Denoising as well as the best of any two denoisers

Erik Ordentlich
Hewlett–Packard Laboratories
Palo Alto, CA 94304
erik.ordentlich@hp.com

***Abstract*—Given two arbitrary sequences of denoisers for block lengths tending to infinity we ask if it is possible to construct a third sequence of denoisers with an asymptotically vanishing (in block length) excess expected loss relative to the best expected loss of the two given denoisers for all clean channel input sequences. As in the setting of DUDE [1], which solves this problem when the given denoisers are sliding block denoisers, the construction is allowed to depend on the two given denoisers and the channel transition probabilities. We show that under certain restrictions on the two given denoisers the problem can be solved using a straightforward application of a known loss estimation paradigm. We then show by way of a counter-example that the loss estimation approach fails in the general case. Finally, we show that for the binary symmetric channel, combining the loss estimation with a randomization step leads to a solution to the stated problem under *no* restrictions on the given denoisers.**

I. Problem statement

Given alphabets $\mathcal{X}$ and $\mathcal{Z}$, an *n-block denoiser* is a mapping $\hat{X} : \mathcal{Z}^n \to \mathcal{X}^n$. For any $z^n \in \mathcal{Z}^n$, let $\hat{X}(z^n)[i]$ denote the i-th term of the sequence $\hat{X}(z^n)$. Fixing a per symbol loss function $\Lambda(\cdot,\cdot)$, for a noiseless input sequence x^n and the observed output sequence z^n, the *normalized cumulative loss* $L_{\hat{X}}(x^n, z^n)$ of the denoiser $\hat{X}$ is

$$L_{\hat{X}}(x^n, z^n) = \frac{1}{n}\sum_{i=1}^{n} \Lambda\Big(x_i, \hat{X}(z^n)[i]\Big).$$

Given a discrete memoryless channel (DMC) with transition probability matrix Π between $\mathcal{X}^n$ and $\mathcal{Z}^n$ (i.e., the setting of DUDE [1]) and two sequences of denoisers $\hat{X}_1 : \mathcal{Z}^n \to \mathcal{X}^n$ and $\hat{X}_2 : \mathcal{Z}^n \to \mathcal{X}^n$, we ask if there always exists a sequence of denoisers $\hat{X}_U$ whose expected losses $L_{\hat{X}_U}$ satisfy

$$\limsup_{n\to\infty} \max_{x^n} E(L_{\hat{X}_U}(x^n, Z^n)) - \min\{E(L_{\hat{X}_1}(x^n, Z^n)), E(L_{\hat{X}_2}(x^n, Z^n))\} = 0. \quad (1)$$

Such a denoiser $\hat{X}_U$ would then perform, in an expected sense and asymptotically, as well as the best of $\hat{X}_1$ and $\hat{X}_2$ for any channel input sequence(s).

The analogous problem in the settings of prediction [2], noisy prediction [3], and filtering (i.e., causal denoising) [4] has been solved. DUDE [1] is a solution when the two denoisers are sliding window denoisers (each denoised symbol is a function of a window of noisy symbols centered at the corresponding noisy symbol). We are not aware of a solution to the problem at the stated level generality. In the sequel, we analyze the successes and limitations of the loss estimator approach developed in [4] for filtering and extended to the denoising setting in [5], in the context of the above problem. We show that while a direct application of this approach fails in general, a certain randomized version of the approach does, in fact, solve the above problem for the case of the binary symmetric channel (BSC) (though, for now, not in a computationally practical way). The approach should be applicable to other DMCs, as will be addressed in future work.

II. Implications for error correction

In a channel coding setting, we can set the two target sequences of denoisers to the decoders of *any* two sequences of channel codes with vanishing maximal error probability. A denoiser with the universality property (1) acts like a super-decoder that when applied to the decoding of the union of the two sequences of codebooks achieves asymptotically vanishing *bit-error rate* with respect to the transmitted codeword. It would be interesting to know if such a super-decoder can be constructed without relying on randomization, as we do herein.

III. Loss estimator based approach

A loss estimator for a denoiser $\hat{X}$ is a mapping $\hat{L}_{\hat{X}} : \mathcal{Z}^n \to \mathbb{R}$ that, given a noisy sequence z^n, estimates the loss $L_{\hat{X}}(x^n, z^n)$ incurred by $\hat{X}$ to be $\hat{L}_{\hat{X}}(z^n)$.

Given a loss estimator, let $\hat{j}^*(z^n)$ denote the index $j \in \{1, 2\}$ of the denoiser $\hat{X}_j$ attaining the smallest estimated loss. That is $\hat{j}^*(z^n) = \arg\min_{j\in\{1,2\}} \hat{L}_{\hat{X}_j}(z^n)$. Consider the loss estimator based denoiser $\hat{X}^n_U(z^n) = \hat{X}_{\hat{j}^*(z^n)}(z^n)$.

