Chapter 9

Discrete Denoising

9.1 Discrete Denoising

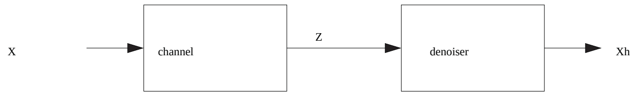

Figure 9.1: Discrete denoising system

We consider the discrete denoising system shown in Figure 9.1. The clean source $\mathbf{X} = (X_1, X_2, \ldots)$ (or $(\ldots, X_{-1}, X_0, X_1, \ldots)$ if $\mathbf{X}$ is a double-sided sequence) is corrupted by a noisy channel. The denoiser observes the output of the channel $\mathbf{Z} = (Z_1, Z_2, \ldots)$ (or double-sided $(\ldots, Z_{-1}, Z_0, Z_1, \ldots)$) and reconstructs the sequence $\hat{\mathbf{X}} = (\hat{X}_1, \hat{X}_2, \ldots)$ (or double-sided $(\ldots, \hat{X}_{-1}, \hat{X}_0, \hat{X}_1, \ldots)$). The sequences $\mathbf{X}$, $\mathbf{Z}$, and $\hat{\mathbf{X}}$ are discrete, that is, $X_i \in \mathcal{X}$, $Z_i \in \mathcal{Z}$, $\hat{X}_i \in \hat{\mathcal{X}}$, and $\mathcal{X}$, $\mathcal{Z}$, and $\hat{\mathcal{X}}$ are finite alphabets. We are interested in sequences of block length n and denote the n-block source sequence, noisy sequence, and reconstruction sequence by $X^n = (X_1, X_2, \ldots, X_n)$, $Z^n = (Z_1, Z_2, \ldots, Z_n)$, and $\hat{X}^n = (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n)$, respectively. Now we introduce some definitions.

Definition 9.1.1. *An "n-block denoiser" is a mapping* $\hat{X}^n : \mathcal{Z}^n \to \hat{\mathcal{X}}^n$.

There are some other terms used through the literature: estimation, noncausal filtering, and smoothing. Given a "loss function" or "distortion criterion" $\Lambda : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ e.g. Hamming loss, and given denoiser $\hat{X}^n(\cdot)$, and particular sequences x^n and z^n, define the per-symbol loss

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

where $\hat{X}^n(z^n)[i]$ is the i-th component of the n-tuple $\hat{X}^n(z^n)$, and it can also be denoted by $\hat{X}_i(z^n)$. If (X^n, Z^n) are jointly distributed according to some distribution, then $\mathsf{E}L_{\hat{X}^n}(X^n, Z^n)$ is our measure of performance.

9.2 Optimum Performance

We first present some notation for probability distributions. If A is a discrete random variable, then $P_A(a) = \Pr(A = a)$ is the probability mass function of A. Assume that $A \in \mathcal{A}$ and the alphabet set $\mathcal{A} = \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ is ordered. Then P_A is a column vector of dimension $|\mathcal{A}|$ with i-th component $P_A(a_i)$. Given another jointly distributed random variable B, define the conditional probability of A given B as

$$P_{A|B}(a) = \Pr(A = a|B) = \mathsf{E}\left[1_{\{A=a\}} \middle| B\right].$$

The column vector $P_{A|B}$ is thus a random simplex vector of dimension $|\mathcal{A}|$. We need the following definitions to characterize the optimum performance.

Definition 9.2.1. *The "Bayes envelope" of P_X is defined as*

$$U(P_X) = \min_{\hat{x}\in\hat{\mathcal{X}}} \sum_{x\in\mathcal{X}} P_X(x)\Lambda(x,\hat{x}) = \min_{\hat{x}\in\hat{\mathcal{X}}} \mathsf{E}\Lambda(X,\hat{x}) = \min_{\hat{x}\in\hat{\mathcal{X}}} P_X^T \lambda_{\hat{x}},$$

where $\lambda_{\hat{x}}$ is the column of the loss matrix associated with $\hat{x}$:

$$\lambda_{\hat{x}} = \begin{pmatrix} \Lambda(a_1,\hat{x}) \\ \Lambda(a_2,\hat{x}) \\ \vdots \\ \Lambda\left(a_{|\mathcal{A}|},\hat{x}\right) \end{pmatrix}.$$

The Bayes envelope is the minimum expected loss achievable in guessing the value of $X \sim P_X$. We generalize the definition to accommodate any vector v of dimension $|\mathcal{X}|$: $U(v) = \min_{\hat{x}\in\hat{\mathcal{X}}} v^t \lambda_{\hat{x}}$.

Definition 9.2.2. *The minimizer of the Bayes envelope is called the "Bayes response"*

$$\hat{X}_{Bayes}(v) = \operatorname*{argmin}_{\hat{x}\in\hat{\mathcal{X}}} v^T \lambda_{\hat{x}},$$

where ties are resolved lexicographically, that is, the symbol with the smallest index is chosen. Note that $\hat{X}_{Bayes}(v) = \hat{X}_{Bayes}(\alpha v)$ $\forall \alpha > 0$.

Exercise 9.2.3. *Show that the Bayes envelope has the following properties.*

(a) $U(\cdot)$ is concave.

(b) "Data processing inequality": If $Y = f(Z)$, then $\mathsf{E}U\left(P_{X|Z}\right) \le \mathsf{E}U\left(P_{X|Y}\right)$.

(c) Generalize (b) to $X - Z - Y$, that is, X and Y are conditionally independent given Z.

Now we can express the optimum performance in terms of the Bayes envelope.

Theorem 9.2.4. *If (X^n, Z^n) are arbitrarily jointly distributed, then*

$$\min_{\hat{X}^n \in \mathcal{D}_n} \mathsf{E} L_{\hat{X}^n}(X^n, Z^n) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E} U\left(P_{X_i|Z^n}\right),$$

where $\mathcal{D}_n$ is the set of all n-block denoisers. The minimum is achieved by $\hat{X}_i(Z^n) = \hat{X}_{Bayes}\left(P_{X_i|Z^n}\right)$.

Proof For any n-block denoiser, the per-symbol loss is equal to

$$\begin{aligned}
\mathsf{E} L_{\hat{X}^n}(X^n, Z^n) &= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\Lambda\left(X_i, \hat{X}_i(Z^n)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left[\mathsf{E}\left[\Lambda\left(X_i, \hat{X}_i(Z^n)\right)\Big| Z^n\right]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left[\sum_{x\in\mathcal{X}} P_{X_i|Z^n}(x)\Lambda(x, \hat{X}_i(Z^n))\right] \\
&\overset{(a)}{\geq} \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left[\min_{\hat{x}\in\hat{\mathcal{X}}} \sum_{x\in\mathcal{X}} P_{X_i|Z^n}(x)\Lambda(x, \hat{x})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E} U(P_{X_i|Z^n}),
\end{aligned}$$

where the last equality follows from the definition of the Bayes envelope $U(P_{X_i|Z^n})$. The equality in (a) is achieved by

$$\hat{X}_i(Z^n) = \operatorname*{argmin}_{\hat{x}\in\hat{\mathcal{X}}} \sum_{x\in\mathcal{X}} P_{X_i|Z^n}(x)\Lambda(x, \hat{x}) = \hat{X}_{\text{Bayes}}\left(P_{X_i|Z^n}\right)$$

□

9.3 Optimum Performance for Stationary Sources

If the clean source process and the noisy process are jointly stationary, we are interested in the optimum performance of n-block denoiser as $n \to \infty$.

Definition 9.3.1. *Suppose that $(\mathbf{X}, \mathbf{Z}) = \{(X_i, Z_i)\}_{i=-\infty}^{\infty}$ are jointly stationary. Define "denoisability of $\mathbf{X}$ based on $\mathbf{Z}$" as*

$$\mathbb{D}(\mathbf{X}, \mathbf{Z}) = \lim_{n\to\infty} \min_{\hat{X}^n \in \mathcal{D}_n} \mathsf{E} L_{\hat{X}^n}(X^n, Z^n) \tag{9.1}$$

The following exercises show properties of $\mathbb{D}(\mathbf{X}, \mathbf{Z})$.

Exercise 9.3.2. *Prove that the limit in* (9.1) *exists.*
Hint:

(a) A sequence $\{a_n\}_{n\geq 1}$ *is called "sub-additive" if* $a_{n+m} \leq a_n + a_m$.

(b) "Sub-additive lemma:" $\forall$ *sub-additive sequence* $\{a_n\}$, $\lim_{n\to\infty}(a_n/n)$ *exists and is equal to* $\inf_{n\geq 1}(a_n/n)$.

