Uniform Concentration of the Loss Estimator for Neural DUDE

Taesup Moon  
College of Information and Communication Engineering  
School of Electronic and Electrical Engineering  
Sungkyunkwan University (SKKU), Suwon, Korea 16419  
tsmoon@skku.edu

Abstract

We give a theoretical justification of the concentration property observed for the recently developed neural network-based sliding window discrete denoiser, Neural DUDE. Namely, we rigorously prove that the estimated loss devised for Neural DUDE, computed solely from the noisy data, concentrates on the true denoising loss, which can only be evaluated with the underlying clean data. The concentration is shown to hold in a strong sense, i.e., it uniformly holds over all the bounded network parameters including the parameters of Neural DUDE and over all possible underlying clean data, with high probability. Moreover, we characterize the sufficient condition for the concentration, in terms of the sliding window size $k$ and the data size $n$, as $k = o(\sqrt{n})$, which is a much weaker condition than that of DUDE, the predecessor of Neural DUDE. For the proof, we make a novel application of the tools of the learning theory, e.g., Rademacher complexity. We conclude with experimental results that highlight the theoretical results and advocate the hyperparameter selection method of Neural DUDE.

1 Introduction

Cleaning noise-corrupted discrete-valued data, also known as discrete denoising, has diverse different application areas, e.g., image denoising and DNA sequence denoising. Universal discrete denoising, which assumes nothing about the data except for the noise mechanism being a discrete memoryless channel (DMC), has been first considered in [1]. Discrete Universal DEnoiser (DUDE), which operates as a sliding-window denoiser, proposed in [1] enjoyed the rigorous theoretical performance guarantees in the universal setting as well as strong empirical performance for several applications, e.g., binary image denoising [2] and DNA sequence denoising [3].

While DUDE had strong results, one of its major drawbacks was that the performance of the method is highly sensitive on the choice of its hyperparameter, i.e., the sliding-window size $k$, and it had to be hand-tuned without any systematic rule. In order to overcome this drawback, which is mainly due to the nonparametric nature of DUDE that separately obtains the empirical distribution of the noisy symbols for different contexts, [4] introduced a neural network based sliding-window denoiser, dubbed as Neural DUDE. By devising novel “pseudo-labels”, Neural DUDE trains a single neural network, solely from the noisy data, that shares information from similar contexts through its parameters. As a result, the experiments in [4] showed that Neural DUDE can significantly outperform DUDE and maintain robustness with respect to $k$.

One by-product of Neural DUDE was that it has a reliable estimator of the true denoising performance; namely, an estimated loss of Neural DUDE can be computed from the noisy data, and it tends to concentrate on the true loss. Therefore, in practice, in which the true loss can never be evaluated, we can select the hyperparameter $k$ of Neural DUDE by selecting the one that minimizes the estimated loss that should concentrate on the true loss. While such concentration of the estimated loss can be
also observed for DUDE for some small $k$’s as well, it quickly fails as the window size $k$ grows. In contrast, for Neural DUDE, it has been observed in [4] that the concentration stably holds for much larger $k$ values.

In this paper, we give a rigorous proof of such concentration property of the estimated loss for Neural DUDE to justify its usage for the hyperparameter selection. In fact, we show a much stronger result; namely, we show that the concentration of the estimated loss holds with high probability uniformly over all neural network based denoisers that have the same architecture as the Neural DUDE. The concentration is also universal for all underlying clean sequences. We also explicitly specify the sufficient condition on the window size $k$ and the data size $n$ for the concentration to hold, i.e., $k = o(\sqrt{n})$. By comparing this condition to that of DUDE, $k = o(\log n)$, our result justifies the empirical finding of [4] that the concentration for Neural DUDE holds for much larger $k$’s for a fixed $n$ than those for DUDE. In our proof, we utilize the tools from statistical learning theory, e.g., Rademacher complexity [5], in a novel way; i.e., while the typical learning theory focuses on the uniform concentration of the training error on the test error, we show the uniform concentration of the estimated loss on the true loss in the denoising setting.

2 Notations and Preliminaries

2.1 Notations and problem setting

We will mainly follow the notations in [4]. Throughout the paper, an $n$-tuple sequence is denoted as, e.g., $a^n = (a_1, \ldots, a_n)$, and $a^i$ refers to the subsequence $(a_i, \ldots, a_j)$. The uppercase letters will stand for the random variables, and the lowercase letters will stand for either the realizations of the random variables or the individual symbols. We denote $\Delta^n$ as the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$.

In universal/anonymous denoising, the clean, underlying source data will be denoted as an individual sequence $x^n$ as we do not assume any probabilistic models on it. We assume each component $x_i$ takes a value in some finite set $\mathcal{X}$. For example, for binary data, $\mathcal{X} = \{0, 1\}$, and for DNA data, $\mathcal{X} = \{A, C, G, T\}$.

When the source sequence is corrupted by a Discrete Memoryless Channel (DMC), namely, the index-independent noise, it results in a noisy version of the source, $Z^n$, of which each $Z_i$ takes a value in, again, a finite set $\mathcal{Z}$. The DMC is completely characterized by the channel transition matrix $\Pi \in \mathbb{R}^{|\mathcal{Z}| \times |\mathcal{X}|}$, of which the $(x, z)$-th element stands for $\Pr(Z = z | X = x)$, i.e., the conditional probability of the noisy symbol taking value $z$ given the source symbol was $x$. An essential but natural assumption we make is that $\Pi$ is of the full rank. We also denote $\Pi^\dagger = (\Pi \Pi^\top)^{-1}$ as the Moore-Penrose pseudoinverse of $\Pi$. In our setting, $\Pi$ is assumed to be known to the denoiser.

Upon observing the entire noisy data $Z^n$, a discrete denoiser reconstructs the original data with $\hat{X}^n = (\hat{X}_1(Z^n), \ldots, \hat{X}_n(Z^n))$, where each reconstructed symbol $\hat{X}_i(Z^n)$ takes its value in a finite set $\hat{\mathcal{X}}$. The goodness of the reconstruction by $\hat{X}^n$ is measured by the average loss, $L_{\hat{X}_n}(x^n, Z^n) = \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, \hat{X}_i(Z^n))$, where $\Lambda(x_i, \hat{x}_i)$ is a loss function that measures the loss incurred by estimating $x_i$ with $\hat{x}_i$. The loss function is fully represented with a loss matrix $\Lambda \in \mathbb{R}^{|\hat{\mathcal{X}}| \times |\mathcal{X}|}$.