Lemma 1: If for all $\epsilon > 0$, $\hat{L}_{\hat{X}_j}$ satisfies

$$\limsup_{n\to\infty} \max_{x^n} \max_{j\in\{1,2\}} Pr(|\hat{L}_{\hat{X}_j}(Z^n) - L_{\hat{X}_j}(x^n, Z^n)| \geq \epsilon) = 0 \quad (2)$$

then $\hat{X}_U$ satisfies (1).

The proof of the lemma is similar to that of Lemma 4 below, so we omit it. Lemma 1 suggests that one solution to the problem of asymptotically tracking the best of two denoisers is to estimate the loss of each denoiser from the noisy sequence and denoise using the one minimizing the estimated loss. This would work provided the loss estimator could be shown to satisfy (2).

The following is one potential estimator, first proposed in [5]. The estimate of the loss incurred by *any* denoiser $\hat{X}$ proposed in [5] is given by

$$\hat{L}_{\hat{X}}(z^n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathcal{X}} h(x, z_i) \sum_{z\in\mathcal{Z}} \Lambda(x, \hat{x}_i(z))\Pi(x, z) \quad (3)$$

where we use $\hat{x}_i(z)$ to abbreviate $\hat{X}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i]$ and $h(\cdot,\cdot)$ satisfies $\sum_z \Pi(x,z)h(x',z) = 1(x = x')$.

Example 1: For a DMC with invertible Π, $h(x,z) = \Pi^{-T}(x,z)$, uniquely.

Example 2: Specializing the previous example to a BSC with crossover probability δ,

$$h(x,z) = \begin{cases} \frac{\overline{\delta}}{1-2\delta} & (x,z) \in \{(0,0),(1,1)\} \\ \frac{-\delta}{1-2\delta} & (x,z) \in \{(0,1),(1,0)\} \end{cases}$$

where $\overline{x}$ defaults to $1-x$.

Example 3: For the binary erasure channel, h with the above property is not unique. Consider a symmetric binary erasure channel with erasure probability $1/2$. One example of a valid h is:

$$h(x,z) = 2 \cdot 1(x = z). \tag{4}$$

In this case, the estimator (3) assumes an especially intuitive form: for each unerased symbol, determine what the denoiser would have denoised that symbol to if had been erased, average the total losses over all symbols. Formally,

$$\hat{L}_{\hat{X}}(z^n) = \frac{1}{n} \sum_{i: z_i \neq e} \Lambda(z_i, \hat{x}_i(e)) \tag{5}$$

and note that $z_i = x_i$ for $i : z_i \neq e$.

Conditional unbiasedness. The loss estimator (3) has been shown to be conditionally unbiased in the following sense. Let

$$\tilde{\Lambda}_{i,\hat{X}}(z^n) \triangleq \sum_{x \in \mathcal{X}} h(x, z_i) \sum_{z \in \mathcal{Z}} \Lambda(x, \hat{x}_i(z)) \Pi(x,z)$$

denote the estimate of the loss incurred on the i-th symbol. Then $\hat{L}_{\hat{X}}(z^n) = \frac{1}{n}\sum_{i=1}^n \tilde{\Lambda}_{i,\hat{X}}(z^n)$.

Lemma 2: [5], [6], [7] For all x^n, all denoisers $\hat{X}$, and all i, $1 \leq i \leq n$, z_1^{i-1}, z_{i+1}^n

$$\begin{aligned} &E\Big[\tilde{\Lambda}_{i,\hat{X}}(Z^n) \,\big|Z_1^{i-1} = z_1^{i-1}, Z_{i+1}^n = z_{i+1}^n\Big] \\ &\quad = E\Big[\Lambda\Big(x_i, \hat{X}(Z^n)[i]\Big) \,\big|Z_1^{i-1} = z_1^{i-1}, Z_{i+1}^n = z_{i+1}^n\Big] \end{aligned} \tag{6}$$

and therefore $E\Big[\hat{L}_{\hat{X}}(z^n)\Big] = E\big[L_{\hat{X}}(x^n, Z^n)\big]$.

IV. Success stories

In this section, we review some special cases for which the loss estimator (3) exhibits the concentration property (2) and hence for which the loss estimation paradigm solves the universal denoising problem.

Example 4: Causal denoisers. In this case, $\hat{X}(z^n)[i]$ is a function of only $z_1, \ldots, z_i$. It follows that $\Delta_i(z^n) = \tilde{\Lambda}_{i,\hat{X}}(z^n) - \Lambda(x_i, \hat{X}(z^n)[i])$ is also causal. As shown in [4], causality combined with the conditional unbiasedness property (6) can be shown to imply that the Doob martingale of $\hat{L}_{\hat{X}}(Z^n) - L(x^n, \hat{X}(Z^n))$ with respect to $Z_1, Z_2, \ldots, Z_i$ has bounded differences (which, in fact are the $\Delta_i(\cdot)$). Thus, by Azuma's inequality [8], the required concentration property (2) holds with exponential decay. The Doob martingale differences can be similarly bounded also for non-causal denoisers with a bounded or slowly growing lookahead and thus concentration follows similarly for such cases.