Exercise 9.3.3. *Prove* $\mathbb{D}(\mathbf{X},\mathbf{Z}) = \mathsf{E}U(P_{X_0|\mathbf{Z}})$.
Hint:

(a) $\mathbb{D}(\mathbf{X},\mathbf{Z}) \leq \mathsf{E}U(P_{X_0|Z_{-k}^m})$ $\forall k, m > 0$, *where*

$$Z_m^n = \begin{cases} (Z_m, Z_{m+1}, \ldots, Z_n) & \textit{if } m \leq n, \\ \emptyset & \textit{otherwise.} \end{cases}$$

Note that, by stationarity, $\mathsf{E}U(P_{X_i|Z^n}) = \mathsf{E}U(P_{X_i|Z_{-(i-1)}^{n-i}})$.

(b) $\mathbb{D}(\mathbf{X},\mathbf{Z}) \geq \mathsf{E}U(P_{X_0|\mathbf{Z}})$.

(c) $\lim\limits_{\substack{m\to\infty \\ k\to\infty}} \mathsf{E}U(P_{X_0|Z_{-k}^m}) = \mathsf{E}U(P_{X_0|\mathbf{Z}})$.

Although the optimum performance can be expressed in terms of the Bayes envelope and is achieved by the corresponding Bayes response, usually it is difficult to compute the Bayes response $\hat{X}_{\text{Bayes}}(P_{X_i|\mathbf{Z}})$ given the source distribution P_{X^n} and the noisy channel $P_{Z^n|X^n}$. In addition, the source distribution P_{X^n} is unknown in practice.

Table 9.1: Qualitative comparison of compression and denoising

Compression	Denoising
$\min \mathsf{E}\left[\frac{1}{n}l_n(X^n)\right]$ [1]	$\min\limits_{\hat{X}^n\in\mathcal{D}_n} \mathsf{E}L_{\hat{X}^n}(X^n, Z^n)$
$\big($Achiever of min: Huffman Code tailored to P_{X^n} $\big)$	$\big($Achiever of min: $\hat{X}_i(Z^n) = \hat{X}_{\text{Bayes}}(P_{X_i\|Z^n})\big)$
$= \frac{1}{n}H(X^n) + o(n)$	
$= \frac{1}{n}\sum_{i=1}^{n} H(X_i\|X^{i-1})$	$= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}U(P_{X_i\|Z^n})$
$\overset{n\to\infty}{\longrightarrow} H(X_0\|X_{-\infty}^{-1})$ [2]	
$= \mathsf{E}\left[\frac{1}{P_{X_0\|X_{-\infty}^{-1}}(X_0)}\right]$	$\overset{n\to\infty}{\longrightarrow} \mathsf{E}U(P_{X_0\|\mathbf{Z}})$ [3]

To summarize, we make an analogy between the concepts we have just seen and the familiar ones from information theory.

9.4 Caveats

(a) Computation of Posterior Distribution is hard. Bayes' rule gives

$$P_{X_i|Z^n}(x_i) = \frac{\sum_{x^{n\backslash i}} P_{X^n}(x^n)P_{Z^n|X^n}(z^n)}{\sum_{x^n} P_{X^n}(x^n)P_{Z^n|X^n}(z^n)}$$

where $x^{n\backslash i} = (x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots x_n)$. For certain special cases, the computation is less complex. Example: For a Markov source corrupted by DMC, the posterior distribution can be computed efficiently using forward-backward Recursion which is an instance of Dynamic Programming.

(b) Bayes' optimal solution requires the knowledge of the prior distribution which is rarely available in practice.

[1] The minimization is over all length functions associated with uniquely decodable codes for X^n.

[2] Assuming that $\mathbf{X}$ is stationary (see [1, Chapter 4]).

[3] Assuming that $(\mathbf{X}, \mathbf{Z})$ are jointly stationary.

These drawbacks call for a low complexity universal Denoiser, one that will essentially achieve optimum performance for any prior.

Bibliography

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory,* Wiley-Interscience, 2006.

Chapter 10

DUDE

10.1 Towards Discrete Universal DEnoiser (DUDE)

The setting is similar to the one before except that we consider an iid input process: Let $X \sim P_X$ represent a source symbol and let $\Pi_{|\mathcal{X}|\times|\mathcal{Z}|}$ denote the DMC channel matrix, where $\pi(x,z) = P(Z=z|X=x)$. Let $Z \sim P_Z$ denote the output symbol. Then, by total probability theorem, $P_Z(z) = \sum_x P_x(x)\pi(x,z)$, or $P_Z^T = P_X^T\Pi$.

Let's first assume that $|\mathcal{X}| = |\mathcal{Z}|$ and that the DMC matrix Π is invertible. Then we have $P_X^T = P_Z^T\Pi^{-1}$.

$$\begin{aligned} P_{X|Z=z}(x) &= \frac{P_X(x)\pi(x,z)}{P_Z(z)} \\ &= \frac{[\Pi^{-T}P_z](x)\pi(x,z)}{P_Z(z)} \end{aligned}$$

Define π_z as the column of Π that corresponds to the symbol z:

$$\pi_z = \begin{pmatrix} \pi(x_1,z) \\ \pi(x_2,z) \\ \vdots \\ \pi(x_{|\mathcal{X}|},z) \end{pmatrix}$$

Further, for vectors $v_1, v_2 \in R^n$, define $v_1 \odot v_2 \in R^n$ as the component-wise multiplication of v_1 and v_2 (also called Schur Product): $(v_1 \odot v_2)_i = v_{1i}v_{2i}$.

Then

$$P_{X|Z=z} = \frac{\Pi^{-T}P_Z \odot \pi_z}{P_Z(z)} \tag{10.1}$$

Since the input is iid and the channel is memoryless, symbol by symbol decoding is optimal. The optimal denoising function is the one that satisfies

$$\phi_{opt}(z) = \arg\min_{\phi} E\Lambda(X,\phi(Z)) \tag{10.2}$$

We proved in the previous lecture that $\phi_{opt}(z) = \hat{X}_{Bayes}(P_{X|Z=z})$. Hence, substituting from (10.1),

$$\phi_{opt}(z) = \hat{X}_{Bayes}(\Pi^{-T} P_Z \odot \pi_z) \tag{10.3}$$

Note that the normalization constant $P_Z(z)$ that appears in the denominator of (10.1) can be readily dropped due to the invariance of the Bayes' response to multiplication by a scalar.

We will now assume more generally that Π is of full row rank (and hence $|\mathcal{X}| \leq |\mathcal{Z}|$).

In that case, $P_Z^T = P_X^T \Pi \Rightarrow P_Z^T \Pi^T = P_X^T \Pi\, \Pi^T$ which yields $P_X^T = P_Z^T \Pi^T \left(\Pi\, \Pi^T\right)^{-1}$ where the full row rank condition guarantees the invertibility of $\Pi\, \Pi^T$. Thus (10.3) can be written more generally as

$$\begin{aligned}\phi_{opt}(z) =& \hat{X}_{Bayes}\left(\left(\Pi\, \Pi^T\right)^{-1} \Pi\, P_Z \odot \pi_z\right) \\ &= \arg\min_{\hat{x} \in \hat{\mathcal{X}}} \lambda_{\hat{x}}^T \left[\left(\Pi\, \Pi^T\right)^{-1} \Pi\, P_Z \odot \pi_z\right] \\ &\triangleq \Phi(\Lambda, \Pi, P_Z, z) \end{aligned} \tag{10.4}$$

10.2 Our Setup

We now describe our setup. We assume the following:

(a) $\mathcal{X}$, $\mathcal{Z}$, $\hat{\mathcal{X}}$ are finite alphabets representing source, output and reconstruction symbols respectively.

(b) The source $\mathbf{X}$ is unknown.

(c) The channel is a known DMC. Further, the DMC matrix Π is of full row-rank as described above. Note:

    (a) Condition 3 is benign. For example,

        i. *Binary Symmetric Channel* ($BSC(\delta)$: Condition holds $\iff \delta \neq \frac{1}{2}$.

        ii. *Binary Erasure Channel* ($BEC(\epsilon)$): Condition holds $\iff \epsilon < 1$.

        iii. *Z channel ($Z(p)$)*: Condition holds $\iff p < 1$.

        Refer fig. 10.1 for a description of these DMCs.

    (b) Condition 3 is necessary in the universality setting. Note that in a universal setting, a decoder can only see the output and is assumed to know nothing about the input distribution. Hence the decoder may estimate the output distribution, but cannot directly determine the input distribution. However one can argue intuitively that to construct a code that performs as well as Bayes' optimal solution, the Denoiser must be able to determine the input distribution uniquely.