The $k$-th order sliding window denoisers are the denoisers that are defined by a time-invariant mapping $s_k : Z^{2k+1} \to \hat{\mathcal{X}}$. That is, $\hat{X}_i(Z^n) = s_k(Z_{i-k}^{i+k})$. We also denote the tuple $(Z_{i-k}^{i+k}, Z_{i+1}^{i+k+1}) \triangleq C_i$ as the $k$-th order double-sided context around the noisy symbol $Z_i$, and we let $C|k|$ as the set of all such contexts. As discussed in [4], both DUDE in (1) and Neural DUDE are sliding window denoisers. We also denote $\mathcal{S} \triangleq \{s : \mathcal{Z} \to \hat{\mathcal{X}}\}$ as the set of single-symbol denoisers that are sliding window denoisers with $k = 0$. Note $|\mathcal{S}| = |\hat{\mathcal{X}}| \times |\mathcal{Z}|$. Then, an alternative view of of $s_k(\cdot)$ is that $s_k(C_i, \cdot) \in \mathcal{S}$ is a single symbol denoiser defined by $C_i$ and applied to $Z_i$.

The basic building block of Neural DUDE is the unbiased estimated loss function as described in [4] Section 3.1. That is, based on the known $\Pi$ assumption, we can devise an estimated loss $L = \Pi \rho \in \mathbb{R}^{|\mathcal{Z}| \times |\mathcal{S}|}$, in which $\rho \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{S}|}$ with the $(x, s)$-th element is $\mathbb{E}_x \Lambda(x, s(Z))$. The notation $\mathbb{E}_x(\cdot)$ stands for the expectation with respect to the conditional distribution $\Pr(Z = s | X = x)$ defined by the $x$-th row of $\Pi$. Then, as shown in [4], $L$ has the unbiased property, $\mathbb{E}_x L(Z, s) = \mathbb{E}_x \Lambda(x, s(Z))$.

Note we are using the uppercase notation $C_i$ to highlight the randomness as opposed to $c_i$ used in [4].
2.2 Neural DUDE

Neural DUDE of [4] defines a single fully-connected neural network \( p^k(w, \cdot) \) \( : \mathcal{Z}^{2k} \rightarrow \Delta^{\mid S \mid} \) that works as a sliding-window denoiser. That is, at location \( i \), the network takes the double-sided context \( C_i \in \mathbb{C}[k] \) as input and outputs the probability distribution on the single symbol denoisers to apply to \( Z_i \), \( w \) stands for all the parameters in the network. Figure 1(a) shows an example architecture of Neural DUDE in which \( L \) is the total number of layers and \( n_L \) is the number of nodes in the \( \ell \)-th layer. By carrying out the one-hot encoding of each noisy symbol, the total number of parameters of Neural DUDE then becomes \( 2k\mid Z \mid n_1 + \prod_{\ell=2}^{L} n_{\ell-1} n_{\ell} + n_L \mid S \mid \). Furthermore, the ReLU (Rectified Linear Unit), \( f(x) = \max(0, x) \), is used as the activation function for all intermediate hidden nodes. For the output layer, the usual softmax function is used.

![Figure 1: (a) An example network architecture \( p^k(w, \cdot) \) of Neural DUDE with \( L \) layers. (b),(c) The Bit Error Rate (BER) and the estimated BER plots for DUDE and Neural DUDE for the synthetic binary example in [4] Section 5.1. Note the scales of the y-axis are very different in the two plots.](image)

In order to train the network parameters, Neural DUDE computes the matrix \( L_{\text{new}} \in \mathbb{R}^{\mid Z \mid \times \mid S \mid} \) defined as \( L_{\text{new}} \triangleq -L + L_{\text{max}} 1_{\mid Z \mid} 1_{\mid S \mid}^\top \), in which \( L_{\text{max}} \triangleq \max_{z,s} L(z, s) \), and \( 1_{\mid Z \mid} \) and \( 1_{\mid S \mid} \) stand for the all-1 vector with \( \mid Z \mid \) and \( \mid S \mid \), respectively. We note that all the elements in \( L_{\text{new}} \) are designed to be non-negative. Once \( L_{\text{new}} \) is computed, Neural DUDE uses the objective function as in [4] Eq.(7), that uses \( L_{\text{new}} 1_{Z_i} \in \mathbb{R}^{\mid S \mid} \) as a “pseudo-label” for the single-symbol denoiser at the \( i \)-th location. In order to learn the parameter \( w \), the training procedure of Neural DUDE is carried out equivalently as the ordinary supervised learning setting that uses the back-propagation and variants of mini-batch SGD (e.g., [7]) to minimize the objective function. For more details on Neural DUDE, we refer the readers to [4].

Once the objective function converges, we denote \( w^* \) as the converged parameters and denote the single-letter mapping defined by Neural DUDE for the context \( C_i \in \mathbb{C}[k] \) as \( s_{k,\text{Neural DUDE}}(C_i, \cdot) \). Thus, the reconstruction at location \( i \) becomes \( \hat{X}_i, \text{Neural DUDE}(Z^n) = s_{k,\text{Neural DUDE}}(C_i, Z_i) \).

3 Main Result

One interesting phenomenon that [4] states without rigorous proof is the concentration property of the estimated loss for Neural DUDE. That is, as given in [4] Equation (9) and restated below, the average estimated loss of Neural DUDE tends to concentrate on the true loss even for very large \( k \)'s:

\[
\frac{1}{n} \sum_{i=1}^{n} L(Z_i, s_{k,\text{Neural DUDE}}(C_i, \cdot)) \approx \frac{1}{n} \sum_{i=1}^{n} L_i(x_i, s_{k,\text{Neural DUDE}}(C_i, Z_i)).
\]  

(1)

In contrast, the average estimated loss of DUDE quickly diverges from the true loss as \( k \) grows. Such sharp contrast is again shown in Figure 1(b) and 1(c) for the synthetic binary example.

The property (1) makes Neural DUDE practical as it enables to choose the best window size \( k \) based on the average estimated loss, which can be always computed from the noisy observation \( Z^n \). The main focus of this paper is to prove the property (1) rigorously with specifying the convergence rate in terms of \( k \) and \( n \).