Example 5: Each noisy sample affects only a few denoised values. If in a non-causal denoiser, the number of denoised values affected by each noisy sample is $o(\sqrt{n})$, then the Doob martingale differences are bounded by McDiarmid's condition [8], resulting in exponentially decaying concentration.

Example 6: Each denoised value depends only on a few noisy samples. A similar conclusion to the previous example can be shown (though less immediately) to hold assuming that for all i, $\hat{X}(z^n)[i]$ depends on only $c_n = o(\sqrt{n})$ of the z_j.

The following proposition improves on this last example in terms of expanding the number of noisy variables each denoising function can depend on, but at the expense of non-exponential concentration.

Proposition 1: Suppose for each i, $\hat{X}(z^n)[i]$ is a function of only (but any) $o(n)$ of the z^n. Then for any clean sequence x^n

$$\max_{x^n} E([\hat{L}_{\hat{X}}(Z^n) - L_{\hat{X}}(x^n, Z^n)]^2) = o(1) \tag{7}$$

where the expectation is with respect to the noise.

Remark Note that, via an application of Chebyshev's inequality, this proposition implies (2).

Proof: Let $\Delta_i = \Delta_i(z^n) = \tilde{\Lambda}_{i,\hat{X}}(z^n) - \Lambda(x_i, \hat{X}(z^n)[i])$ so that

$$\hat{L}_{\hat{X}}(z^n) - L_{\hat{X}}(x^n, z^n) = \frac{1}{n}\sum_{i=1}^n \Delta_i(z^n). \tag{8}$$

Let T_i denote the subset of indices i, such that $\hat{X}(z^n)[i]$ is a function of z_j with $j \in T_i$. We then have that Δ_i is a function of z_j with $j \in T_i' = T_i \cup \{i\}$. We then have that for i and $j \notin T_i'$,

$$\begin{aligned} E(\Delta_i(Z^n)\Delta_j(Z^n)) &= E(E(\Delta_i(Z^n)\Delta_j(Z^n)|Z^{j-1}, Z_{j+1}^n)) \\ &= E(\Delta_i(Z^n)E(\Delta_j(Z^n)|Z^{j-1}, Z_{j+1}^n)) \qquad (9) \\ &= 0 \qquad (10) \end{aligned}$$

where (9) follows from the fact that $\Delta_i(Z^n)$ is completely determined by Z^{j-1}, Z_{j+1}^n, since $j \notin T_i'$, and (10) follows from the conditional unbiasedness (6).

We then have

$$\begin{aligned} E\big(\big(\sum_{i=1}^n \Delta_i\big)^2\big) &= \sum_{i=1}^n \sum_{j=1}^n E(\Delta_i \Delta_j) \\ &= \sum_{i=1}^n \sum_{j \in T_i'} E(\Delta_i \Delta_j) \qquad (11) \\ &= O(n \max_i |T_i'|) = o(n^2), \qquad (12) \end{aligned}$$

where (11) follows from (10) and (12) follows from the assumption of the proposition. □

V. Problematic cases

The following are some problematic cases for the above loss estimator based approach.

Binary erasure channel with erasure probability 1/2. Consider the loss estimator based scheme with $h(x,z)$ as given by (4) applied to tracking the two denoisers

$$\hat{X}_1(z^n)[i] = \sum_{j=1}^{n} 1(z_j = 0) \mod 2$$
$$\hat{X}_2(z^n)[i] = 1 + \sum_{j=1}^{n} 1(z_j = 0) \mod 2 \tag{13}$$

for each i that $z_i = e$, under the Hamming loss. Thus, denoiser 1 denoises to all 0's if the number of 0's in z^n is even and to all 1's, otherwise, and denoiser 2 does precisely the opposite. Suppose the input sequence x^n is the all zero sequence. In this case (actually all cases), the expected (unnormalized) loss of each denoiser is $n/4$. It turns out, however, that the loss estimator based denoiser always makes the worst possible choice. Suppose z^n has an even number of 0's. Denoiser 1 in this case achieves 0 loss, while denoiser 2 achieves a loss of N_e (denoting the number of erasures). Following (5), the estimated unnormalized loss of denoiser 1, on the other hand, is N_0 and of denoiser 2, 0. The loss estimator based denoiser will thus elect to follow denoiser 2, incurring a loss of N_e. The loss estimator goes similarly astray for z^n with an odd number of 0's, and the average denoiser loss is thus $n/2$, failing to track the $n/4$ average performance.