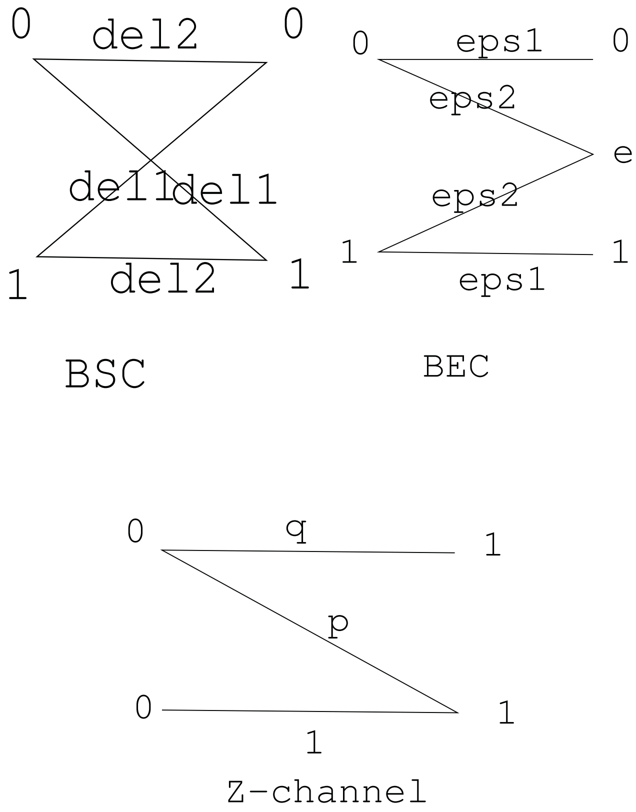

Figure 10.1: Discrete denoising system

The only way one can determine P_X is from the matrix equation $P_Z^T = P_X^T \Pi$. Hence it is reasonable to assume that Π is of full row rank and that $|\mathcal{X}| \leq |\mathcal{Z}|$.

(d) $\Lambda(x, \hat{x}) \geq 0$ is a loss function. Note that the non-negativity stipulation entails no essential loss of optimality since any loss function can be made to satisfy the condition by the addition of a sufficiently large constant.

We now define the Discrete Universal DEnoiser (DUDE). The idea behind DUDE is to "Correct by the Context". Define

$$m(z^n, l^k, r^k)[z] = \left|\left\{k+1 \leq i \leq n-k | z_{i-k}^{i+k} = (l^k, z, r^k)\right\}\right|$$

Here m is a $|\mathcal{Z}|-$dimensional column vector denoting the count of the symbol $z \in \mathcal{Z}$ in the double-sided context l^k, r^k.

The denoiser $DUDE(k)$ is the function

$$\hat{X}_i\left(Z^n\right) = \Phi\left(\Lambda, \Pi, m\left(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}\right), z_i\right)$$

where $\Phi(\,)$ is given by (10.4) with P_Z replaced by the context-based count $m(\,)$. Note that, up to an inconsequential normalization constant, m serves as an estimate of the conditional probability of the output symbol given the context.

10.3 The Setup: Universal Discrete Denoising

As depicted in Fig. 10.2, a sequence x^n from an unknown source passes through a discrete memoryless channel (DMC) characterized by a full-row-rank transition matrix Π. The denoiser produces an estimate $\widehat{X}^n$ so as to minimize loss as defined by a given loss matrix Λ.

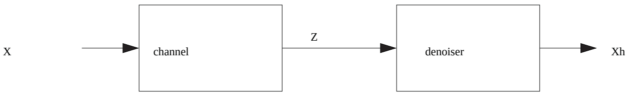

Figure 10.2: Discrete denoising system

10.4 The DUDE's operation

The denoiser operates in two phases.

(a) In the first pass, it computes the kth-order context statistics $m(z^n, l^k, r^k)$, defined as follows:

$$m(z^n, l^k, r^k)[z] = \left|\left\{k+1 \leq i \leq n-k : z_{i-k}^{i+k} = (l^k, z, r^k)\right\}\right|.$$

In other words, $m(z^n, l^k, r^k)$ is a histogram of those elements in z^n that have contexts l^k on the left and r^k on the right.

(b) In the second pass, the context statistics are used to denoise z^n. Formally, the kth-order estimate is given by a denoising function Φ:

$$\widehat{X}_i(z^n) = \Phi(\Lambda, \Pi, m(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}), z_i).$$

Φ is chosen to be the Bayes response for a source distribution derived from the channel matrix Π and the context statistics. Letting v denote the context-conditional histogram $m(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})$, the Bayes response takes the form

$$\Phi(\Lambda, \Pi, v, z) = \widehat{X}_{\text{Bayes}}(\Pi^{-T} v \odot \pi_z),$$

where π_z indicates the column of the channel matrix Π associated with the symbol z. For a non-square transition matrix Π, Π^{-T} is generalized to $(\Pi\Pi^T)^{-1}\Pi$.

Φ can be made more explicit if we allow $\lambda_{\hat{x}}$ to denote a column of the loss matrix Λ:

$$\Phi(\Lambda, \Pi, v, z) = \operatorname*{argmin}_{\hat{x}} \lambda_{\hat{x}}^T \Pi^{-T} v \odot \pi_z.$$

Exercise 10.4.1. *Suppose* Π *is a binary symmetric channel (BSC) with crossover probability* $\delta < 1/2$*, and suppose that we are interested in the Hamming loss function. That is,*

$$\Pi = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}, \Lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Show the following:

$$\Phi(\Lambda, \Pi, v, z) = \begin{cases} z \text{ if } \frac{v(z)}{v(\overline{z})} \geq \frac{2\delta(1-\delta)}{\delta^2+(1-\delta)^2} \\ \overline{z} \text{ otherwise.} \end{cases}$$

Exercise 10.4.2. *Let* Π *be an erasure channel with erasure probability* ϵ*. That is, the output alphabet* $\mathcal{Z}$ *is the input alphabet* $\mathcal{X}$ *augmented by an erasure symbol* $\{e\}$*, and the transition probabilities are given by*

$$\Pi(x, z) = \begin{cases} 1-\epsilon \text{ if } z = x \\ \epsilon \text{ if } z = e \end{cases}.$$

Furthermore, assume Λ *is still the Hamming loss matrix, as above. Show* Φ *outputs* z *if* z *is not* e*, and outputs* $\underset{\hat{x}}{\operatorname{argmax}}\, v(\hat{x})$ *if* z *is* e*.*

10.5 Choosing the DUDE's context length

A long context length k is desirable, as more contextual information is incorporated into the denoising. However, a large context length also results in fewer counts and, therefore, less reliable context statistics. To be concrete, we select the following

$$k = k_n = \left\lceil \frac{1}{5} \frac{\log n}{\log |\mathcal{Z}|} \right\rceil,$$

which will be justified through the performance guarantees. Denote the resulting denoiser by $\hat{X}^n_{\text{DUDE}}$.

10.6 The DUDE's Performance

We are interested in both the stochastic and semi-stochastic settings.

10.6.1 The Stochastic Setting

Here, we assume that the source is random, stationary, and possesses an unknown distribution. The channel is specified by a known transition matrix Π. In this setting, the DUDE is universally optimal in the following sense.

Theorem 10.6.1. *For any stationary process* $\mathbf{X}$*,*

$$\lim_{n\to\infty} L_{\widehat{X}^n_{DUDE}}(X^n, Z^n) = \lim_{n\to\infty} \min_{\hat{x}^n \in D_n} \mathsf{E} L_{\widehat{X}^n}(X^n, Z^n) = \mathbb{D}(\mathbf{X}, \mathbf{Z}),$$

where D_n *denotes the family of all* n*-block denoisers.*

10.6.2 The Semi-Stochastic Setting

Oftentimes, it is inappropriate to assume that the source is random, let alone stationary. In the semi-stochastic setting, we assume that the source $\mathbf{x}$ is an unknown deterministic sequence. However, we continue to characterize the channel as a DMC with a known probability transition matrix Π. The distribution of the noisy data z^n is then given by $\Pr(Z^n = z^n) = \prod_{i=1}^n \Pi(x_i, z_i)$.

To what standard can we compare the DUDE's performance in this semi-random context? First, consider the class of all functions $f : \mathcal{Z}^{2k+1} \to \widehat{\mathcal{X}}$ that estimate a source symbol x_i from the received symbol z_i and its context $(z_{i-k}^{i-1}, z_{i+1}^{i+k})$. We compare the DUDE's performance to that of the best element in this class as selected by a "genie" with access to the input sequence:

$$D_k(x^n, z^n) = \min_{f:\mathcal{Z}^{2k+1}\to\widehat{\mathcal{X}}} \frac{1}{n} \sum_{i=k+1}^{n-k} \Lambda\left(x_i, f(z_{i-k}^{i+k})\right)$$

Theorem 10.6.2. *For every sequence* $\mathbf{x}$

$$\lim_{n\to\infty} \left[L_{\widehat{X}^n_{DUDE}}(x^n, Z^n) - D_{k_n}(x^n, Z^n)\right] = 0 \quad \textit{w.p. } 1$$

Note that this theorem is stronger than Theorem 3. In fact, as we show below, Theorem 3 is proven as a corollary to Theorem 4.