Before introducing some additional notations and stating our main theorem, we first state a proposition that shows the weak concentration property of the original DUDE in [1].
Proposition 1 Let $S_k$ denote the class of all the $k$-th order sliding window denoisers, $s_k$, and let $\delta > 0$. Then, for all $x^n \in \mathcal{X}^n$, with probability at least $1 - \delta$, we have

$$
\max_{s_k \in S_k} \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \left[ \mathcal{L}(Z_i, s_k(C_i, \cdot)) - \Lambda(x_i, s_k(C_i, Z_i)) \right] = O\left( \sqrt{\frac{k|Z|^{2k} \log(|S|/\delta)}{n}} \right). \quad (2)
$$

Remark: We observe that the upper bound in the proposition vanishes when $k = o(\log n)$. Since DUDE with window size $k$ is also in $S_k$, the proposition gives the justification of the poor concentration property of DUDE with respect to $k$ for fixed $n$ in Figure 1(b).

Proof: For the notational brevity, we denote

$$
(*) : \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \left[ \mathcal{L}(Z_i, s_k(C_i, \cdot)) - \Lambda(x_i, s_k(C_i, Z_i)) \right].
$$

Then, for $\epsilon > 0$, we have

$$
\Pr \left( \max_{s_k \in S_k} (*) > \epsilon \right) \leq \sum_{s_k \in S_k} \Pr \left( (*) > \epsilon \right) \leq |S_k|(k+1) \exp \left( - \frac{2(n-2k)\epsilon^2}{(k+1)C_{\max}^2} \right)
$$

(3)

$$
\leq |S| |Z|^{2k} (k+1) \exp \left( - \frac{2(n-2k)\epsilon^2}{(k+1)C_{\max}^2} \right)
$$

(4)

in which $C_{\max} \triangleq \max_{z,s} |\mathcal{L}(z,s)| + \max_{z,\hat{z}} |\Lambda(x, \hat{z})|$, the first inequality in (3) follows from the union bound, the second inequality in (3) follows from [8, Lemma 2], and (4) follows from computing the size of $|S_k|$. Now, by equating (4) with $\delta$ and solving for $\epsilon$, we obtain

$$
\epsilon = \sqrt{\frac{(k+1)C_{\max}^2}{2(n-2k)}} \log \left( \frac{k+1}{\delta} \right) + |Z|^{2k} \log \left( \frac{|S|}{\delta} \right).
$$

(5)

Thus, we have proven the proposition. ■

Now, we introduce some additional notations. First, denote $\mathcal{W}_k^{L,N}$ as the set of the parameters of the neural networks that has the architecture in Figure 1(a), namely, take $c \in \mathbb{C}[k]$ as input, have $L$ layers and $N$ nodes. Note $N = \sum_{\ell=1}^{L} n_\ell$. We assume that all the weight parameters in $\mathcal{W}_k^{L,N}$ are bounded by some constant $B < \infty$, and the parameters of Neural DUDE are also in $\mathcal{W}_k^{L,N}$.

For each $w \in \mathcal{W}_k^{L,N}$ and the context $c \in \mathbb{C}[k]$, we denote

$$
s_k[w](c, \cdot) = \arg \max_{s \in S} p^k(w, c) s
$$

as the single-symbol denoiser that is defined by the context $c$, the network parameters $w$, and the neural network architecture $p^k(w, \cdot)$. Note the notation in (6) highlights the dependency on $w$. Then, we can rewrite the definition of Neural DUDE for location $i$ as

$$
s_k,\text{Neural DUDE}(C_i, \cdot) = s_k[w^*](C_i, \cdot).
$$

(7)

With above definitions and notations, the following theorem summarizes our main result.

Theorem 2 Let $\delta, \gamma > 0$. Then, for all $x^n \in \mathcal{X}^n$, with probability at least $1 - \delta$,

$$
\sup_{w \in \mathcal{W}_k^{L,N}} \frac{1}{n} \sum_{i=1}^{n} \left[ \mathcal{L}(Z_i, s_k[w](C_i, \cdot)) - \Lambda(x_i, s_k[w](C_i, Z_i)) \right] \leq \sqrt{2}C_{\max}^2 \gamma |S|^2 + \frac{C_{\max}(2B)^L \sqrt{|Z||S|}}{\gamma} \left( \prod_{\ell=1}^{L} n_\ell \right) \frac{k}{\sqrt{n}} \frac{(2k+1)C_{\max}}{n} \sqrt{\frac{2\log(2/\delta)}{n}},
$$

(8)

in which $C_{\max} \triangleq \max_{z,s} |\mathcal{L}(z,s)| + \max_{z,\hat{z}} |\Lambda(x, \hat{z})|$.}

Remark: From the theorem, for sufficiently small $\gamma$, we have the concentration of the average estimated loss to the average true loss with high probability, uniformly for all $w \in \mathcal{W}_k^{L,N}$ and $x^n \in \mathcal{X}^n$, provided that $k = o(\sqrt{n})$. Therefore, the theorem backs the claim that [1] holds for Neural DUDE for much larger $k$’s for fixed $n$ than those of DUDE, as has been observed in Figure 1(c).
4 Proof of the Main Theorem

4.1 Additional notations and definitions

Before proving the theorem, we introduce a few more notations and definitions. First, for any \( x \in \mathcal{X} \), \( z \in \mathcal{Z} \) and \( s \in \mathcal{S} \), we define the per-symbol regret as \( r_{(x,z)}[s] \triangleq L(z,s) - \Lambda(x,s(z)) \). Moreover, for schemes that determine the single-symbol denoiser by finding the maximum argument of a probability vector \( p \in \Delta^{|S|} \) as in Neural DUDE, we also denote \( r_{(x,z)}(p) \triangleq r_{(x,z)}[S(p)] \) as the regret function (in \( p \)), by defining \( S(p) = s \) if \( p \in D_s \triangleq \{ p \in \Delta^{|S|} : s = \arg \max_{s' \in S} p_{s'} \} \). Namely, \( D_s \) is the decision region for \( s \).

With above notations, we can express the \( i \)-th per-symbol regret in Theorem 2 as

\[
L(Z_i, s_k[w](C_i, \cdot)) - \Lambda(x_i, s_k[w](C_i, Z_i)) = r_{(x_i, Z_i)}(p^k(w, C_i)) \triangleq r_i[w].
\]

In (9), we introduced the notation \( r_i[w] \) for brevity and to highlight the dependency on \( w \). By denoting \( R_n[w] = \frac{1}{n} \sum_{i=1}^{n} r_i[w] \), it becomes clear that Theorem 2 obtains a bound on \( \sup_{w \in W_k, n} |R_n[w]| \).