Binary symmetric channel. It turns out that the above example fails to break the loss estimator based denoiser for the BSC and Hamming loss and a more complicated example is required. For the BSC with crossover probability δ, the loss estimate of denoiser $\hat{X}$ is

$$\begin{aligned} &n\hat{L}_{\hat{X}}(z^n) \\ &= \sum_{i:z_i=0} \Big[\frac{\overline{\delta}}{1-2\delta}(\delta\Lambda(0,\hat{X}(z^n\oplus\mathbf{e}_i)[i])+\overline{\delta}\Lambda(0,\hat{X}(z^n)[i]) \\ &\quad -\frac{\delta}{1-2\delta}(\delta\Lambda(1,\hat{X}(z^n)[i])+\overline{\delta}\Lambda(1,\hat{X}(z^n\oplus\mathbf{e}_i)[i])\Big] \\ &\quad + \sum_{i:z_i=1} \Big[\frac{\overline{\delta}}{1-2\delta}(\delta\Lambda(1,\hat{X}(z^n\oplus\mathbf{e}_i)[i])+\overline{\delta}\Lambda(1,\hat{X}(z^n)[i]) \\ &\quad -\frac{\delta}{1-2\delta}(\delta\Lambda(0,\hat{X}(z^n)[i])+\overline{\delta}\Lambda(0,\hat{X}(z^n\oplus\mathbf{e}_i)[i])\Big], \end{aligned} \tag{14}$$

where $\mathbf{e}_i$ denotes the "indicator" sequence, with $\mathbf{e}_i[j] = 0$ if $j \neq i$ and $\mathbf{e}_i[i] = 1$ and $\oplus$ denotes componentwise modulo two addition. We can express (14) in terms of the joint type of the three sequences $z^n, \hat{X}(z^n)$, and $\{\hat{X}(z^n \oplus \mathbf{e}_i)[i]\}_{i=1}^n$. Specifically, for $b_k \in \{0,1\}$, $k = 0,1,2$, define

$$N_{b_0b_1b_2} = |\{i : z_i = b_0, \hat{X}(z^n)[i] = b_1, \hat{X}(z^n \oplus \mathbf{e}_i)[i] = b_2\}|,$$
$$N_{b_0b_1} = \sum_{b_2} N_{b_0b_1b_2}, \text{ and } N_{b_0} = \sum_{b_1} N_{b_0b_1}.$$

After some simplification, we can then express (14) as

$$\begin{aligned} n\hat{L}_{\hat{X}}(z^n) = &-\frac{\delta}{1-2\delta}(N_{000}+N_{111}) + \delta(N_{001}+N_{110}) \\ &+ \overline{\delta}(N_{010}+N_{101}) + \frac{\overline{\delta}}{1-2\delta}(N_{011}+N_{100}) \end{aligned} \tag{15}$$

For our example, we will set $\hat{X}_1(z^n)[i] = \hat{X}_2(z^n)[i] = 0$ for all i and z^n with even parity. Thus, for even parity, the two denoisers will be identical, resulting in identical losses for any clean sequence. For z^n with odd parity, this implies that the corresponding $N_{b_0b_1b_2} = 0$ for $b_2 = 1$ so that $N_{b_0b_1} = N_{b_0b_10}$. We will next assume that the clean sequence is the all 0 sequence and specify the behavior of the two denoisers for odd parity z^n taking this into account. Under this assumption on the clean sequence, with probability tending to 1, $N_1 = \delta n + o(n)$ and $N_0 = \overline{\delta}n + o(n)$, so that for odd parity z^n, with probability tending to 1, we can write

$$N_{10} = n\delta - N_{11} + o(n) \text{ and } N_{00} = n\overline{\delta} - N_{01} + o(n).$$

Using the above, we can further simplify (15) to

$$\begin{aligned} n\hat{L}_{\hat{X}}(z^n) &= -\frac{\delta}{1-2\delta}N_{00} + \delta N_{11} + \overline{\delta}N_{01} + \frac{\overline{\delta}}{1-2\delta}N_{10} \\ &= N_{01}\left(\frac{\delta}{1-2\delta}+\overline{\delta}\right) + N_{11}\left(\delta - \frac{\overline{\delta}}{1-2\delta}\right) + o(n) \\ &= (N_{01}-N_{11})\left(\frac{\delta}{1-2\delta}+\overline{\delta}\right) + o(n). \end{aligned} \tag{16}$$

The two denoisers will then, respectively, denoise z^n with odd parity so that:

$$\hat{X}_1(z^n) \to N_{01} = 0, N_{11} = N_1$$
$$\hat{X}_2(z^n) \to N_{01} = \lfloor \delta N_0 \rfloor, N_{11} = 0.$$