10.6.3 A bit of terminology

(a) Let $\{a_n\}$ be a sequence. Then lim sup and lim inf are defined as follows:

$$\limsup_{n\to\infty} a_n = \overline{\lim}_{n\to\infty} a_n \equiv \lim_{n\to\infty} \sup_{m\geq n} a_m.$$

$$\liminf_{n\to\infty} a_n = \underline{\lim}_{n\to\infty} a_n \equiv \lim_{n\to\infty} \inf_{m\geq n} a_m.$$

(b)

Exercise 10.6.3. *Prove that* $\overline{\lim}_{n\to\infty} a_n = \underline{\lim}_{n\to\infty} a_n$ *if and only if* $\lim_{n\to\infty} a_n$ *exists.*

(c) Fatou's Lemma.

Lemma 10.6.4. *Let* $\{R_n\}$ *be a sequence of nonnegative random variables. Then*

$$\mathsf{E}[\underline{\lim}_{n\to\infty} R_n] \leq \underline{\lim}_{n\to\infty} \mathsf{E}[R_n].$$

c.f. [2] or [3] for a proof.

10.6.4 Proof of Theorem 1 Using Theorem 2

Proof First, we note that Theorem 4 directly implies

$$\lim_{n\to\infty}\left[L_{\widehat{X}^n_{\text{DUDE}}}(X^n,Z^n)-D_{k_n}(X^n,Z^n)\right]=0 \text{ w.p. } 1.$$

Next, fix some constant integer l and take the expectation of the above expression.

$$\begin{aligned}
0 &= \mathsf{E}\lim_{n\to\infty}\left[L_{\widehat{X}^n_{\text{DUDE}}}(X^n,Z^n)-D_{k_n}(X^n,Z^n)\right]\\
&= \mathsf{E}\overline{\lim}_{n\to\infty}\left[L_{\widehat{X}^n_{\text{DUDE}}}(X^n,Z^n)-D_{k_n}(X^n,Z^n)\right]\\
&\geq \overline{\lim}_{n\to\infty}\left[\mathsf{E}L_{\widehat{X}^n_{\text{DUDE}}}(X^n,Z^n)-\mathsf{E}D_{k_n}(X^n,Z^n)\right]\\
&\geq \overline{\lim}_{n\to\infty}\left[\mathsf{E}L_{\widehat{X}^n_{\text{DUDE}}}(X^n,Z^n)-\mathsf{E}D_{l}(X^n,Z^n)\right] \qquad (10.5)
\end{aligned}$$

Line two follows from existence of the limit in question, line three is a consequence of Fatou's lemma, and line four is true because k_n will eventually exceed the finite l (and obviously $D_l(x^n,Z^n)\geq D_k(x^n,z^n)$ for all n,x^n,Z^n and $l\leq k$).

We now upper bound the rightmost term in 10.5.

$$\begin{aligned}
\mathsf{E}D_l(X^n,Z^n) &= \mathsf{E}\min_{f:\mathcal{Z}^{2l+1}\to\widehat{\mathcal{X}}}\frac{1}{n}\sum_{i=l+1}^{n-l}\Lambda(X_i,f(Z_{i-l}^{i+l}))\\
&\leq \min_{f:\mathcal{Z}^{2l+1}\to\widehat{\mathcal{X}}}\frac{1}{n}\sum_{i=l+1}^{n-l}\mathsf{E}\Lambda(X_i,f(Z_{i-l}^{i+l}))\\
&= \min_{f:\mathcal{Z}^{2l+1}\to\widehat{\mathcal{X}}}\frac{n-2l}{n}\mathsf{E}\Lambda(X_0,f(Z_{-l}^{l}))\\
&= \frac{n-2l}{n}\mathsf{E}U(P_{X_0|Z_{-l}^{l}}) \qquad (10.6)
\end{aligned}$$

The second line is valid because the minimum of an expectation is greater than the expectation of the minimum, the line that follows is due to stationarity, and the last line follows from the definition of the Bayes Envelope (see lecture 2 notes).

Combining 10.6 and 10.5 yields:

$$\begin{aligned}
\overline{lim}_{n\to\infty}\mathsf{E}L_{\widehat{X}^n_{\text{DUDE}}} &\leq \lim_{n\to\infty}\frac{n-2l}{n}\mathsf{E}U(P_{X_0|Z_{-l}^{l}})\\
&= \lim_{n\to\infty}\mathsf{E}U(P_{X_0|Z_{-l}^{l}}).
\end{aligned}$$

The arbitrariness of l implies

$$\begin{aligned}
\overline{lim}_{n\to\infty}\mathsf{E}L_{\widehat{X}^n_{\text{DUDE}}} &\leq \lim_{l\to\infty}\mathsf{E}U(P_{X_0|Z_{-l}^{l}})\\
&= \mathbb{D}(\mathbf{X},\mathbf{Z}) \qquad (10.7)
\end{aligned}$$

where Eq. 10.7 is one of the HW exercises.

This completes the proof when combined with the obvious lower bound

$$\underline{\lim}_{n\to\infty} EL_{\widehat{X}^n_{\text{DUDE}}} \geq \underline{\lim}_{n\to\infty} \min_{\widehat{x}^n\in D^n} \mathsf{E}L_{\widehat{X}^n} = \mathbb{D}(\mathbf{X},\mathbf{Z}).$$

□

10.7 The Semi-Stochastic Setting

Here we show the optimality of the DUDE algorithm in the semi-stochastic setting described in Lecture 4. In particular, we prove Theorem 10.7.1 through a series of exercises. In the following we assume for simplicity that $|\mathcal{X}| = |\mathcal{Z}|$ and hence $\Phi(\Lambda,\Pi,v,z) = \widehat{X}_{Bayes}((\Pi^{-T}v)\odot\Pi_z)$. The more general case is handled similarly.

Theorem 10.7.1. *For every sequence* $\boldsymbol{x}$

$$\lim_{n\to\infty}\left[L_{\widehat{X}^n_{DUDE}}(x^n,Z^n) - D_{k_n}(x^n,Z^n)\right] = 0 \qquad w.p. \qquad 1$$

We begin by introducing some notation. Let $\Lambda_{max} = \max_{x,\hat{x}}\Lambda(x,\hat{x}) < \infty$. Also, let the count vector q be defined by

$$q(x^n,z^n,u_{-k}^k)[x] = |\{k+1\leq i\leq n-k : z_{i-k}^{i+k} = u_{-k}^k, x_i = x\}|.$$

In the following exercise we express the error achieved by the "genie" aided algorithm using the count vector q.

Exercise 10.7.2. *:*

(a) Show that

$$D_k(x^n,z^n) = \frac{1}{n}\sum_{u_{-k}^k} U(q(x^n,z^n,u_{-k}^k))$$

(b) Write $L_{\widehat{X}^n}(x_{k+1}^{n-k},z^n) = \frac{1}{n}\sum_{i=k+1}^{n-k}\Lambda(x_i,\widehat{X}_i(z^n))$. *For any* $w : \mathcal{Z}^n\times\mathcal{Z}^{2k+1}\to\mathbb{R}^{|\mathcal{X}|}$, *let* $\widehat{X}^n$ *satisfy*

$$\widehat{X}_i(z^n) = \widehat{X}_{Bayes}(w(z^n,z_{i-k}^{i+k})) \qquad k+1\leq i\leq n-k.$$

Then, $\forall x^n, z^n$, *show that*

$$0\leq L_{\widehat{X}^n}(x_{k+1}^{n-k},z^n) - D_k(x^n,z^n) \leq \frac{\Lambda_{max}}{n}\sum_{u_{-k}^k}||q(x^n,z^n,u_{-k}^k) - w(z^n,u_{-k}^k)||_1.$$

The second part of Exercise 12.4.2 bounds the suboptimality of a Bayes estimator that uses weights w using the ℓ_1-norm between q and w. Recall that,

$$\widehat{X}_{DUDE}(z^n)[i] = \widehat{X}_{Bayes}\left(\left(\Pi^{-T}m(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})\right) \odot \Pi_{z_i}\right)$$

From part (b) of Exercise 12.4.2 , it then follows that $\forall x^n, \epsilon > 0$,

$$\mathcal{P}\left(|L_{\widehat{X}^n_{DUDE}}(x^n, Z^n) - D_k(x^n, Z^n)| > \epsilon\right) \leq \mathcal{P}\left(\frac{\Lambda_{max}}{n} \sum_{u_{-k}^k} ||q(x^n, Z^n, u_{-k}^k - \Pi^{-T}m(Z^n, u_{-k}^{-1}, u_1^k) \odot \Pi_{u_0}||_1 > \epsilon\right) \quad (10.8)$$