One important point to note is that the regret function \( r_{(x,z)}(p) \) defined above is not a continuous function in \( p \) for any \( x \) and \( z \). Due to the technical necessity required in the later analyses, we need to define another function \( \tilde{r}^\gamma_{(x,z)}(p) \), which is a pointwise upper bound on \( r_{(x,z)}(p) \) and is Lipschitz continuous in \( p \). To do that, for each \( s \in \mathcal{S} \), we first let

\[
B(s) \triangleq \{ p \in \Delta^{|S|} : p_s = \max_{s' \neq s} p_{s'} \}
\]

be the set of decision boundaries for \( D_s \) defined above. Furthermore, for each clean-noisy pair \( (x, z) \) and a single-symbol denoiser \( s \), define

\[
\mathcal{N}_{(x,z)}(s) \triangleq \{ s' \in S : \mathcal{B}(s') \cap \mathcal{B}(s) \neq \emptyset \text{ and } r_{(x,z)}[s'] \geq r_{(x,z)}[s] \}
\]

as the set of the neighboring single symbol denoisers that share decision boundaries with \( s \) and has larger per-symbol regret for \( (x, z) \). Then, we define

\[
\mathcal{B}(\mathcal{N}_{(x,z)}(s)) \triangleq \bigcup_{s' \in \mathcal{N}_{(x,z)}(s)} \mathcal{B}(s') \cap \mathcal{B}(s)
\]

as a subset of \( B(s) \) that only contains boundaries between \( s \) and the single-symbol denoisers in \( \mathcal{N}_{(x,z)}(s) \). Note when \( \mathcal{N}_{(x,z)}(s) = \emptyset \), that is, when \( s = \arg \max_{s' \in S} r_{(x,z)}[s'] \), then \( \mathcal{B}(\mathcal{N}_{(x,z)}(s)) = \emptyset \) as well. Then, for small \( \gamma > 0 \), we define the \( \gamma \)-margin for \( (x, z) \) as

\[
U^\gamma_{(x,z)} \triangleq \{ p \in \Delta^{|S|} : \text{dist}(p, \mathcal{B}(\mathcal{N}_{(x,z)}(S(p)))) \leq \gamma \},
\]

in which \( \text{dist}(p, \mathcal{E}) \triangleq \min_{\mathcal{E} \in S} \| p - p' \|_2 \) for a set \( \mathcal{E} \in \Delta^{|S|} \). Also, we define \( \text{dist}(p, \emptyset) = \infty \). In words, \( U^\gamma_{(x,z)} \) is the set of the probability vectors that are within distance \( \gamma \) from the boundaries for the single-symbol denoisers that have larger per-symbol regrets than \( S(p) \). A simple example that describes the defined notations for \( |S| = 3 \) is given in Figure 2(a).

With the definition (13), we now define the interpolated regret function for \( p \in \Delta^{|S|} \)

\[
\tilde{r}^\gamma_{(x,z)}(p) \triangleq r_{(x,z)}(p) + \Delta r^\gamma_{(x,z)}(p),
\]

in which \( \Delta r^\gamma_{(x,z)}(p) \geq 0 \) for all \( p \in \Delta^{|S|} \) and is defined as

\[
\Delta r^\gamma_{(x,z)}(p) = \begin{cases} \text{Lin.Interpol.} \left( r_{(x,z)}[S(p)], \mathcal{N}_{(x,z)}(S(p)) \right) & \text{if } p \in U^\gamma_{(x,z)} \\ 0 & \text{otherwise.} \end{cases}
\]

The notation \( \text{Lin.Interpol.} \left( r_{(x,z)}[S(p)], \mathcal{N}_{(x,z)}(S(p)) \right) \) in (15) stands for the appropriate linear interpolation value between the per-symbol regret value \( r_{(x,z)}[S(p)] \) and the larger values in \( \{ r_{(x,z)}[s] : s \in \mathcal{N}_{(x,z)}(S(p)) \} \), determined by \( \text{dist}(p, \mathcal{B}(\mathcal{N}_{(x,z)}(S(p)))) \). While the specific function form in (15) is not important, the important part is that \( \tilde{r}^\gamma_{(x,z)}(p) \) becomes Lipschitz continuous in \( p \); that is, for the constant \( C_{\max} \triangleq \max_{x,s} |L(z,s)| + \max_{x,\hat{x}} |\Lambda(x, \hat{x})| \), we have

\[
|\tilde{r}^\gamma_{(x,z)}(p) - \tilde{r}^\gamma_{(x,z)}(p')| \leq \frac{C_{\max}}{\gamma} \| p - p' \|_2
\]

for all \( p, p' \in \Delta^{|S|} \). The property (16) becomes necessary in proving Lemma 3 below and Theorem 2. The example of \( \tilde{r}^\gamma_{(x,z)}(p) \) on \( \Delta^{|S|} \) is given in Figure 2(b).
With (14), we define the interpolated per-symbol regret as
\[ \tilde{r}_i^\gamma[w] \triangleq \tilde{r}_i^\gamma(x_i, Z_i)[p_k(w, C_i)] \]
and the corresponding Lipschitz-continuous average regret as \( \tilde{R}_n^\gamma[w] = \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^\gamma[w] \). Since \( \tilde{r}_i^\gamma(x_i, Z_i)[p] \geq r_i^\gamma(x_i, Z_i)[p] \) for all \( (x, z) \) and \( p \), we have \( \tilde{R}_n^\gamma[w] \geq R_n[w] \) a.s. for all \( w \in \mathcal{W}_k^{L,N} \). Thus, we have the upper bound
\[
\sup_{w \in \mathcal{W}_k^{L,N}} R_n[w] \leq \sup_{w \in \mathcal{W}_k^{L,N}} \tilde{R}_n^\gamma[w] \triangleq G_n \quad \text{a.s.} \tag{18}
\]
Note the bound in (18) is in the almost sure sense since both \( R_n[w] \) and \( \tilde{R}_n^\gamma[w] \) are random variables. In (18), we defined \( G_n \) to denote the right-hand side of the inequality. Now, before proving our theorem, we present three lemmas which are essentially on bounding \( G_n \). The full proof of the lemmas are given in the Supplementary Material, and we only give the high-level proof ideas below.