Thus, denoiser 1, for z^n with odd parity, sets $\hat{X}_1(z^n)[i] = z_i$, while denoiser 2 sets $\hat{X}_2(z^n)[i] = 0$ if $z_i = 1$ and $\hat{X}_2(z^n)[i] = 1$ for an arbitrary fraction δ of those i for which $z_i = 0$. Under the assumption that x^n is all 0, the following summarizes the actual losses and estimated losses for z^n with odd parity and $N_1 = n\delta + o(n)$:

Denoiser	$nL_{\hat{X}}(x^n, z^n)$	$n\hat{L}_{\hat{X}}(z^n)$
1	$\delta n + o(n)$	$-\left(\frac{\delta}{1-2\delta}+\overline{\delta}\right)\delta n + o(n)$
2	$\delta\overline{\delta}n + o(n)$	$\left(\frac{\delta}{1-2\delta}+\overline{\delta}\right)\delta\overline{\delta}n + o(n)$

Thus, we see that the estimated loss for denoiser 1 is smaller (negative in fact) while its actual loss is larger. Since the above scenario (odd parity z^n and $N_1 = \delta n + o(n)$) occurs roughly with probability 1/2, and for z^n with even parity the two denoisers both incur zero loss, it follows that the expected loss of the loss estimator based denoiser fails to track the expected loss of the best denoiser, namely denoiser 2, in this case.

VI. Smoothed denoisers

The misbehavior of the loss estimator in the previous section appears to be the result of an excessive sensitivity of the target denoisers to the noisy sequence. Our path forward for the BSC is to first "smooth" the target denoisers via a randomization procedure in a way that does not significanly alter their average case performance on any sequence. The expected performance (with respect to the randomization) of the smoothed denoisers, in turn, will be shown to be more amenable to accurate loss estimation. To this end, for the BSC-δ case, let W^n be i.i.d. Bernoulli-q_n for some q_n vanishing (with n). Given a denoiser $\hat{X}$, the randomized (smoothed) version is taken to be

$$\hat{X}'(z^n) = \hat{X}(z^n \oplus W^n). \tag{17}$$

Conditioned on $Z^n = z^n$, the expected loss (with respect to W^n) of this randomized denoiser is

$$\overline{L}_{\hat{X}'}(x^n, z^n) \stackrel{\triangle}{=} \frac{1}{n}\sum_i E_{W^n}\Lambda(x_i, \hat{X}(z^n \oplus W^n)[i]). \quad (18)$$

We can readily adapt the above loss estimator to estimate $\overline{L}_{\hat{X}'}(x^n, z^n)$ as

$$\hat{\overline{L}}_{\hat{X}'}(z^n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathcal{X}} h(x, z_i)$$
$$\times \sum_{z\in\mathcal{Z}} E_{W^n}\Lambda(x, \hat{X}((z^{i-1}, z, z_{i+1}^n) \oplus W^n))[i]\Pi(x, z). \quad (19)$$

The summands (over i) of this estimate of the expected loss of the randomized denoiser also have a conditional unbiasedness property. Specifically, letting

$$\tilde{\overline{\Lambda}}_{i,\hat{X}'}(z^n) \stackrel{\triangle}{=} \sum_{x\in\mathcal{X}} h(x, z_i)$$
$$\times \sum_{z\in\mathcal{Z}} E_{W^n}\Lambda(x, \hat{X}((z^{i-1}, z, z_{i+1}^n) \oplus W^n))[i]\Pi(x, z), \quad (20)$$

we have

$$E\Big[\tilde{\overline{\Lambda}}_{i,\hat{X}'}(Z^n)\,\big|Z_1^{i-1} = z_1^{i-1}, Z_{i+1}^n = z_{i+1}^n\Big]$$
$$= E\Big[E_{W^n}\Lambda\Big(x_i, \hat{X}(Z^n \oplus W^n)[i]\Big)\,\big|Z_1^{i-1}=z_1^{i-1}, Z_{i+1}^n=z_{i+1}^n\Big] \quad (21)$$

We can then prove (see below) the following key lemma.

Lemma 3: For all δ and $\hat{X}'$ as in (17) with $q_n = n^{-\nu}$ and $0 < \nu < 1$,

$$\max_{x^n} E\left(\overline{L}_{\hat{X}'}(x^n, Z^n) - \hat{\overline{L}}_{\hat{X}'}(Z^n)\right)^2 = o(1) \quad (22)$$

with $\overline{L}_{\hat{X}'}$ and $\hat{\overline{L}}_{\hat{X}'}$ as in (18) and (19) and where the expectation is with respect to the BSC-δ induced Z^n.