To estimate the ℓ_1 norm of the difference between the count vectors, we write

$$q(x^n, Z^n, u_{-k}^k)[x] = \sum_{i=k+1}^{n-k} \mathbf{1}_{\{Z_{i-k}^{i+k} = u_{-k}^k, x_i = x\}}$$

and

$$\begin{aligned}
\Pi^{-T}m(Z^n, u_{-k}^{-1}, u_1^k) \odot \Pi_{u_0}[x] &= \Pi^{-T}m(Z^n, u_{-k}^{-1}, u_1^k)[x]\Pi(x, u_0) \\
&= \Pi(x, u_0) \sum_{\tilde{u}_0} \Pi^{-T}(x, \tilde{u}_0) m(Z^n, u_{-k}^{-1}, u_1^k)[\tilde{u}_0] \\
&= \Pi(x, u_0) \sum_{\tilde{u}_0} \Pi^{-T}(x, \tilde{u}_0) \sum_{i=k+1}^{n-k} \mathbf{1}_{\{Z_{i-k}^{i+k} = (u_{-k}^{-1}, \tilde{u}_0, u_1^k)\}}
\end{aligned}$$

In the following exercise, we establish that the random variable $\left(q(x^n, Z^n, u_{-k}^k)[x] - \Pi^{-T}m(Z^n, u_{-k}^{-1}, u_1^k)[x]\Pi(x, u_0)\right)$ is a sum of the form $\sum_{i=k+1}^{n-k} f_i(Z_{i-k}^{i+k})$ where the f_i have zero mean.

Exercise 10.7.3. *Prove that $\forall x^n$, u_{-k}^k, and $k+1 \leq i \leq n-k$,*

$$\mathbb{E}(\mathbf{1}_{\{Z_{i-k}^{i+k} = u_{-k}^k, x_i = x\}}) = \mathbb{E}\left(\Pi(x, u_0) \sum_{\tilde{u}_0} \Pi^{-T}(x, \tilde{u}_0) \mathbf{1}_{\{Z_{i-k}^{i+k} = (u_{-k}^{-1}, \tilde{u}_0, u_1^k)\}}\right)$$

We further note that $\{f_i\}$ are not only of zero mean by are also bounded and that f_i and f_j are independent when $|i-j| > 2k$. The following exercise is a consequence of these properties of $\{f_i\}$ combined with Hoeffding's inequality.

Exercise 10.7.4. *Show that,*

$$\begin{aligned}
\mathcal{P}\left(\frac{1}{n}\left|q(x^n, Z^n, u_{-k}^k)[x] - \Pi^{-T}m(Z^n, u_{-k}^{-1}, u_1^k)[x]\Pi(x, u_0)\right| \geq \epsilon\right) &= \mathcal{P}\left(\frac{1}{n}\left|\sum_{i=k+1}^{n-k} f_i(Z_{i-k}^{i+k})\right| \geq \epsilon\right) \\
&\leq 2(2k+1)\exp\left(-\frac{(n-2k)\epsilon^2}{2(2k+1)(1+|\mathcal{X}|\,||\Pi^{-1}||_\infty)^2}\right)
\end{aligned}$$

Hint : *Use the mean and depended structure of $\{f_i\}$ noted above, combined with*

Theorem 10.7.5. *(Hoeffding inequality) Let $V_1, V_2, \ldots V_n$ be independent random variables with $\mathbb{E}V_i = 0$ and $|V_i| \le c$ for all i. Then,*

$$\mathcal{P}\left(\frac{1}{n}|\sum_{i=1}^{n} V_i| \ge \epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{2c^2}\right)$$

cf. [4] for a simple proof of Hoeffding's inequality.

10.8 The DUDE's Perfomance

10.8.1 Proof of DUDE's Optimality in Semi-Stochastic Setting: Continued

Consider denoisers of the form

$$\hat{X}^n : \hat{X}_i\left(z^n\right) = \hat{X}_{Bayes}\left(w\left(z^n, z_{i-k}^{i+k}\right)\right), \quad k+1 \le i \le n-k, \tag{10.9}$$

where $w(\cdot,\cdot)$ takes values in $\mathbb{R}^{|\mathcal{X}|}$. Note that DUDE is of this form with $w = w_{DUDE}$ explicitly given by

$$w_{DUDE}(z^n, z_{i-k}^{i+k}) = \left(\Pi^{-T} m\left(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}\right)\right) \odot \pi_{z_i}. \tag{10.10}$$

Furthermore, recall that the difference in the average per-symbol loss of the *genie-aided* denoiser and a legitimate denoiser with some w is upper-bounded by where $\Lambda_{max} \triangleq \max_{x,\hat{x}} \Lambda(x, \hat{x})$ and $q(x^n, z^n, u_{-k}^k)$ is the genie-aided statistic defined by

$$q(x^n, z^n, u_{-k}^k)[x] = |\{k+1 \le i \le n-k : z_{i-k}^{i+k} = u_{-k}^k, x_i = x\}|, \quad x \in \mathcal{X} \tag{10.11}$$

Now, consider the performance of DUDE in our semi-stochastic setting with a deterministic x^n and a random output Z^n. Note that hereafter we will use a specific k for DUDE, which is given by

$$k = k_n = \lfloor \frac{1}{5}\frac{\log n}{\log |\mathcal{Z}|} \rfloor, \tag{10.12}$$

and we will use $\hat{X}^n_{DUDE}$ and $\hat{X}^{n,k_n}_{\text{DUDE}}$ interchangably. Then, we have the following upper bounds.

$$\begin{aligned}
&\Pr\left\{|L_{\hat{X}^n_{DUDE}}\left(x_{k+1}^{n-k}, Z^n\right) - D_k\left(x^n, Z^n\right)| \geq \epsilon\right\} \\
&\quad \overset{(a)}{\leq} \Pr\left\{\frac{\Lambda_{max}}{n}\sum_{u_{-k}^k}\sum_{x\in\mathcal{X}} \left|q(x^n, Z^n, u_{-k}^k)[x] - w_{DUDE}(Z^n, u_{-k}^k)[x]\right| \geq \epsilon\right\}, \\
&\quad \overset{(b)}{\leq} \Pr\left\{\bigcup_{u_{-k}^k, x}\left\{\frac{1}{n}\left|q(x^n, Z^n, u_{-k}^k)[x] - w_{DUDE}(Z^n, u_{-k}^k)[x]\right| \geq \frac{\epsilon}{|\mathcal{X}|^{2k+2}\Lambda_{max}}\right\}\right\}, \\
&\quad \overset{(c)}{\leq} \sum_{u_{-k}^k}\sum_{x\in\mathcal{X}}\underbrace{\Pr\left\{\frac{1}{n}\left|q(x^n, Z^n, u_{-k}^k)[x] - w_{DUDE}(Z^n, u_{-k}^k)[x]\right| \geq \frac{\epsilon}{|\mathcal{X}|^{2k+2}\Lambda_{max}}\right\}}_{:=J\left(x, u_{-k}^k\right)}.
\end{aligned} \tag{10.13}$$

The inequality (a) holds from (**??**), and the inequality (b) is obtained by applying the following union-type bound

$$\Pr\left(\sum_{i=1}^m v_i \geq \epsilon\right) \leq \Pr\left(\bigcup_{i=1}^m \left\{v_i \geq \frac{\epsilon}{m}\right\}\right), \tag{10.14}$$

and the inequality (c) is a direct result of the union bound. Define a function $B(n, k, \epsilon)$ as

$$B(n, k, \epsilon) \triangleq 2(2k+1)\exp\left(-\frac{n-2k}{2k+1}\cdot\frac{\epsilon^2}{2(1+|\mathcal{X}|\|\Pi^{-1}\|_\infty)^2}\right). \tag{10.15}$$

Then, from the exercise in the previous lecture, each $J\left(x, u_{-k}^k\right)$, $x \in \mathcal{X}$ and $u_{-k}^k \in \mathcal{X}^{2k+1}$, is upper-bounded by

$$J\left(x, u_{-k}^k\right) \leq B\left(n, k, \frac{\epsilon}{|\mathcal{X}|^{2k+2}\Lambda_{max}}\right). \tag{10.16}$$

Since $B\left(n, k, \frac{\epsilon}{|\mathcal{X}|^{2k+2}\Lambda_{max}}\right)$ in (10.16) does not depend on a particular choice of u_{-k}^k and x, combining (10.13) and (10.16), we have

$$\Pr\left\{|L_{\hat{X}^n_{DUDE}}\left(x_{k+1}^{n-k}, Z^n\right) - D_k\left(x^n, Z^n\right)| \geq \epsilon\right\} \leq |\mathcal{X}|^{2k+2} B\left(n, k, \frac{\epsilon}{|\mathcal{X}|^{2k+2}\Lambda_{max}}\right). \tag{10.17}$$

Suppose that k is fixed, or grows to infinity, but sufficiently slowly with n. Then, the RHS of (10.17) can be made to vanish quite quickly as n tends to infinity. To exploit that fact properly, we need the following lemma.