**Lemma 1** For \( \epsilon > 0 \) and \( C_{\text{max}} \triangleq \max_{z,s} |L(z, s)| + \max_{x,\tilde{x}} |A(x, \tilde{x})| \), we have
\[
\Pr\left( G_n - \mathbb{E}(G_n) > \epsilon \right) \leq \exp\left( -\frac{n\epsilon^2}{2(2k+1)^2C_{\text{max}}^2} \right). \tag{19}
\]

**Proof:**
The lemma directly follows from the McDiarmid’s inequality\[^9\]. That is, since we assume in the universal setting that the source sequence \( x^n \) is an *individual sequence*, we highlight that the randomness in \( G_n \) is determined by the random variables \( (Z_1, \ldots, Z_n) \) as follows:
\[
G_n = \sup_{w \in \mathcal{W}_k^{L,N}} \tilde{R}_n^\gamma[w] \triangleq g(Z_1, \ldots, Z_n). \tag{20}
\]
Then, by considering a different noisy sequence \( (Z_1, \ldots, Z'_j, \ldots, Z_n) \) which is identical to \( (Z_1, \ldots, Z_n) \) except for the \( j \)-th location replaced with \( Z'_j \), we have the following bound on the difference:
\[
\left| g(Z_1, \ldots, Z_j, \ldots, Z_n) - g(Z_1, \ldots, Z'_j, \ldots, Z_n) \right| \leq \left| \sup_{w \in \mathcal{W}_k^{L,N}} \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^\gamma[w] \right) - \sup_{w \in \mathcal{W}_k^{L,N}} \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^\gamma[w] + \frac{1}{n} \sum_{i=j-k}^{j+k} \{ \tilde{r}_i^\gamma[w] - \tilde{r}_i^\gamma[w] \} \right) \right| \tag{21}
\]
\[
\leq \sup_{w \in \mathcal{W}_k^{L,N}} \left( \frac{1}{n} \sum_{i=j-k}^{j+k} \{ \tilde{r}_i^\gamma[w] - \tilde{r}_i^\gamma[w] \} \right) \leq \frac{2(2k+1)}{n} C_{\text{max}}, \tag{22}
\]
Then, for \((29)\) is from the bound \((21)\), \((21)\) follows from the fact that \(\tilde{r}_i^{\gamma}[w] \) in [Manuscript, Eq.(14)] is only affected by the \((2k+1)\)-tuple, \(Z_{i+k}^{\gamma}\), the first inequality in \((22)\) follows from separately applying the supremum for the second term, and the second inequality in \((22)\) follows from the fact that for all \(w \in \mathcal{W}_k^{L,N}\),

\[
\left| \frac{1}{n} \sum_{i=j-k}^{j+k} \{ \tilde{r}_i^{\gamma}[w] - \tilde{r}_i^{\gamma}[w] \} \right| \leq \frac{1}{n} \sum_{i=j-k}^{j+k} \{ |\tilde{r}_i^{\gamma}[w]| + |\tilde{r}_i^{\gamma}[w]| \} \leq \frac{2(2k+1)}{n} C_{\max}. \tag{23}
\]

Now, since the noisy observations \((Z_1, \ldots, Z_n)\) are independent random variables given the underlying clean source \((x_1, \ldots, x_n)\) from the memoryless channel assumption, we can directly apply the McDiarmid’s inequality and obtain the lemma. \(\blacksquare\)

**Lemma 2** Let \(c = (z_{-k}^1, z_k^1) \in \mathbb{C}[k]\) and let \(A \triangleq \{(x, z_0, c) \rightarrow \tilde{r}_i^{\gamma}(x, z_0)(p^k(w, c)): w \in \mathcal{W}_k^{L,N}\}\) be the function class of the interpolated per-symbol regret functions, parameterized by \(w \in \mathcal{W}_k^{L,N}\). Then, for \(\gamma > 0\) used in \((13)\) and \(C_{\max}\) defined in Lemma 7 we have

\[
\mathbb{E}(G_n) \leq 2\mathcal{R}_n(A) + \sqrt{2}C_{\max} \gamma |S|^2, \tag{24}
\]

in which \(\mathcal{R}_n(A)\) is the Rademacher complexity of \(A\) defined as

\[
\mathcal{R}_n(A) \triangleq \mathbb{E}\left( \sup_{w \in \mathcal{W}_k^{L,N}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{r}_i^{\gamma}[w] \right). \tag{25}
\]

In \((25)\), \(\{\sigma_i\}^n_{i=1}\) are drawn i.i.d. uniform over \(\{+1, -1\}\), and \(\tilde{r}_i^{\gamma}[w]\) is as defined in \((17)\).

**Proof:** First, by referring to the notations [Manuscript, Eq.(6)] and [Manuscript, Eq.(14)], we have

\[
\tilde{r}_i^{\gamma}[w] = r_i[w] + \Delta r_i^{\gamma}[w] \tag{26}
\]

by defining \(\Delta r_i^{\gamma}[w] \triangleq \Delta r_i^{\gamma}(x_i, z_i)(p^k(w, C_i))\). Then, we have the following inequalities:

\[
\mathbb{E}(G_n) = \mathbb{E}\left( \sup_{w \in \mathcal{W}_k^{L,N}} \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^{\gamma}[w] \right) \tag{27}
\]

\[
= \mathbb{E}\left( \sup_{w \in \mathcal{W}_k^{L,N}} \left[ \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^{\gamma}[w] - \mathbb{E}\left( \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^{\gamma}[w] \right) \right] + \mathbb{E}\left( \frac{1}{n} \sum_{i=1}^{n} \Delta r_i^{\gamma}[w] \right) \right) \tag{28}
\]

\[
\leq \mathbb{E}\left( \sup_{w \in \mathcal{W}_k^{L,N}} \left[ \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^{\gamma}[w] - \mathbb{E}\left( \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^{\gamma}[w] \right) \right] \right) + \sqrt{2}C_{\max} \gamma |S|^2, \tag{29}
\]

in which \((28)\) follows from \((22)\) and the fact

\[
\mathbb{E}(r_i[w]) = \mathbb{E}\left( \mathbb{L}(Z_i, s_k[w]|C_i, \cdot) - \mathbf{A}(x_i, s_k[w]|C_i, Z_i) \right) = 0 \tag{30}
\]

for all \(i\), since \((30)\) is from the fact that \(\mathbb{L}(Z, s)\) is an unbiased estimate of \(\mathbb{E}\mathbf{A}(x, s(Z))\) as defined in [Manuscript, Section 2.1]. Thus, \(\mathbb{E}(\tilde{r}_i^{\gamma}[w]) = \mathbb{E}(\Delta r_i^{\gamma}[w])\) for all \(i\), and we have \((28)\). Furthermore, \((29)\) is from the bound

\[
\mathbb{E}(\Delta r_i^{\gamma}[w]) \leq \sqrt{2}C_{\max} \gamma |S|^2, \text{ for all } i \tag{31}
\]

since \(\sqrt{2}C_{\max} \gamma\) is the crude upper bound on the maximum contribution to the expectation at each decision boundary, and there are no more than \(|S|^2\) decision boundaries in the simplex.

