
&= \frac{2 \sum_{k=\frac{n}{2}+1}^{n-m_n} \binom{n-m_n}{k}}{2^{n-m_n}} \\
&= \frac{2 \sum_{k=\frac{n-m_n}{2}}^{n-m_n} \binom{n-m_n}{k} - 2 \sum_{k=\frac{n-m_n}{2}}^{\frac{n}{2}} \binom{n-m_n}{k}}{2^{n-m_n}} \\
&\geq 1 - \frac{2 \sum_{k=\frac{n-m_n}{2}}^{\frac{n}{2}} \binom{n-m_n}{k}}{2^{n-m_n}} \\
&\geq 1 - 2 \left(\frac{m_n}{2} + 1 \right) \frac{\binom{n-m_n}{\frac{n-m_n}{2}}}{2^{n-m_n}}
\end{aligned}$$

$$\begin{aligned}
&\geq 1 - 2m_n \frac{\binom{n-m_n}{\frac{n-m_n}{2}}}{2^{n-m_n}} \\
&\geq 1 - 2\sqrt{\frac{4}{\pi(n-m_n)}}m_n,
\end{aligned} \tag{75}$$

where the last inequality is from Lemma 19. Recalling that $m_n = O(n^{1/4-\rho})$

$$\begin{aligned}
H(V_{n-m_n+1}^n | V_1^{n-m_n}) &\geq m_n - \frac{4m_n^2}{\sqrt{\pi(n-m_n)}} \\
&= m_n - o(1).
\end{aligned} \tag{76}$$

From symmetry, $H(V_1^{m_n} | V_{m_n}^{n-m_n})$ can be evaluated to the exact same expression, and this leads to the required result. \blacksquare

Proof of Lemma 9: Let

$$r_k := \frac{(n-k+1)}{2} + (n-k+1)^{1/2+\rho}. \tag{77}$$

Then, for some $c, c' > 0$

$$\begin{aligned}
\mathbb{P} \left[W_H(V_1^k) \leq \frac{n}{2} - r_k \right] &= \mathbb{P} \left[\left\{ W_H(V_1^k) \leq \frac{n}{2} - r_k \right\} \cap \left\{ W_H(V_{k+1}^n) \geq r_k \right\} \right] \\
&\quad + \mathbb{P} \left[\left\{ W_H(V_1^k) \leq \frac{n}{2} - r_k \right\} \cap \left\{ W_H(V_{k+1}^n) < r_k \right\} \right] \\
&= \mathbb{P} \left[\left\{ W_H(V_1^k) \leq \frac{n}{2} - r_k \right\} \cap \left\{ W_H(V_{k+1}^n) \geq r_k \right\} \right] \\
&\leq \mathbb{P} \left[W_H(V_{k+1}^n) \geq r_k \right] \\
&\leq \frac{\sum_{l=r_k}^{n-k} \binom{n-k}{l} \cdot 2^k}{2^{n-1}} \\
&\stackrel{(a)}{\leq} \frac{n}{2^{n-k-1}} \binom{n-k}{r_k} \\
&\stackrel{(b)}{\leq} \frac{n}{2^{n-k-1}} 2^{(n-k)\mathbf{h}_b(\frac{r_k}{n-k})} \\
&\stackrel{(c)}{\leq} 2n \cdot 2^{-c'(n-k)^{2\rho}} \\
&\leq 2n \cdot 2^{-c' \cdot m_n^{2\rho}} \\
&\leq 2^{-c \cdot m_n^{2\rho}},
\end{aligned} \tag{78}$$

where (a) is since $r_k \geq \frac{n-k}{2}$, (b) is using Lemma 19, and (c) is using Taylor expansion of the binary entropy function at $\frac{1}{2}$. \blacksquare

B. Noisy case

Proof of Lemma 10: We have

$$M(W) = M(V + Z)$$

$$\begin{aligned}
&= M(\beta * \alpha) \\
&= [\beta(1 - \alpha) + (1 - \beta)\alpha] \cdot [\beta\alpha + (1 - \beta)(1 - \alpha)] \\
&= \alpha(1 - \alpha) + (1 - 2\alpha)^2 \cdot \beta(1 - \beta) \\
&= \alpha(1 - \alpha) + (1 - 2\alpha)^2 \cdot M(V).
\end{aligned} \tag{79}$$