Lemma 10.8.1. *(Borel-Cantelli Lemma [1]): Let $\{E_n\}$ be a sequence of events satisfying*

$$\sum_{n=1}^{\infty} \Pr(E_n) < \infty, \qquad \textit{(infinitely summable)} \tag{10.18}$$

and define $\limsup_{n\to\infty}$ *of E_n as*

$$\limsup_{n\to\infty} E_n \triangleq \bigcap_{n=1}^{\infty} \left(\bigcup_{m=n}^{\infty} E_m \right). \tag{10.19}$$

Note that $\limsup_{n\to\infty} E_n$ *can be interpreted as the event that infinitely many of the events $\{E_n\}$ occur. Then,*

$$\Pr(\limsup_{n\to\infty} E_n) = 0. \tag{10.20}$$

Proof By the definition, (10.19), for every m,

$$P(\limsup_{n\to\infty} E_n) \leq P(\bigcup_{m=n}^{\infty} E_n) \leq \sum_{m=n}^{\infty} P(E_n). \tag{10.21}$$

Since the LHS of (10.21) is independent of m, it is bounded by the limit of the RHS as $m \to \infty$, which is 0 by (10.18). □

Exercise 10.8.2. *Let $C(n,k,\epsilon)$ denote the RHS of (10.17). Then, verify that*

$$\sum_{n=1}^{\infty} C(n,k,\epsilon) < \infty, \quad \textit{for any} \quad \epsilon > 0, \tag{10.22}$$

when $k = k_n$ as in (10.12).

Let now E_n be the event

$$|L_{\hat{X}^n_{DUDE}}\left(x_{k+1}^{n-k}, Z^n\right) - D_k\left(x^n, Z^n\right)| \geq \epsilon. \tag{10.23}$$

Then, combining the result of Exercise 10.8.2 and (10.17), we have

$$\sum_{n=1}^{\infty} \Pr(E_n) \leq \sum_{n=1}^{\infty} C(n,k,\epsilon) < \infty. \tag{10.24}$$

Using Borel-Cantelli Lemma, we have

$$\Pr\left(\lim_{n\to\infty} \sup |L_{\hat{X}^n_{DUDE}}\left(x_{k+1}^{n-k}, Z^n\right) - D_k\left(x^n, Z^n\right)| \geq \epsilon\right) = 0 \quad w.p.1 \tag{10.25}$$

implying

$$\lim_{n\to\infty} L_{\hat{X}^n_{DUDE}}\left(x_{k+1}^{n-k}, Z^n\right) - D_k\left(x^n, Z^n\right) = 0 \quad w.p.1, \tag{10.26}$$

by the arbitrariness of $\epsilon > 0$. This completes the optimality proof of DUDE in the semi-stochastic setting.

10.9 Lower Limits of Discrete Universal Denoising

If we look at the DUDE algorithm from previous lectures, the loss incurred by DUDE is not much worse than that of the best k$^{\text{th}}$ order sliding window denoiser. Compared to the benchmark (the best k$^{\text{th}}$ order sliding window denoiser), is there another denoiser whose excess loss is much smaller than that incurred by the DUDE? To answer this question, we will derive a lower bound on the excess loss of any denoiser compared to the same benchmark.

10.9.1 Background and Notation

The alphabet of the noiseless signal, as well as the noisy observation and the reconstruction is a M-letter alphabet, denoted by $\mathcal{A}$. Denote $\prod(i,j)$ the probability of the output symbol j when the input symbol is i. We assume a given loss function $\Lambda : \mathcal{A}^2 \to [0,\infty)$, where $\Lambda(i,j)$ defines the loss incurred by estimating the symbol i with the symbol j. An n-block denoiser is a mapping $\hat{X}^n : \mathcal{A}^n \to \mathcal{A}^n$. Let $L_{\hat{X}^n}(x^n, z^n)$ denote the normalized cumulative loss when the underlying noiseless sequence is x^n and the observed sequence is $z^n \in \mathcal{A}$, i.e.,

$$L_{\hat{X}^n}(x^n, z^n) = \frac{1}{n}\sum_{i=1}^{n} \Lambda\left(x_i, \hat{X}^n(z^n)[i]\right) \tag{10.27}$$

A *k-th order sliding window denoiser* $\hat{X}^n$ is a denoiser that is defined by a mapping

$$f : \mathcal{A}^{2k+1} \to \mathcal{A}$$

so that for all $z^n \in \mathcal{A}^n$

$$\hat{X}^n(z^n)[i] = f\left(z_{i-k}^{i+k}\right), \quad i = k+1, ..., n-k.$$

Let $\mathcal{S}_k$ be the collection of all k$^{\text{th}}$ order sliding window denoiser.

Question: Is k$^{\text{th}}$ order DUDE in $\mathcal{S}_k$?

The answer is No. Fix a z^n sequence, DUDE acts like a k$^{\text{th}}$ order sliding window denoiser for that sequence. But in general, for different z^n sequences, DUDE applies different sliding window denoiser.

The k$^{\text{th}}$ order minimum loss of (x^n,z^n) is defined as

$$\begin{aligned} D_k(x^n, z^n) &= \min_{\hat{X}^n \in S_k} L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) \\ &= \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, f(z_{i-k}^{i+k})). \end{aligned} \tag{10.28}$$

The expected k$^{\text{th}}$ order minimum loss is defined as

$$\hat{D}_k(x^n) \triangleq \mathsf{E}[D_k(x^n, Z^n)] \tag{10.29}$$

This quantity is the benchmark against which we will compare the loss incurred by other denoisers. Finally, the k$^{\text{th}}$ order regret $\hat{R}_k(\hat{X}^n)$ of any n-block denoiser is defined as follows:

$$\hat{R}_k(\hat{X}^n) = \max_{x^n \in \mathcal{A}^n} \left(E[L_{\hat{X}^n}(x_{k+1}^{n-k}, Z^n)] - \hat{D}_k(x^n) \right) \tag{10.30}$$

Specifically, we know that the k$^{\text{th}}$ order regret of DUDE is upper-bounded by

$$\hat{R}_k(\hat{X}^{n,k}_{\text{DUDE}}) \le C\sqrt{\frac{kM^{2k}}{n}}$$

The question we are going to answer in this lecture is whether there exists any denoiser which gives significantly better regret. We will show that for (all sufficiently large n), and any $\hat{X}^n$

$$\hat{R}_k(\hat{X}^n) \ge C\frac{\alpha^k}{\sqrt{n}}.$$

No denoiser can make regret approaching zero faster than $\mathcal{O}(\frac{1}{\sqrt{n}})$

10.9.2 Main Result

Let X^n be a sequence of i.i.d. random variables with $\mathbb{P}$ denoting the distribution of X_i. The quantity $\mathsf{E}[L_{\hat{X}^n}(X^n, Z^n)] - \hat{D}_k(X^n)$ is then a random variable. A key observation is that we can use the expectation of this random variable to lower bound the regret, i.e.,

$$\hat{R}_k(\hat{X}^n) \ge \mathsf{E}\left[\mathsf{E}[L_{\hat{X}^n}(X^n, Z^n)] - \hat{D}_k(X^n)\right] = \mathsf{E}\left[L_{\hat{X}^n}(X^n, Z^n)\right] - \mathsf{E}\left[\hat{D}_k(X^n)\right]$$

The first term on the RHS of above equation can be lower bounded as

$$\mathsf{E}[L_{\hat{X}^n}(x^n, Z^n)] \ge \min_{\hat{X}^n} \mathsf{E}[L_{\hat{X}^n}(X^n, Z^n)]$$

The minimizer is the Bayes response, i.e.,

$$\begin{aligned} \hat{X}^n_{\text{opt}}(z^n)[i] &= arg\min_{\hat{x}} \lambda_{\hat{x}}^T P_{X_i|z_i} \\ &= arg\min_{\hat{x}^n} \lambda_{\hat{x}}^T \frac{(\mathbb{P} \odot \pi_{z_i})}{\mathbb{P}^T \pi_{z_i}} \end{aligned}$$

Then the optimal loss becomes

$$D_{\text{opt}}(\mathbb{P}) = \min_{\hat{x}} \lambda_{\hat{x}}^T P_{X_i|z_i}$$

Using D_{opt} , the kth order regret is lower bounded by

$$\hat{R}_k(\hat{X}^n) \ge D_{\text{opt}}(\mathbb{P}) - \mathsf{E}[\hat{D}_k(X^n)] \tag{10.31}$$