The lemma implies that for any BSC the estimate (19) of the randomized denoiser conditional expected loss concentrates for all clean sequences and all underlying denoisers, including those in which the estimate of the underlying denoiser loss does not. This motivates an estimation minimizing randomized denoiser which departs from the approach of Section III as follows. Given denoisers $\hat{X}_1$ and $\hat{X}_2$ let $\hat{X}_1'$ and $\hat{X}_2'$ denote their respective randomized versions according to the above randomization. Next, define $\hat{j}'^*(z^n)$ to be

$$\hat{j}'^*(z^n) = \arg\min_{j\in\{1,2\}} \hat{\overline{L}}_{\hat{X}_j'}(z^n)$$

with $\hat{\overline{L}}_{\hat{X}_j'}$ in (19) above. The estimation minimizing randomized denoiser is then defined as

$$\hat{X}_{RU}^n(z^n) = \hat{X}'_{\hat{j}'^*(z^n)}(z^n) = \hat{X}_{\hat{j}'^*(z^n)}(z^n \oplus W^n). \quad (23)$$

This denoiser thus determines the denoiser whose randomized version yields the smallest estimated expected loss computed according to (19) and denoises using the randomized version of the selected denoiser. We then have the following.

Lemma 4: If for all $\epsilon > 0$, $\hat{\overline{L}}_{\hat{X}_j'}$ satisfies

$$\limsup_{n\to\infty} \max_{x^n} \max_{j\in\{1,2\}} Pr(|\hat{\overline{L}}_{\hat{X}_j'}(Z^n) - \overline{L}_{\hat{X}_j'}(x^n, Z^n)| \geq \epsilon) = 0 \quad (24)$$

then $\hat{X}_{RU}$ satisfies

$$\limsup_{n\to\infty} \max_{x^n} E(L_{\hat{X}_{RU}}(x^n, Z^n))$$
$$- \min\{E(L_{\hat{X}_1'}(x^n, Z^n)), E(L_{\hat{X}_2'}(x^n, Z^n))\} = 0, \quad (25)$$

where the expectations are with respect to the channel output Z^n *and* the randomization W^n.

Proof. Let j'^* denote

$$j'^*(x^n, z^n) = \arg\min_{j\in\{1,2\}} \overline{L}_{\hat{X}_j'}(x^n, z^n).$$

Suppose for x^n and z^n, $|\hat{\overline{L}}_{\hat{X}_j'}(z^n) - \overline{L}_{\hat{X}_j'}(x^n, z^n)| \leq \epsilon$ for $j \in \{1, 2\}$. We then have

$$\overline{L}_{\hat{X}'_{\hat{j}'^*}}(x^n, z^n) - \overline{L}_{\hat{X}'_{j'^*}}(x^n, z^n)$$
$$= \overline{L}_{\hat{X}'_{\hat{j}'^*}}(x^n, z^n) - \hat{\overline{L}}_{\hat{X}'_{\hat{j}'^*}}(z^n) + \hat{\overline{L}}_{\hat{X}'_{\hat{j}'^*}}(z^n) - \overline{L}_{\hat{X}'_{j'^*}}(x^n, z^n)$$
$$\leq \overline{L}_{\hat{X}'_{\hat{j}'^*}}(x^n, z^n) - \hat{\overline{L}}_{\hat{X}'_{\hat{j}'^*}}(z^n) + \hat{\overline{L}}_{\hat{X}'_{j'^*}}(z^n) - \overline{L}_{\hat{X}'_{j'^*}}(x^n, z^n)$$
$$\leq 2\epsilon,$$

implying, via a union bound, that

$$Pr(\overline{L}_{\hat{X}'_{\hat{j}'^*}}(x^n, Z^n) - \overline{L}_{\hat{X}'_{j'^*}}(x^n, Z^n) \geq 2\epsilon)$$
$$\leq \sum_{j=1}^{2} Pr(|\hat{\overline{L}}_{\hat{X}_j'}(Z^n) - \overline{L}_{\hat{X}_j'}(x^n, Z^n)| \geq \epsilon). \quad (26)$$

Noting that $\hat{X}'_{\hat{j}'^*} = \hat{X}_{RU}$, it follows that, for all $\epsilon > 0$,

$$\max_{x^n} E(L_{\hat{X}_{RU}}(x^n, Z^n))$$
$$- \min\{E(L_{\hat{X}_1'}(x^n, Z^n)), E(L_{\hat{X}_2'}(x^n, Z^n))\}$$
$$\leq \max_{x^n} E(\overline{L}_{\hat{X}'_{\hat{j}'^*}}(x^n, Z^n) - \overline{L}_{\hat{X}_{j'^*}}(x^n, Z^n)) \quad (27)$$
$$\leq 2\epsilon + \Lambda_{\max} \max_{x^n} Pr(\overline{L}_{\hat{X}'_{\hat{j}'^*}}(x^n, Z^n) - \overline{L}_{\hat{X}_{j'^*}}(x^n, Z^n) \geq 2\epsilon) \quad (28)$$

where $\Lambda_{\max}$ denotes the maximum loss and (27) follows from $E(L_{\hat{X}_{RU}}(x^n, Z^n)) = E(\overline{L}_{\hat{X}'_{\hat{j}'^*}}(x^n, Z^n))$ and