Now, similarly as in the arguments for the generalization bound and Rademacher complexity in learning theory [10], we introduce a ghost noisy observation sequence \((Z_1^1, \ldots, Z_n^1)\), which is an independent realization of the noisy observations given the same underlying clean source sequence
We will show the bound with consecutive compositions of the layers. First, for

\[ \ell \]

we have

\[ \sup_{w \in W_\ell} \left[ \frac{1}{n} \sum_{i=1}^n \bar{r}_i^\gamma[w] - E\left( \frac{1}{n} \sum_{i=1}^n \bar{r}_i^\gamma[w] \right) \right] \]

\[ = E\left( \sup_{w \in W_\ell} \left[ \frac{1}{n} \sum_{i=1}^n \bar{r}_i^\gamma[w] - E\left( \frac{1}{n} \sum_{i=1}^n \bar{r}_i^{\gamma*}[w] \right) \right] \right) \]

\[ = E\left( \sup_{w \in W_\ell} E\left( \frac{1}{n} \sum_{i=1}^n \bar{r}_i^\gamma[w] - \frac{1}{n} \sum_{i=1}^n \bar{r}_i^{\gamma*}[w] \bigg| Z^n \right) \right) \]

\[ \leq E\left( \sup_{w \in W_\ell} \frac{1}{n} \sum_{i=1}^n \left[ \bar{r}_i^\gamma[w] - \bar{r}_i^{\gamma*}[w] \right] \right) \]

in which (32) is from the fact that both \( \frac{1}{n} \sum_{i=1}^n \bar{r}_i^\gamma[w] \) and \( \frac{1}{n} \sum_{i=1}^n \bar{r}_i^{\gamma*}[w] \) follow the identical distribution, (33) follows from the fact that \( (Z_1, \ldots, Z_n) \) and \( (Z_1^\gamma, \ldots, Z_n^\gamma) \) are independent, (34) follows from pushing the supremum inside the expectation, and (35) follows from carrying out the iterated conditional expectation.

Now, we introduce the Rademacher variables \( \sigma_1, \ldots, \sigma_n \) that are independent of \( (Z_1, \ldots, Z_n) \) and \( (Z_1^\gamma, \ldots, Z_n^\gamma) \), and each \( \sigma_i \) is i.i.d and uniform over \( \{ +1, -1 \} \). Then, continuing the inequalities from (35) yields

\[ E\left( \sup_{w \in W_\ell} \frac{1}{n} \sum_{i=1}^n \sigma_i \left[ \bar{r}_i^\gamma[w] - \bar{r}_i^{\gamma*}[w] \right] \right) \]

\[ = E\left( \frac{1}{n} \sum_{i=1}^n \sigma_i \bar{r}_i^\gamma[w] \right) + \sup_{w \in W_\ell} \frac{1}{n} \sum_{i=1}^n \sigma_i \bar{r}_i^{\gamma*}[w] \]

\[ = 2E\left( \sup_{w \in W_\ell} \frac{1}{n} \sum_{i=1}^n \sigma_i \bar{r}_i^\gamma[w] \right) \]

\[ = 2R_n(A) \]

in which (36) follows from the fact that the distribution of \( \bar{r}_i^\gamma[w] - \bar{r}_i^{\gamma*}[w] \) is symmetric around 0 and multiplying \( \sigma_i \) does not change the distribution, (37) holds by the inequality \( \sup_{a \in \mathbb{R}} \{ a[w] - b[w] \} \leq \sup_{a \in \mathbb{R}} a[w] + \sup_{a \in \mathbb{R}} \{ -b[w] \} \), (38) follows from the linearity of expectation and the fact that \( \sigma_i \) and \( -\sigma_i \) have the same distribution, and (39) is from the definition of the Rademacher complexity (25).

Now, by combining (39) and (29), the lemma is proven. \( \square \)

**Lemma 3** For \( \gamma > 0 \) used in (13), we have the following bound on \( \mathcal{R}_n(A) \) in (25):

\[ \mathcal{R}_n(A) \leq \frac{C_{\max}(2B)^L \sqrt{|Z||S|}}{\gamma} \left( \prod_{\ell=1}^L n_\ell \right)^{\frac{k}{\sqrt{n}}} \]

**Proof:** First, note that the definition of the Rademacher complexity in [Manuscript, Eq.(18)] is slightly different from the ordinary definition in [10, Eq.(218)] in that the summands \( \bar{r}_i^\gamma[w] \)'s are not independent due to the overlapping contexts. However, we can still utilize the general tools of Rademacher complexity to obtain the bound [Manuscript, Eq.(19)] of the lemma.