10.9.3 BSC example

$$X_i \quad \sim \mathbb{P} = \begin{bmatrix} 1-P \\ P \end{bmatrix} \qquad \delta < \frac{1}{2}$$

$$\text{when} \qquad z_i = 1, \quad \hat{X}_{\text{opt}}(z^n)(i) = \begin{cases} 0 & P < \delta \\ 1 & P > \delta \\ \textit{either} & P = \delta \end{cases}$$

$$\text{when} \qquad z_i = 0, \quad \hat{X}_{\text{opt}}(z^n)(i) = \begin{cases} 0 & P < 1-\delta \\ 1 & P > 1-\delta \\ \textit{either} & P = 1-\delta \end{cases}$$

We can get $\hat{X}_{\text{opt}}$ and $D_{\text{opt}}(\mathbb{P})$by combining two cases

$$\hat{X}_{\text{opt}} = \begin{cases} \text{"always say 0"} & P \leq \delta \\ \text{"say what you see"} & \delta \leq P \leq 1-\delta \\ \text{"always say 1"} & P \geq 1-\delta \end{cases}$$

$$D_{\text{opt}}(\mathbb{P}) = \begin{cases} P & P \leq \delta \\ \delta & \delta \leq P \leq 1-\delta \\ 1-P & P \geq 1-\delta \end{cases}$$

Observe that when $P = \delta$, the crossover probability, there are two Bayes optimal denoisers, namely, the "always say o" and the "say-what-you-see" denoiser. We will try to lower bound the regret for $P = \delta$.

$$\text{when} \quad \mathbb{P} = \begin{bmatrix} 1-\delta \\ \delta \end{bmatrix}, \quad \hat{R}_k(\hat{X}^n) \geq \delta - E[\min_{\hat{X} \in S_0} L_{\hat{X}}(X^n, Z^n)]$$

The second term of the RHS of the above equation can be handled as follows:

$$\begin{aligned} \mathsf{E}\left[\min_{\hat{X} \in S_0} L_{\hat{X}}(X^n, Z^n)\right] &\leq \mathsf{E}\left[\min\{L_{always0}(X^n, Z^n), L_{swys}(X^n, Z^n)\}\right] \\ &= \mathsf{E}\left[\min\{\frac{1}{n}\sum_{i=1}^{n} 1\{X_i = 1\}, \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \neq Z_i\}\}\right] \\ &= \delta + \frac{1}{\sqrt{n}} E\left[\min\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1\{X_i = 1\} - \delta), \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1\{X_i \neq Z_i\} - \delta)\}\right] \end{aligned}$$

Note that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1\{X_i = 1\} - \delta)$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1\{X_i \neq Z_i\} - \delta)$ are sums of independent random variables. Further each set of random variables are independent of each other. Therefore, by the Central Limit Theorem, they coverge in distribution to independent zero mean Gaussian random variables $(N(0, \delta(1-\delta)))$.

Therefore,

$$\mathsf{E}[\min_{\hat{X}\in S_0} L_{\hat{X}}(X^n, Z^n)] \approx \delta - \frac{C}{\sqrt{n}}, \quad C > 0.$$

Now we have

$$\begin{aligned}\hat{R}_k(\hat{X}^n) &\geq \delta - \mathsf{E}[\min_{\hat{X}\in S_0} L_{\hat{X}}(X^n, Z^n)] \\ &\geq \delta - (\delta - \frac{C}{\sqrt{n}}) \\ &= \frac{C}{\sqrt{n}} \end{aligned} \tag{10.32}$$

10.9.4 Proof of Lower bound

Definition: (π, λ) is *neutralizable* if $\exists$ channel output symbol t $\in \mathcal{A}$, s.t. for some $\mathbb{P} \in \mathcal{M}$ ($\mathcal{M}$: simplex M-dimensional)

(a) $\lambda_i^T(\mathbb{P} \odot \pi_t) = \lambda_j^T(\mathbb{P} \odot \pi_t) = \min_{\hat{x}} \lambda_{\hat{x}}^T(\mathbb{P} \odot \pi_t)$

(b) $(\lambda_i - \lambda_j) \odot P \odot \pi_t \neq 0$

Further the distribution $\mathbb{P}$ that satisfies the two equations is termed *loss-neutral.*

Theorem 10.9.1. *: For any neutralizable (*π*,*λ*), and any sequence of denoisers* $\{\hat{X}^n\}$

$$\hat{R}_k(\hat{X}^n) \geq \frac{C}{\sqrt{n}}\Big(\sum_a \sqrt{(P^*)^T\pi_a}\Big)^{2k}(1 + o(1))$$

where P^* *is any loss-neutral distribution and* C *is a positive function of* (π, λ) *and* P^**.*

Proof: Let

$$q(z^n, x^n, c_{-k}^k)[\alpha] \quad = \quad \frac{|\{i : z_{i-k}^{i+k} = c_{-k}^k, x_i = \alpha\}|}{n-2k}, \quad \text{where} \quad \alpha \in \mathcal{A}.$$

Suppose X^n is iid with $X_i \sim \mathbb{P}$

$$\begin{aligned} E[q(z^n, x^n, c_{-k}^k)] &= (\mathbb{P} \odot \pi_{c_0})\Pi_{\substack{i=-k \\ i\neq 0}}^{k} \mathbb{P}^{\mathbb{T}}\pi_{c_i} \\ D_k(x^n, z^n) &= \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} \frac{1}{n-2k}\sum_{i=k+1}^{n-k} \Lambda(x_i, f(z_{i-k}^{i+k})) \\ &= \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}} \min_{\hat{x}\in\mathcal{A}} \sum_j \Lambda(j, \hat{x}) q(z^n, x^n, c_{-k}^k)[j] \\ &= \sum_{c_{-k}^k} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T q(z^n, x^n, c_{-k}^k) \end{aligned}$$

Definition: $X_1, \ldots, X_n$ is m-dependent if for all $s > r + m$, $X_1, \ldots, X_r$ and $X_s, \ldots, X_n$ are independent

Theorem 10.9.2. *For a stationary m-dependent sequence X^n s.t. $E[X_i] = 0$ and $E[|X_i|^3] < \infty$,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \to N(0, V), \text{ as } n \to \infty,$$

where $V = E[X_1^2] + 2\sum_{k=2}^{m+1} E[X_1, X_k]$

This theorem is proved by Hoeffding and Robbins [5]. The following lemma is a consequence of the theorem. The proof of the lemma can be found in [6].

Lemma 10.9.3. *X^n is iid and $X_i \sim \mathbb{P}$. Then, for any $\alpha \in \mathbb{R}^n$ and any $c_{-k}^k \in \mathcal{A}^{2k+1}$ s.t. $\alpha^T(\mathbb{P} \odot \pi_{c_0}) = 0$,*

$$\lim_{n\to\infty} \mathsf{E}_{\mathbb{P}}\left[\sqrt{n}\left|\alpha^T q(z^n, x^n, c_{-k}^k)\right|\right] = \sqrt{\frac{\alpha V}{\pi}}$$

where $V = (\alpha \odot \alpha)^T(\mathbb{P} \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} \mathbb{P}^T \pi_{c_i}$.

Let us complete the proof of Theorem 1.

Recall Eq. 10.31,

$$\hat{R}_k(\hat{X}^n) \geq D_{\text{opt}}(\mathbb{P}) - \mathsf{E}[\hat{D}_k(X^n)]$$

If $X_i \sim \mathbb{P}^* \to$ loss-neutral w.r.t. $(\pi_t, \lambda_i, \lambda_j)$,

$$\hat{R}_k(\hat{X}^n) \geq \sum_{c_0 \in \mathcal{A}} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T(\mathbb{P}^* \odot \pi_{c_0}) - E[\hat{D}_k(X^n)] \tag{10.33}$$

Let us find an upper bound for the second term of R.H.S. of the above equation.

$$\begin{aligned}
\mathsf{E}[\hat{D}_k(X^n)] &= \mathsf{E}[D_k(X^n, Z^n)] \\
&= \mathsf{E}\Big[\sum_{c_{-k}^k} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T q(Z^n, X^n, c_{-k}^k)\Big] \\
&= \sum_{c_{-k}^k, c_0 \neq t} \mathsf{E}[\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T q(Z^n, X^n, c_{-k}^k)] + \sum_{c_{-k}^k, c_0 = t} \mathsf{E}[\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T q(Z^n, X^n, c_{-k}^k)]
\end{aligned} \tag{10.34}$$

The first term on the R.H.S. of Eq. 10.34 can be upper bounded as follows:

$$\begin{aligned}
\sum_{c_{-k}^k, c_0 \neq t} \mathsf{E}[\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T q(Z^n, X^n, c_{-k}^k)] &\leq \sum_{c_{-k}^k, c_0 \neq t} \min_{\hat{x}\in\mathcal{A}} E[\lambda_{\hat{x}}^T q(Z^n, X^n, c_{-k}^k)] \\
&= \sum_{c_{-k}^k, c_0 \neq t} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T(\mathbb{P} \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} \mathbb{P}^{\mathbb{T}} \pi_{c_i} \\
&= \sum_{c_0 \neq t} \min_{\hat{x}} \lambda_{\hat{x}}^T(\mathbb{P} \odot \pi_{c_0})
\end{aligned} \tag{10.35}$$