$$\min\{E(L_{\hat{X}_1'}(x^n, Z^n)), E(L_{\hat{X}_2'}(x^n, Z^n))\}$$
$$= \min\{E(\overline{L}_{\hat{X}_1'}(x^n, Z^n)), E(\overline{L}_{\hat{X}_2'}(x^n, Z^n))\}$$
$$\geq E(\min\{\overline{L}_{\hat{X}_1'}(x^n, Z^n), \overline{L}_{\hat{X}_2'}(x^n, Z^n)\}) = E(\overline{L}_{\hat{X}'_{j'^*}}(x^n, Z^n)).$$

The lemma now follows from (26), (24), and the fact that (28) holds for all $\epsilon > 0$. $\square$

This lemma shows that the loss estimation minimizing randomized denoiser exhibits the same asymptotic expected performance as the best of two randomized denoisers, and if the expected performance of each such randomized denoiser were, in turn, close to the expected performance of the corresponding original denoiser, the estimation minimizing randomized denoiser would solve our original problem. The proof of Lemma 3 is sketched in the next section, while the latter property is contained in the following.

Lemma 5: For a BSC-δ, Hamming loss, and any denoiser $\hat{X}$, if W^n is i.i.d. Bernoulli-q_n with $q_n = n^{-\nu}$ for $\nu > 1/2$,

$$\max_{x^n} |E(L(x^n, \hat{X}(Z^n \oplus W^n))) - E(L(x^n, \hat{X}(Z^n)))| = o(1)$$

where the first expectation is with respect to the channel and the randomization.

The proof of the lemma, which is too long to include, involves showing that the L_1 distance between the distributions of the random variables Z^n and $Z^n \oplus W^n$ vanishes uniformly for all input sequences x^n. Thus, we have the following.

Theorem 6: For ν satisfying $1/2 < \nu < 1$, the loss estimation minimizing randomized denoiser $\hat{X}_{RU}$ given by (23), with W^n i.i.d. Bernoulli-$n^{-\nu}$, satisfies

$$\limsup_{n\to\infty} \max_{x^n} E(L_{\hat{X}_{RU}}(x^n, Z^n)) - \min\{E(L_{\hat{X}_1}(x^n, Z^n)), E(L_{\hat{X}_2}(x^n, Z^n))\} = 0.$$

VII. Proof of Lemma 3

We begin by defining, for any $f : \{0,1\}^n \to \mathbb{R}$, the x^n-dependent total influence (terminology inspired by a related quantity in [9]) of f as

$$I(f) = \sum_{j=1}^{n} E(|f(Z^n) - f(Z^{j-1}, \tilde{Z}_j, Z_{j+1}^n)|)$$

where $(\tilde{Z}^n, Z^n)$ constitute an i.i.d. pair of random variables with Z^n distributed according to the channel with input x^n (hence the dependence on x^n). The proof of Lemma 3 hinges on the following result.

Proposition 2: For all $0 < \nu < 1$ and $\hat{X}'(z^n)$ defined as in (17),

$$\max_{x^n, i} I(\overline{\hat{X}'}(\cdot)[i]) = o(n), \tag{29}$$

where $\overline{\hat{X}'}(z^n)[i] = E_{W^n}(\hat{X}'(z^n)[i]) = E_{W^n}(\hat{X}(z^n \oplus W^n)[i])$, with the expectation taken with respect to W^n.

The proof, which is too long to include, involves showing that $\max_f \max_{z^n} \sum_{j=1}^n |\overline{f}(z^n) - \overline{f}(z^n \oplus \mathbf{e}_j)| = o(n)$, where the outer maximization is over all functions $f : \{0,1\}^n \to [0,1]$, with $\overline{f}(z^n) \stackrel{\triangle}{=} E_{W^n}(f(z^n \oplus W^n))$.

Proof sketch of Lemma 3: The proof is similar to that of Proposition 1, except the correlations appearing in (9) are handled using Proposition 2. Define

$$\overline{\Delta}_i(z^n) \stackrel{\triangle}{=} \overline{\overline{\Lambda}}_{i,\hat{X}'}(z^n) - E_{W^n}\Lambda\Big(x_i, \hat{X}(z^n \oplus W^n)[i]\Big) \tag{30}$$

with $\overline{\overline{\Lambda}}_{i,\hat{X}'}$ as in (20). We claim that

$$\max_{x^n, i} I(\overline{\Delta}_i(\cdot)) = o(n). \tag{31}$$

To see this, note that for the binary/Hamming loss case, $E_{W^n}\Lambda(x, \hat{X}(z^n \oplus W^n)[i]) = \overline{\hat{X}'}(z^n)[i]$ if $x = 0$ and $1 - \overline{\hat{X}'}(z^n)[i]$ if $x = 1$. Using this, it can be shown that $I(\overline{\Delta}_i(\cdot)) \le d_1 + d_2 I(\overline{\hat{X}'}(\cdot)[i])$, for some constants d_1 and d_2. The claim (31) then follows from Proposition 2.