■

Proof of Lemma 11: We will prove by induction. The relation holds (with equality) for $n = 1$ from Lemma

10. We assume that the property hold up to $n - 1$. Now,

$$\begin{aligned}
M(W^n) &= \sum_{i=1}^{n-1} M(W_i|W_1^{i-1}) + M(W_n|W_1^{n-1}) \\
&\geq \sum_{i=1}^{n-1} M(W_i|W_1^{i-1}) + M(W_n|W_1^{n-1}, Z_1^{n-1}) \\
&= \sum_{i=1}^{n-1} M(W_i|W_1^{i-1}) + M(V_n + Z_n|V_1^{n-1}, Z_1^{n-1}) \\
&\stackrel{(a)}{=} \sum_{i=1}^{n-1} M(W_i|W_1^{i-1}) + M(V_n + Z_n|V_1^{n-1}) \\
&\stackrel{(b)}{=} \sum_{i=1}^{n-1} M(W_i|W_1^{i-1}) + \alpha(1 - \alpha) + (1 - 2\alpha)^2 \cdot M(V_n|V_1^{n-1}) \\
&\stackrel{(c)}{\geq} (n - 1)\alpha(1 - \alpha) + (1 - 2\alpha)^2 \cdot M(V_1^{n-1}) + \alpha(1 - \alpha) + (1 - 2\alpha)^2 \cdot M(V_n|V_1^{n-1}) \\
&= n\alpha(1 - \alpha) + (1 - 2\alpha)^2 \cdot M(V^n),
\end{aligned} \tag{80}$$

where (a) is since $(V_n, Z_n)|V_1^{n-1}, Z_1^{n-1}$, (b) is using a conditional version of (79) (which holds since the pointwise relation holds), and (c) is using the induction assumption. Equality clearly holds when V^n is a memoryless random vector. ■

Proof of Lemma 12: The proof is quite similar to the proof of (6) in Section III. Let $\rho \in (0, 1/2)$ and $\eta \in [0, \frac{1}{2})$ be given. For any given $k \in [n - m_n]$ let us define the events

$$\begin{aligned}
\mathcal{A}_k &:= \left\{ W_H(V_1^k) \geq \frac{k-1}{2} - (n-k+1)^{1/2+\rho/3} \right\} \\
&= \left\{ W_H(V_1^k) \geq \frac{n}{2} - r_k + 1 \right\},
\end{aligned} \tag{81}$$

where $r_k := \frac{(n-k+1)}{2} + (n-k+1)^{1/2+\rho/3}$. Let us analyze $M(V_k|V_1^{k-1} = v_1^{k-1})$ for $1 \leq k \leq m_n$ when $v_1^{k-1} \in \mathcal{A}_{k-1}$. Conditioning on $v_1^{k-1} \in \mathcal{A}_{k-1}$, we have that V_k^n is a t -majority vector of length $n - k + 1 \geq n - m_n + 1$, and its threshold is less than

$$t \leq \frac{r_k}{n - k + 1} = \frac{1}{2} + \frac{1}{(n - k + 1)^{1/2-\rho/3}}. \tag{82}$$

Let $P_k := \mathbb{P}[V_k = 1|V_1^{k-1}]$. Assuming that n is sufficiently large, Lemma 6 (with $\eta < \frac{\rho}{3}$) implies that conditioned

on the event $V^{k-1} \in \mathcal{A}_k$

$$\begin{aligned} \frac{1}{2} \leq P_k &\leq \frac{1}{2} + \frac{1}{(n - m_n + 1)^{1/2 - \rho/3}} + O_\eta \left(\frac{1}{(n - m_n + 1)^{1/2 - \eta}} \right) \\ &\leq \frac{1}{2} + O_\eta \left(\frac{1}{n^{1/2 - \rho/3}} \right) \end{aligned} \quad (83)$$

for all $k \in [n - m_n]$, and n sufficiently large. Consequently,

$$\mathbb{M}(V_k | V_1^{k-1} = v_1^{k-1}) = P_k(1 - P_k) \geq \frac{1}{4} - O_\eta \left(\frac{1}{n^{1 - 2\rho/3}} \right). \quad (84)$$