The second term on the R.H.S. of Eq. 10.34 can be upper bounded in the following way:

$$\begin{aligned}
& \sum_{c_{-k}^{k}, c_0=t} \mathsf{E}\left[\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T q(Z^n, X^n, c_{-k}^k)\right] \\
\leq \quad & \sum_{c_{-k}^{k}, c_0=t} \mathsf{E}\left[\min\{\lambda_i^T q(Z^n, X^n, c_{-k}^k), \lambda_j^T q(Z^n, X^n, c_{-k}^k)\}\right] \\
\overset{(a)}{=} \quad & \frac{1}{2} \sum_{c_{-k}^{k}, c_0=t} \Big(\mathsf{E}[\lambda_i^T q(Z^n, X^n, c_{-k}^k)] + \mathsf{E}[\lambda_j^T q(Z^n, X^n, c_{-k}^k)] - \mathsf{E}\left[\left|(\lambda_i - \lambda_j)^T q(Z^n, X^n, c_{-k}^k)\right|\right] \Big) \\
\overset{(b)}{=} \quad & \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T (\mathbb{P} \odot \pi_t) - \frac{1}{2} \sum_{c_{-k}^{k}, c_0=t} \mathsf{E}[|(\lambda_i - \lambda_j)^T q(Z^n, X^n, c_{-k}^k)|] \\
\overset{(c)}{=} \quad & \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T (\mathbb{P} \odot \pi_t) - \sum_{c_{-k}^{k}, c_0=t} \frac{C(1+o(1))}{\sqrt{n}} \sqrt{\frac{2V_{c_{-k}^k}}{\pi}} \qquad (10.36)
\end{aligned}$$

where $V_{c_{-k}^k} = ((\lambda_i - \lambda j) \odot (\lambda_i - \lambda j))^T (\mathbb{P} \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} \mathbb{P}^T \pi_{c_i}$, (a) follows because $\min\{x, y\} = \frac{x+y-|x-y|}{2}$, (b) follows from the defintion of neutralizable and loss-neutrality, and (c) is due to Lemma 3.

Finally, substituting Eq. 10.35 and Eq. 10.36 into Eq. 10.34 and then using the obtained upper bound of $E[\hat{D}_k(X^n)]$ in Eq. 10.33, we have

$$\begin{aligned}
\hat{R}_k(\hat{X}^n) \quad \geq \quad & D_{\text{opt}}(\mathbb{P}) - \mathsf{E}[\hat{D}_k(X^n)] \\
= \quad & \sum_{c_{-k}^k} \min_{\hat{x}} \lambda_{\hat{x}}^T (\mathbb{P} \odot \pi_t) - \sum_{c_{-k}^k} \min_{\hat{x}} \lambda_{\hat{x}}^T (\mathbb{P} \odot \pi_t) + \frac{C(1+o(1))}{\sqrt{n}} \sum_{c_{-k}^{k}, c_0=t} \sqrt{\frac{2V_{c_{-k}^k}}{\pi}} \\
= \quad & \sum_{c_{-k}^{k}, c_0=t} \frac{C(1+o(1))}{\sqrt{n}} \sqrt{\frac{2V_{c_{-k}^k}}{\pi}}. \qquad (10.37)
\end{aligned}$$

Observe that

$$\sum_{c_{-k}^{k} \in \mathcal{A}^{2k+1}, c_0=t} \sqrt{\frac{2V_{c_{-k}^k}}{\pi}} = \sqrt{\frac{2}{\pi}} \sqrt{((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbb{P}^* \odot \pi_t)} \sum_{a_1^{2k} \in \mathcal{A}^{2k}} \left(\prod_{i=1}^{2k} (\mathbb{P}^*)^T \pi_{a_i} \right)^{\frac{1}{2}}.$$

Note that since $\mathbb{P}^*$ is a loss-neutral distribution $(\lambda_i - \lambda_j) \odot \mathbb{P}^* \odot \pi_t \neq 0$, and therefore

$$((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbb{P} \odot \pi_t) > 0.$$

Also observe that

$$\sum_{a_1^{2k} \in \mathcal{A}^{2k}} \left(\prod_{i=1}^{2k} (\mathbb{P}^*)^T \pi_{a_i} \right)^{\frac{1}{2}} = \left(\sum_{a\in\mathcal{A}} \sqrt{(\mathbb{P}^*)^T \pi_a} \right)^{2k}.$$

For more details and discussions see [6].

10.10 Some concluding remarks on DUDE

We have studied DUDE in the setup of the universal discrete denoising with a known channel: how it works and what is the performance it can achieve. It is shown that, with the jointly stationary process $(\mathbf{X}, \mathbf{Z})$, DUDE can do remarkably well compared to the optimal denoiser, which can access both to the noisy output sequence and the underlying clean source sequence. Indeed, in the limit of n tends to infinity, DUDE with the context length k_n, $\hat{X}^{n,k_n}_{\mathrm{DUDE}}$, can achieve the same performance as the optimal denoiser of the same context length in terms of the average per-symbol loss. We conclude this section with some discussion on the context length and the non-asymptotic performance of DUDE.

10.10.1 Context Length of DUDE

So far we have considered the explicit context length of $k_n = \lfloor \frac{1}{5} \frac{\log n}{\log |\mathcal{Z}|} \rfloor$ to be concrete with our analysis. However this is not necessarily the best k_n that we can choose, and indeed we can do better choices based on more refined analysis. Here are some theoretical and practical guidelines for choosing k_n.

Remark

a) The relation $k_n|\mathcal{Z}|^{2k_n} = o(n \log n)$ can be shown to be sufficient for the validity of the Theorem 2, by using a better exponential-type bound [2, 3], which decreases more rapidly w.r.t k_n. So, e.g., $k_n = c\frac{\log n}{\log |\mathcal{Z}|}, \quad c < \frac{1}{2}$, suffices to guarantee that DUDE competes with the genie of order k_n. Note that this is true despite the fact that $k_n|\mathcal{Z}|^{2k_n}$ does not suffice to guarantee the convergence of the union bound in (10.17).

b) In practice, k_n can be determined in various ways.

   - Data-dependent selection
     We can rely on a *compressibility heuristic* to try a number of different context lengths and select the k_n resulting in the most compressible reconstruction, or *dynamic context* of the data can be taken into account, meaning that k_n may vary depending on the location of a symbol that we want to denoise.
   - Other possible tweaks
     We can perform *context aggregation* or *iterated DUDE* to improve the denoising performance. In particular, in case of the iterated DUDE, the equivalent channel becomes no longer memoryless as the number of iteration increases, and we cannot use the same technique used in DMC to analyze the performance of DUDE. However, some empirical results show that the denoising performance can be improved up to a certain number of iterations.

10.10.2 Non-Asymptotic Performance

In practice the sequence length n is fixed and finite, thus the asymptotic performance of DUDE with n tending to infinity may not be a useful performance indicator. Therefore, we present the non-asymptotic upper bound and lower bound on the average per-symbol loss of DUDE as

- Non-asymptotic upper bound

$$E\left[L^n_{\hat{X}_{DUDE}}\left(x^n, Z^n\right) - D_k\left(x^n, Z^n\right)\right] \leq C_u \cdot \sqrt{\frac{k|\mathcal{Z}|^{2k}}{n}}, \quad \forall x^n \text{ and } k \tag{10.38}$$

- Non-asymptotic lower bound

$$\max_{x^n} E\left[L^n_{\hat{X}_{DUDE}}\left(x^n, Z^n\right) - D_k\left(x^n, Z^n\right)\right] \geq C_l \cdot \frac{c^k}{\sqrt{n}}, \quad c > 1, \quad \forall k \tag{10.39}$$

where C_u and C_l are some constants, which only depend on Π and Λ. See [6] for the details.

Bibliography

[1] T.M. Cover and J.A Thomas, “Elements of Information Theory”, Wiley-Interscience, 2006.

[2] R. Durrett, Probability: Theory and Examples, third edition, Thomson, Brooks/Cole, 2003.

[3] A. Dembo and O. Zeitouni, ”Large Deviations: Techniques and Applicatoins”, second edition, Springer, 1998

[4] L. Devroye, L. Gyorfi and G. Lugosi, “A Probabilistic Theory of Pattern Recognition”, *Springer*, 1996.

[5] W. Hoeffding and H. Robbins, “The central limit theorem for dependent random variables,”*Duke Math. J.*, vol 15, no. 3, pp. 773-780.

[6] K. Viswanathan, Erik Ordentlich,“Lower Limits of Discrete Universal Denoising”,*IEEE Trans. Inf. Theory*, VOL 55, No.3 March 2009.