We will show the bound with consecutive compositions of the layers. First, for \( \ell \geq 2 \), we denote \( w_{\ell,m} \in \mathbb{R}^{n_{\ell-1}} \) as the weight parameter vector associated with the \( m \)-th node in the \( \ell \)-th layer. For \( \ell = 1 \), we have \( w_{1,m} \in \mathbb{R}^{2|Z|^k} \) as the input to the network, \( C_i = (Z_{i-k}, Z_{i+k}) \), has dimension \( 2|Z|^k \), since each noisy symbol is one-hot encoded with dimension \( |Z| \).
From our assumption that each weight parameter in $W_{k}^{L,N}$ is bounded by $B$, we have $\|w_{1,j}\|_2 \leq \sqrt{2|Z|k}B$. Furthermore, we have $\|C_i\|_2 \leq \sqrt{2k}$ a.s. due to the one-hot encoding of the noisy symbol. Since we are using the ReLU function, $f(x) = \max\{0, x\}$, as the nonlinearity in the network, we first consider the Rademacher complexity of the function class $F = \{(x, z_0, c) \mapsto w^Tc : \|w\|_2 \leq \sqrt{2|Z|k} \cdot B\}$ in which $c = (z_{k}^{-1}, z_{k}^{1}) \in C[k]$. That is, we have

$$\mathcal{R}_n(F) = \frac{1}{n} \mathbb{E} \left( \sup_{\|w\|_2 \leq \sqrt{2|Z|k} \cdot B} \sum_{i=1}^{n} \sigma_i (w^T C_i) \right)$$

(41)

$$\leq \frac{\sqrt{2|Z|k} \cdot B}{n} \mathbb{E} \left( \sum_{i=1}^{n} \|\sigma_i C_i\|_2^2 \right) \leq \frac{\sqrt{2|Z|k} \cdot B}{n} \mathbb{E} \left( \sum_{i=1}^{n} \|C_i\|_2^2 \right)$$

(42)

$$\leq \frac{\sqrt{2|Z|k} \cdot B}{n} \sqrt{2kn} = (2B)\sqrt{|Z|} \frac{k}{\sqrt{n}}$$

(43)

in which the first inequality of (42) follows from applying Cauchy-Schwartz inequality to $w$ and $\sum_{i=1}^{n} \sigma_i C_i$, the second inequality of (42) follows from the concavity of $\sqrt{\cdot}$, the inequality of (43) follows from the independence of $\sigma_i$, the equality of (43) follows from the fact that $\sigma_i$ does not affect the norm, and (44) follows the bound on $\|C_i\|_2$. Note that the independence of $C_i$ is not required for the bounds. Now, from the Lipschitz composition property of the Rademacher complexity [10, Theorem 11], we upper bound the Rademacher complexity of the hidden node in the first layer as

$$\mathcal{R}_n\left(\{(x, z_0, c) \mapsto f(w^T c) : \|w\|_2 \leq \sqrt{2|Z|k} \cdot B\}\right) \leq (2B)\sqrt{|Z|} \frac{k}{\sqrt{n}}$$

(45)

since $f(x) = \max\{0, x\}$ is 1-Lipschitz continuous.

Continuing to the second layer, by denoting $h_1 = [f(w_{1,1}^Tc), \ldots, f(w_{1,m_1}^Tc)]^T \in \mathbb{R}^{m_1}$ as the vector of the first hidden layer node values for the input $c \in C[k]$, the value of the $m$-th node in the second layer is $f(w_{2,m}h_1)$ with $\|w_{2,m}\|_2 \leq \sqrt{m_1}B$. Then, by following the similar argument as in [10, Theorem 43], we have

$$\mathcal{R}_n\left(\{(x, z_0, c) \mapsto f(w_{2,m}^T h_1) : \|w_{2,m}\|_2 \leq \sqrt{m_1}B\}\right) \leq (2B)^2 \sqrt{|Z|} \frac{k}{\sqrt{n}}$$

(46)

again by the Lipschitz continuity of $f(x)$ and the fact that $f(0) = 0$.

By continuing above argument, we can see that the Rademacher complexity of the $s$-th node in the final output layer, $o_s$, before the softmax function can be bounded by

$$(2B)^{L+1} \sqrt{|Z|} \left( \prod_{\ell=1}^{L} n_\ell \right) \frac{k}{\sqrt{n}}.$$  

(47)

Since we can show that the softmax function $o \mapsto p = \text{softmax}(o_1, \ldots, o_{|S|})$ is $\frac{|S|}{2}$-Lipschitz, we have

$$\mathcal{R}_n\left(\{(x, z_0, c) \mapsto p^k(w, c) : w \in W_{k}^{L,N}\}\right) \leq (2B)^L \sqrt{|Z||S|} \left( \prod_{\ell=1}^{L} n_\ell \right) \frac{k}{\sqrt{n}}.$$  

(48)

Finally, since $p^\gamma_{(x,z)}(p)$ in [Manuscript, Eq.(11)] is designed to be $\frac{|S|}{2}$-Lipschitz as in [Manuscript, Eq.(13)], we can apply the Lipschitz composition property one more time and have the bound in the lemma.

4.3 Proof of Theorem[2]

We now prove the main theorem. First, we have

$$\mathbb{P}\left( \sup_{w \in W_{k}^{L,N}} |R_n(w) | \geq \epsilon \right) \leq \mathbb{P}\left( \sup_{w \in W_{k}^{L,N}} R_n(w) \geq \epsilon \right) + \mathbb{P}\left( \sup_{w \in W_{k}^{L,N}} (-R_n(w)) \geq \epsilon \right)$$

(49)
by applying the union bound. Now, from the pointwise upper bound property in [18], we have

$$\Pr(G_n \geq \epsilon) \leq \exp \left( -\frac{n(\epsilon - \mathbb{E}(G_n))^2}{2(2k+1)^2C_{\max}^2} \right) \leq \exp \left( -\frac{n(\epsilon - (2R_n(A) + \sqrt{2C_{\max}^2}|S|^2))^2}{2(2k+1)^2C_{\max}^2} \right) \leq \frac{\delta}{2}$$

where the inequality [52] holds for $\epsilon \geq 2R_n(A) + \sqrt{2C_{\max}^2}|S|^2$. Now, for the second term in (49), we can analogously define a function, $\tilde{r}_i^j[w]$, that upper bounds the negated per-symbol regret function, $-r_i[w]$, as in [14] and [17]. Then, we obtain the identical result as Lemma 2 for the function class of interpolated per-symbol negated regret functions. That is, with defining $G_n \equiv \sup_{w \in W} \left\{ \frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i^j[w] \right\}$, and $A \equiv \{ (x, z_0, c) \to \tilde{r}_i^j(x, z_0) (p^k(x, w, c)) : w \in W_k \}$ as the function class for the negated regret function, we obtain $\Pr(G_n \geq \epsilon) \leq 2R_n(A) + \sqrt{2C_{\max}^2}|S|^2$, via the exact same derivation in Lemma 2. Since it is clear that $R_n(A)$ and $R_n(A)$ have the same upper bound, we obtain (The second term in (49)) $\leq \frac{\delta}{2}$, in which $\delta$ is as defined in [52]. By solving for $\epsilon$, we obtain

$$\epsilon = 2R_n(A) + \sqrt{2C_{\max}^2}|S|^2 + (2k+1)C_{\max} \sqrt{\frac{2\log(2/\delta)}{n}},$$

and by plugging in the result of Lemma 3 in [53], we have proved the theorem. ■

5 Experimental Results

In this section, we show the experimental results on binary image denoising and DNA sequence denoising that corroborate the theoretical result of the paper. The basic experimental settings are identical to those of [4].