As in Lemma 9 (when replacing m_n , the maximal value of k , with a maximal value of $n - m_n$), there exists $c > 0$ such that

$$\mathbb{P} \left[V^{k-1} \notin \mathcal{A}_{k-1} \right] \leq 2^{-c(n - m_n)^{2\rho/3}} \quad (85)$$

for all $k \in [m_n]$, and then

$$\begin{aligned} \sum_{k=1}^{m_n} \mathbb{M}(V_k | V_1^{k-1}) &\geq \sum_{k=1}^{m_n} \sum_{v^{k-1} \in \mathcal{A}_{k-1}} \mathbb{P} \left[V^{k-1} = v^{k-1} \right] \mathbb{M}(V_k | V_1^{k-1} = v^{k-1}) \\ &\geq \sum_{k=1}^{m_n} \left[1 - 2^{-c(n - m_n)^{2\rho/3}} \right] \left[\frac{1}{4} - O_\eta \left(\frac{1}{n^{1 - 2\rho/3}} \right) \right] \\ &\geq \frac{m_n}{4} - o_\eta(1). \end{aligned} \quad (86)$$

■

Proof of Lemma 13: Let us define the event

$$\mathcal{B}_k := \left\{ W_H(V_1^k) \geq \frac{n}{2} + 1 \right\}. \quad (87)$$

As in the proof of Lemma 7,

$$\begin{aligned} \mathbb{P} \left[V^k \in \mathcal{B}_k \right] &\geq \mathbb{P} \left[W_H(V_1^{n - m_n}) \geq \frac{n}{2} + 1 \right] \\ &\geq 1 - 2\sqrt{\frac{4}{\pi(n - m_n)}} m_n \\ &= 1 - O \left(n^{-1/4 - \rho} \right) \end{aligned} \quad (88)$$

for all $k \in \{n - m_n + 1, \dots, n\}$. Conditioned on $v_1^{k-1} \in \mathcal{B}_k$, all the suffixes v_k^n are possible in order to obtain a majority vector, and hence $\mathbb{P}[V_k = 1 | V_1^{k-1} = v_1^{k-1}] = \frac{1}{2}$. Then,

$$\begin{aligned} \sum_{k=n - m_n + 1}^n \mathbb{M}(V_k | V_1^{k-1}) &\geq \sum_{k=n - m_n + 1}^n \sum_{v_1^{k-1} \in \mathcal{B}_{k-1}} \mathbb{P} \left[V_1^{k-1} = v_1^{k-1} \right] \mathbb{M}(V_k | V_1^{k-1} = v_1^{k-1}) \\ &= \sum_{k=n - m_n + 1}^n \left[1 - O \left(n^{-1/4 - \rho} \right) \right] \cdot \frac{1}{4} \end{aligned}$$

$$\begin{aligned}
&\geq \frac{m_n}{4} - O\left(\frac{1}{n^{2\rho}}\right) \\
&\geq \frac{m_n}{4} - o(1).
\end{aligned} \tag{89}$$

■

Proof of Lemma 14: The entropy is bounded as

$$\begin{aligned}
H(Y^n | \text{Maj}(X^n) = 1) &\stackrel{(a)}{=} H(Y^n | \text{Maj}(X^n)) \\
&= H(\text{Maj}(X^n) | Y^n) + H(Y^n) - H(\text{Maj}(X^n)) \\
&= H(\text{Maj}(X^n) | Y^n) + n - 1 \\
&\stackrel{(b)}{\leq} H(\text{Maj}(X^n) | \text{Maj}(Y^n)) + n - 1 \\
&\stackrel{(c)}{\leq} h_b[\mathbb{P}(\text{Maj}(X^n) = \text{Maj}(Y^n))] + n - 1 \\
&\stackrel{(d)}{\leq} \mu(\alpha) + n - 1 + o(1),
\end{aligned} \tag{90}$$

where (a) follows from symmetry, (b) from the data processing theorem, (c) is from Fano's inequality, and (d) is from [10, Theorem 2.45]. ■