Next, we note that for $(Z^n, \tilde{Z}^n)$ an i.i.d. pair with Z^n distributed according to the channel (as in the definition of total influence above), for all pairs (i, j),

$$\begin{aligned}
&E(\overline{\Delta}_i(Z^{j-1}, \tilde{Z}_j, Z_{j+1}^n)\overline{\Delta}_j(Z^n)) \\
&= E\big(E\big(\overline{\Delta}_i(Z^{j-1}, \tilde{Z}_j, Z_{j+1}^n)\overline{\Delta}_j(Z^n)\big|Z^{j-1}, \tilde{Z}_j, Z_{j+1}^n\big)\big) \\
&= E\big(\overline{\Delta}_i(Z^{j-1}, \tilde{Z}_j, Z_{j+1}^n)E\big(\overline{\Delta}_j(Z^n)\big|Z^{j-1}, \tilde{Z}_j, Z_{j+1}^n\big)\big) \\
&= 0, \qquad (32)
\end{aligned}$$

where this last step follows from the conditional unbiasedness (21) and the distribution of $(Z^n, \tilde{Z}^n)$.

We then have

$$\begin{aligned}
&E\Big(\sum_{i=1}^{n} \overline{\Delta}_i(Z^n)\Big)^2 \\
&= \sum_{i=1}^{n} E\Big(\sum_{j=1}^{n} \overline{\Delta}_i(Z^n)\overline{\Delta}_j(Z^n)\Big) \\
&= \sum_{i=1}^{n} E\Big(\sum_{j=1}^{n} \overline{\Delta}_i(Z^j, \tilde{Z}_j, Z_j^n)\overline{\Delta}_j(Z^n)\Big) \\
&\quad + \sum_{i=1}^{n} E\Big(\sum_{j=1}^{n} (\overline{\Delta}_i(Z^n) - \overline{\Delta}_i(Z^j, \tilde{Z}_j, Z_j^n))\overline{\Delta}_j(Z^n)\Big) \\
&= \sum_{i=1}^{n} E\Big(\sum_{j=1}^{n} (\overline{\Delta}_i(Z^n) - \overline{\Delta}_i(Z^j, \tilde{Z}_j, Z_j^n))\overline{\Delta}_j(Z^n)\Big) \qquad (33) \\
&\le c \sum_{i=1}^{n} E\Big(\sum_{j=1}^{n} (|\overline{\Delta}_i(Z^n) - \overline{\Delta}_i(Z^j, \tilde{Z}_j, Z_j^n)|)\Big) \\
&\le c \sum_{i=1}^{n} I(\overline{\Delta}_i(\cdot)), \qquad (34)
\end{aligned}$$

where (33) follows from (32) and (34) from the fact that $\overline{\Delta}_i(z^n)$ can be bounded by a constant c for all i, z^n and x^n and the definition of total influence. The proof is completed by applying (31). □

References

[1] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger, "Universal discrete denoising: known channel," *IEEE Trans. Inform. Theory,* vol. 51, no. 1, pp. 5–28, Jan. 2005.

[2] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," *Journal of the ACM,* vol. 44, no. 3, pp. 427–485, May 1997.

[3] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inform. Theory,* vol. 47, no. 6, pp. 2151–2173, Sep. 2001.

[4] T. Weissman, E. Ordentlich, M. J. Weinberger, A. Somekh-Baruch, N. Merhav, "Universal filtering via prediction," *IEEE Trans. Inform. Theory,* vol. 53, no. 4, pp. 1253–1264, Apr. 2007.

[5] E. Ordentlich, M. J. Weinberger, and T. Weissman, "Multi-directional context sets with applications to universal denoising and compression," in *Proc. of IEEE Symp. on Info. Theory*, pp. 1270–1274, 2005.

[6] T. Moon and T. Weissman, "Discrete denoising with shifts," *IEEE Trans. Inform. Theory,* vol. 55, no. 11, pp. 5284–5301, Nov. 2009.

[7] E. Ordentlich, K. Viswanathan, M. J. Weinberger, "Twice-universal denoising," *IEEE Trans. Inform. Theory,* vol. 59, no. 1, pp. 526–545, Jan. 2013.

[8] C. McDiarmid, "Concentration," *Probabilistic Methods for Algorithmic Discrete Mathematics*, Springer, 1998.

[9] J. Kahn, G. Kalai, and N. Linial, "The influence of variables on Boolean functions,", in *Proc. 29-th IEEE Symp. on Foundations of Comp. Sci.,* pp. 68–80, 1988.