For the binary image denoising, we took the standard Lena binary image of size $512 \times 512$ as clean image and obtained a noisy image, in which the clean image is corrupted by the binary symmetric channel (BSC) with crossover probability $\delta = 0.1$. We ran both DUDE and Neural DUDE for much larger $k$ values (up to $k = 500$) than those in [4] Section 5 in order to validate the sufficient condition for the concentration obtained from Theorem 2. Furthermore, unlike [4], which first raster scanned the image and utilized the 1-D contexts, we utilized the 2-D contexts as in [11].

Figure 3: (a) Denoising results of binary Lena image ($512 \times 512$) with $\delta = 0.1$. (b),(c) The concentration gap $|R_n[w]|$ of Neural DUDE with varying number of layers ($L$) and nodes ($N$).

Figure 3(a) shows the true BER of DUDE (red) and Neural DUDE (blue) as well as the estimated BER of Neural DUDE (purple). Neural DUDE used 4 layers including the output layer, and 40 nodes.

---

2The boundary was zero-padded such that the denoising can be done at all locations.
were used for each layer. Firstly, we can clearly observe that the performance of Neural DUDE is very robust to the choice of $k$ even for much larger $k$ values than ones in [4]. Secondly, the overall best BER of Neural DUDE is $0.213\delta$ when $k = 6$. The relative improvement over DUDE is much greater than that of [4], which only tested with $k$ up to 40 and with 1-D contexts. Finally, we note that the estimated BER of Neural DUDE does concentrate on the true BER even for large $k$’s. (The estimated BER of DUDE dramatically diverges as in Figure 1(b) so we omitted it in the plot.) Note when $k = 200$, there are $2^{400}$ possible contexts, which is much larger than the data size $n = 2^{18}$, but the concentration still holds since $k < \sqrt{n}$, as has been suggested by Theorem 2. Moreover, once we select the window size $k$ that minimizes the estimated BER, we select $k = 140$ resulting in the true BER of $0.225\delta$, which is just 5% greater than the overall best BER, $0.213\delta$. Hence, the concentration property proven in our theorem certainly justifies the $k$ selection based on the estimated loss. Figure 3(b) and 3(c) show the effects of the number of layers ($L$) and nodes ($N = \sum_{\ell=1}^{L} n_{\ell}$) on the concentration gap $|R_{\alpha_k}(w^*)|$. From both figures, we can observe that although the dependency of the upper bound (8) on $\|Z\|$ and $|S|$ is high in Theorem 2, the true gap does not appear as much on $L$ and $N$ in this example. Furthermore, for all models in the two figures, the gap gradually increases as we increase $k$, which again is consistent with the theory.

For the DNA sequence denoising, we carried out the same experiments as in [4, Section 5.3] with the MinION Nanopore reads with sequence length $n = 2, 469, 111$. For details on the experimental setting, we refer the readers to [4]. Figure 4 shows the true and estimated error rates of Neural DUDE (with $L = 3$) with varying number of nodes. We can observe that the concentration gap increases as the number of nodes increases for fixed $k$. Note the amount of increase is larger than that of Figure 3(c) and we believe this is due to the dependency of the upper bound (8) on $|Z|$ and $|S|$. Furthermore, when $k$ gets close to $\sqrt{n} = 1571$, the concentration gap increases regardless of the number of nodes, which again can be predicted by Theorem 2. However, when the hyperparameters (window size $k$ and the number of nodes $N$ in this case) are chosen by examining the estimated errors, we can still choose the 160-160-160 model with $k = 250$ and achieve the error rate $\sim 0.4\delta$ that gets close to the overall best error rate $\sim 0.36\delta$ (160-160-160 model with $k = 125$). Thus, the hyperparameter selection method motivated from the concentration property still remains practical in the DNA sequence denoising.

6 Concluding Remarks and Future Work

In this paper, we rigorously showed the uniform concentration property of the estimated loss on the true loss for the neural network-based sliding window discrete denoisers. The result justifies the hyperparameter selection method for Neural DUDE proposed in [4]. We utilized techniques from statistical learning theory in a novel way; i.e., while the uniform concentration of the training error on the test error is the focus of the typical learning theory, we showed the uniform concentration of the estimated loss on the true loss for the denoising setting by modifying the arguments using Rademacher complexity. In the future, we plan to find more real-world applications for Neural DUDE and extend the scheme both algorithmically and theoretically to the case of continuous-valued data.
References

[1] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger. Universal discrete denoising: Known channel. *IEEE Trans. on Inform. Theory*, 51(1):5–28, 2005.

[2] E. Ordentlich, G. Seroussi, S. Verdú, M.J. Weinberger, and T. Weissman. A universal discrete image denoiser and its application to binary images. In *IEEE ICIP*, 2003.

[3] B. Lee, T. Moon, S. Yoon, and T. Weissman. DUDE-Seq: Fast, flexible, and robust denoising of nucleotide sequences. *arXiv:1511.04836*, 2016.

[4] T. Moon, S. Min, B. Lee, and S. Yoon. Neural universal discrete denoisier. In *NIPS*, 2016.

[5] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[6] T. Weissman, E. Ordentlich, M. Weinberger, A. Somekh-Baruch, and N. Merhav. Universal filtering via prediction. *IEEE Trans. Inform. Theory*, 53(4):1253–1264, 2007.

[7] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[8] T. Moon and T. Weissman. Discrete denoising with shifts. *IEEE Trans. Inform. Theory*, 55(11):5284–5301, 2009.

[9] C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

[10] P. Liang. CS229T/STAT231: Statistical learning theory. *Stanford CS229T Lecture Note*, 2016.

[11] T. Moon, T. Weissman, and J. Kim. Discrete denoising of heterogeneous two-dimensional data. In *Proceedings of 2011 IEEE International Symposium on Information Theory (ISIT)*, pages 966–970, August 2011.