Proof of Lemma 15: The proof of is based on the Gaussian approximation of the binomial distribution using the Berry-Essen central limit theorem. For simplicity, we assume that n is odd, but the proof can be easily generalized to any n . We begin by denoting

$$a(y^n) := \mathbb{P}[\text{Maj}(X^n) = 1 | Y^n = y^n], \tag{91}$$

we then writing

$$H(\text{Maj}(X^n) | Y^n) = \mathbb{E} \{h_b[a(Y^n)]\}. \tag{92}$$

Since Y^n is the output of a uniform Bernoulli random vector X^n through a BSC with crossover probability α , then $Y^n = X^n + Z^n$ where $Z^n \sim \text{Bern}(\alpha)$. Equivalently, we also have $X^n = Y^n + Z^n$, where Y^n is a uniform Bernoulli random vector, and Z^n and Y^n are independent. We next use the Berry-Essen central limit theorem [11, Chapter XVI.5, Theorem 2] to evaluate $a(y^n)$. To this end, note that $\mathbb{E}[Z_i - \alpha] = 0$, $\mathbb{E}[(Z_i - \alpha)^2] = \alpha(1 - \alpha)$, and $\mathbb{E}[|Z_i - \alpha|^3] = \alpha(1 - \alpha) [\alpha^2 + (1 - \alpha)^2] < \infty$. Then,

$$\begin{aligned}
a(y^n) &= \mathbb{P}\left[W_H(y^n + Z^n) > \frac{n}{2}\right] \\
&= \mathbb{P}\left[\sum_{i \in [n]: y_i=0} Z_i + \sum_{i \in [n]: y_i=1} (1 - Z_i) > \frac{n}{2}\right] \\
&= \mathbb{P}\left\{\sum_{i \in [n]: y_i=0} (Z_i - \alpha) + \sum_{i \in [n]: y_i=1} (\alpha - Z_i) > (1 - 2\alpha) \left[\frac{n}{2} - W_H(y^n)\right]\right\}
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{P} \left\{ \frac{1}{\sqrt{n\alpha(1-\alpha)}} \left(\sum_{i \in [n]: y_i=0} (Z_i - \alpha) + \sum_{i \in [n]: y_i=1} (\alpha - Z_i) \right) > \frac{(1-2\alpha)}{\sqrt{n\alpha(1-\alpha)}} \cdot \left[\frac{n}{2} - W_H(y^n) \right] \right\} \\
&:= \mathbb{P} \left\{ S_n > \frac{(1-2\alpha)}{\sqrt{n\alpha(1-\alpha)}} \cdot \left[\frac{n}{2} - W_H(y^n) \right] \right\}, \tag{93}
\end{aligned}$$

where S_n was implicitly defined. Now, the Berry-Essen central limit theorem implies that for some C_α

$$\sup_{s \in \mathbb{R}} |\mathbb{P}[S_n > s] - \mathbb{P}[G > s]| \leq \frac{C_\alpha}{\sqrt{n}}, \tag{94}$$

where $G \sim \mathcal{N}(0, 1)$. Further, [12, Lemma 2.7] provides a bound on the difference in the entropy of two probability distributions, in terms of the total variation distance between them. In our case, this implies that for all n sufficiently large,

$$\sup_{s \in \mathbb{R}} |\mathbf{h}_b(\mathbb{P}[S_n > s]) - \mathbf{h}_b(\mathbb{P}[G > s])| \leq -\frac{2C_\alpha}{\sqrt{n}} \ln \left(\frac{C_\alpha}{\sqrt{n}} \right) = o(1). \tag{95}$$

Then, denoting

$$H_n := \frac{(1-2\alpha)}{\sqrt{n\alpha(1-\alpha)}} \cdot \left[\frac{n}{2} - W_H(y^n) \right] \tag{96}$$

we have

$$\begin{aligned}
H(\text{Maj}(X^n)|Y^n) &= \mathbb{E} \{ \mathbf{h}_b[a(Y^n)] \} \\
&= \mathbb{E} \{ \mathbf{h}_b(\mathbb{P}[S_n > H_n]) \} \\
&= \mathbb{E} \{ \mathbf{h}_b(\mathbb{P}[G > H_n]) \} + o(1) \\
&= \mathbb{E} \{ \mathbf{h}_b[Q(|H_n|)] \} + o(1) \tag{97}
\end{aligned}$$

where $Q(\cdot)$ is the Gaussian Q-function, and in the last equality we have used the facts that $Q(t) = 1 - Q(|t|)$ for $t < 0$, and $\mathbf{h}_b(p) = \mathbf{h}_b(1-p)$. Now, applying the central limit theorem once again, we have that $H_n \Rightarrow \frac{(1-2\alpha)}{\sqrt{4\alpha(1-\alpha)}} \cdot G$, as $n \rightarrow \infty$, in distribution. To complete the proof, we note that since $\mathbf{h}_b[Q(|t|)]$ is a bounded and continuous function of t , Portmanteau's lemma (e.g. [11, Chapter VIII.1, Theorem 1]) implies that

$$\mathbb{E} \{ \mathbf{h}_b[Q(|H_n|)] \} \rightarrow \mathbb{E} \left\{ \mathbf{h}_b \left[Q \left(\frac{|(1-2\alpha)G|}{\sqrt{4\alpha(1-\alpha)}} \right) \right] \right\}, \tag{98}$$

as $n \rightarrow \infty$, concluding the proof. ■

Proof of Lemma 17: Let us denote $\alpha = \frac{1}{2} - \gamma$ for $\gamma \in (0, \frac{1}{2})$, and then let us inspect

$$\mathbb{E} \{ \mathbf{h}_b[Q(\Gamma)] \} := \mathbb{E} \left\{ \mathbf{h}_b \left[Q \left(\frac{|G|\gamma}{\sqrt{(\frac{1}{2} - \gamma)(\frac{1}{2} + \gamma)}} \right) \right] \right\} \tag{99}$$

as $\gamma \downarrow 0$. Using Leibniz's integral rule, we obtain $Q'(t) = -\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, $Q''(t) = \frac{t}{\sqrt{2\pi}} \cdot e^{-t^2/2}$ and so, there exists

$\bar{c} > 0$ such that for all $t \geq 0$

$$Q(t) \geq \frac{1}{2} - \frac{t}{\sqrt{2\pi}}. \quad (100)$$

Similarly, there exists $\tilde{c}, s_1 > 0$ such that for all $s \in (0, s_1)$

$$h_b\left(\frac{1}{2} - s\right) \geq 1 - \frac{2}{\ln 2}s^2 - \tilde{c}s^4. \quad (101)$$

Hence, for all sufficiently small $t > 0$

$$\begin{aligned} h_b[Q(t)] &= h_b\left[\frac{1}{2} - \left(\frac{1}{2} - Q(t)\right)\right] \\ &\geq 1 - \frac{2}{\ln 2}\left(\frac{1}{2} - Q(t)\right)^2 - \tilde{c}\left(\frac{1}{2} - Q(t)\right)^4 \\ &\geq 1 - \frac{1}{\pi \cdot \ln 2}t^2 - \frac{\tilde{c}}{4\pi^2}t^4. \end{aligned} \quad (102)$$

So, there exists $\hat{c} > 0$ such that for all sufficiently small γ ,

$$\begin{aligned} &\mathbb{E}\{h_b[Q(\Gamma)]\} \\ &\geq \mathbb{P}[|G| \leq \gamma^{-1+\rho}] \cdot \mathbb{E}\{h_b[Q(\Gamma)] \mid |G| \leq \gamma^{-1+\rho}\} \\ &\geq \mathbb{P}[|G| \leq \gamma^{-1+\rho}] \cdot \mathbb{E}\left\{1 - \frac{1}{\pi \cdot \ln 2}\Gamma^2 - \frac{\tilde{c}}{4\pi^2}\Gamma^4 \mid |G| \leq \gamma^{-1+\rho}\right\} \\ &= \int_{-\gamma^{-1+\rho}}^{\gamma^{-1+\rho}} \frac{1}{\sqrt{2\pi}}e^{-t^2/2} \cdot \left[1 - \frac{1}{\pi \cdot \ln 2}\left(\frac{\gamma^2 t^2}{(\frac{1}{2} - \gamma)(\frac{1}{2} + \gamma)}\right) - \frac{\tilde{c}}{4\pi^2}\left(\frac{\gamma^4 t^4}{(\frac{1}{2} - \gamma)^2(\frac{1}{2} + \gamma)^2}\right)\right] \cdot dt \\ &= 1 - 2Q(\gamma^{-1+\rho}) - \int_{-\gamma^{-1+\rho}}^{\gamma^{-1+\rho}} \frac{1}{\sqrt{2\pi}}e^{-t^2/2} \cdot \left[\frac{1}{\pi \cdot \ln 2}\left(\frac{\gamma^2 t^2}{(\frac{1}{2} - \gamma)(\frac{1}{2} + \gamma)}\right) + \frac{\tilde{c}}{4\pi^2}\left(\frac{\gamma^4 t^4}{(\frac{1}{2} - \gamma)^2(\frac{1}{2} + \gamma)^2}\right)\right] \cdot dt \\ &\geq 1 - 2Q(\gamma^{-1+\rho}) - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-t^2/2} \cdot \left[\frac{1}{\pi \cdot \ln 2}\left(\frac{\gamma^2 t^2}{(\frac{1}{2} - \gamma)(\frac{1}{2} + \gamma)}\right) + \frac{\tilde{c}}{4\pi^2}\left(\frac{\gamma^4 t^4}{(\frac{1}{2} - \gamma)^2(\frac{1}{2} + \gamma)^2}\right)\right] \cdot dt \\ &= 1 - 2Q(\gamma^{-1+\rho}) - \frac{1}{\pi \cdot \ln 2}\left(\frac{\gamma^2}{(\frac{1}{2} - \gamma)(\frac{1}{2} + \gamma)}\right) - \frac{\tilde{c}}{4\pi^2}\left(\frac{3\gamma^4}{(\frac{1}{2} - \gamma)^2(\frac{1}{2} + \gamma)^2}\right) \\ &\stackrel{(a)}{\geq} 1 - \frac{1}{\pi \cdot \ln 2}\left(\frac{\gamma^2}{(\frac{1}{2} - \gamma)(\frac{1}{2} + \gamma)}\right) - \hat{c}\gamma^4, \end{aligned} \quad (103)$$

where (a) is since for any $\rho \in (0, 1)$, using $Q(t) \leq \frac{1}{t} \cdot e^{-t^2/2}$ we have

$$\mathbb{P}[|G| \geq \gamma^{-1+\rho}] = 2Q(\gamma^{-1+\rho}) \leq 2\gamma^{1-\rho} \cdot \exp\left(-\frac{1}{2\gamma^{2-2\rho}}\right). \quad (104)$$

■

APPENDIX B

USEFUL RESULTS

Lemma 19 ([7, Lemma 17.5.1]). *For $0 < \alpha < 1$ such that $n\alpha$ is integer*

$$\frac{2^{nh_b(\alpha)}}{\sqrt{8n\alpha(1-\alpha)}} \leq \binom{n}{n\alpha} \leq \frac{2^{nh_b(\alpha)}}{\sqrt{\pi n\alpha(1-\alpha)}}. \quad (105)$$

Lemma 20 ([13, Lemma 1]). *If $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ are all non-negative numbers, then*

$$\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \leq \max_{1 \leq i \leq n} \frac{a_i}{b_i}. \quad (106)$$

Corollary 21. *Under the conditions above and for any integer $l > 0$,*

$$\frac{\sum_{i=1}^{n-l} a_i}{\sum_{i=1}^n b_i} \leq \max_{1 \leq i \leq n-l} \frac{a_i}{b_i}. \quad (107)$$

This can be obtained by replacing a_i with 0 for $n-l+1 \leq i \leq n$.

REFERENCES

- [1] T. A. Courtade and G. R. Kumar, “Which boolean functions maximize mutual information on noisy inputs?” *Information Theory, IEEE Transactions on*, vol. 60, no. 8, pp. 4515–4525, 2014.
- [2] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, “On hypercontractivity and the mutual information between boolean functions,” in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, October 2013, pp. 13–19.
- [3] V. Chandar and A. Tchamkerten, “Most informative quantization functions,” Tech. Rep., 2014, available online: <http://perso.telecom-paristech.fr/~tchamker/CTAT.pdf>.
- [4] O. Ordentlich, O. Shayevitz, and O. Weinstein, “An improved upper bound for the most informative boolean function conjecture,” May 2015, available online: <http://arxiv.org/pdf/1505.05794v2.pdf>.
- [5] A. Samorodnitsky, “On the entropy of a noisy function,” November 2015, available online: <http://arxiv.org/pdf/1508.01464v4.pdf>.
- [6] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.
- [7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.
- [8] O. Ordentlich, Private Communication.
- [9] N. Alon, “On the density of sets of vectors,” *Discrete Mathematics*, vol. 46, no. 2, pp. 199–202, 1983.
- [10] R. O’Donnell, *Analysis of boolean functions*. Cambridge University Press, 2014.
- [11] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: John Wiley & Sons, 1971, vol. 2.
- [12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
- [13] T. M. Cover and E. Ordentlich, “Universal portfolios with side information,” *Information Theory, IEEE Transactions on*, vol. 42, no. 2, pp. 348–363, march 1996.