Supplementary Material for Unsupervised Neural Universal Denoiser for Finite-Input General-Output Noisy Channel

Tae-Eon Park and Taesup Moon
Department of Electrical and Computer Engineering
Sungkyunkwan University (SKKU), Suwon, Korea 16419
{pte1236, tsmoon}@skku.edu

Appendix A Proof of Theorem 1

The following lemma formalizes to prove Eq. (16) in the paper. First, let $\epsilon^* = \epsilon' \sum_{a=0}^{M-1} \|\pi_a^{-1}\|_2$ as shown in the paper and define

$$\hat{\mathbf{P}}(X_0|y_{-k}^k) \triangleq \frac{p(\mathbf{y}_{-0}^{(k)})}{p(y_{-k}^k)} \cdot [\mathbf{p}^k(\mathbf{w}^*, \mathbf{y}_{-0}^{(k)}) \cdot \mathbf{\Pi}^{-1}] \odot \mathbf{f}_{X_0}(y_0). \tag{1}$$

Lemma 2 *Suppose network parameter* $\mathbf{w}^*$ *learned by minimizing* $\mathcal{L}_{\texttt{Gen-CUDE}}$ *satisfies Assumption 2. Then,*

$$\mathbb{E}\|\mathbf{P}(X_0|Y_{-k}^k) - \hat{\mathbf{P}}(X_0|Y_{-k}^k)\|_1 \le \epsilon^*,$$

in which the expectation is with respect to Y_{-k}^k*.*

Proof: We have the following chain of equations.

$$\mathbb{E}\|\mathbf{P}(X_0|Y_{-k}^k) - \hat{\mathbf{P}}(X_0|Y_{-k}^k)\|_1 = \int_{\mathbb{R}^{2k+1}} p(y_{-k}^k) \cdot \|\mathbf{P}(X_0|y_{-k}^k) - \hat{\mathbf{P}}(X_0|y_{-k}^k)\|_1 \, dy_{-k}^k$$

$$= \int_{\mathbb{R}^{2k+1}} p(y_{-k}^k) \cdot \left\| \frac{p(\mathbf{y}_{-0}^{(k)})}{p(y_{-k}^k)} \cdot [\Big(\mathbf{P}(Z_0|\mathbf{y}_{-0}^{(k)}) - \mathbf{p}^k(\mathbf{w}^*, \mathbf{y}_{-0}^{(k)})\Big) \cdot \mathbf{\Pi}^{-1}] \odot \mathbf{f}_{X_0}(y_0) \right\|_1 dy_{-k}^k \tag{2}$$

$$= \int_{\mathbb{R}^{2k+1}} p(y_{-k}^k) \cdot \Big[\sum_{a=0}^{M-1} \Big| \Big(\mathbf{P}(Z_0|\mathbf{y}_{-0}^{(k)}) - \mathbf{p}^k(\mathbf{w}^*, \mathbf{y}_{-0}^{(k)})\Big) \cdot \pi_a^{-1} \cdot f_a(y_0) \Big| \Big] \cdot \frac{p(\mathbf{y}_{-0}^{(k)})}{p(y_{-k}^k)} \, dy_{-k}^k$$

$$= \sum_{a=0}^{M-1} \int_{\mathbb{R}^{2k+1}} \Big| \Big(\mathbf{P}(Z_0|\mathbf{y}_{-0}^{(k)}) - \mathbf{p}^k(\mathbf{w}^*, \mathbf{y}_{-0}^{(k)})\Big) \cdot \pi_a^{-1} \Big| \cdot f_a(y_0) \cdot p(\mathbf{y}_{-0}^{(k)}) \, dy_{-k}^k$$

$$\le \sum_{a=0}^{M-1} \int_{\mathbb{R}^{2k+1}} \Big\| \mathbf{P}(Z_0|\mathbf{y}_{-0}^{(k)}) - \mathbf{p}^k(\mathbf{w}^*, \mathbf{y}_{-0}^{(k)}) \Big\|_2 \cdot \|\pi_a^{-1}\|_2 \cdot f_a(y_0) \cdot p(\mathbf{y}_{-0}^{(k)}) \, dy_{-k}^k \tag{3}$$

$$\le \sum_{a=0}^{M-1} \int_{\mathbb{R}^{2k+1}} \Big\| \mathbf{P}(Z_0|\mathbf{y}_{-0}^{(k)}) - \mathbf{p}^k(\mathbf{w}^*, \mathbf{y}_{-0}^{(k)}) \Big\|_1 \cdot \|\pi_a^{-1}\|_2 \cdot f_a(y_0) \cdot p(\mathbf{y}_{-0}^{(k)}) \, dy_{-k}^k \tag{4}$$

$$\le \epsilon' \cdot \sum_{a=0}^{M-1} \Big[\|\pi_a^{-1}\|_2 \cdot \int_{\mathbb{R}^{2k+1}} f_a(y_0) \cdot p(\mathbf{y}_{-0}^{(k)}) \, dy_{-k}^k \Big] \tag{5}$$

$$= \epsilon' \cdot \sum_{a=0}^{M-1} \Big[\|\pi_a^{-1}\|_2 \cdot \int_{\mathbb{R}} f_a(y_0) \int_{\mathbb{R}^{2k}} p(\mathbf{y}_{-0}^{(k)}) \, d\mathbf{y}_{-0}^{(k)} \, dy_0 \Big] = \epsilon' \cdot \sum_{a=0}^{M-1} \|\pi_a^{-1}\|_2 = \epsilon^*, \tag{6}$$

in which (2) follows from Lemma 1 and (1), (3) follows from the Cauchy-Schwarz inequality, and (4) follows from the fact that L_2-norm is smaller than the L_1-norm, and (5) follows from Assumption 2. ■

Lemma 3 *Let $\mathcal{R}_\delta(\cdot)$ denote the quantizer that rounds each component of the argument probability vector to the nearest integer multiple of δ in $(0,1]$. For $M>0$ and $\delta>0$, denote $\hat{\mathbf{P}}^\delta = \mathcal{R}_\delta(\hat{\mathbf{P}})$. Then,*

$$\left\|\hat{\mathbf{P}}^\delta\left(X_0|y_{-k}^k\right) - \hat{\mathbf{P}}\left(X_0|y_{-k}^k\right)\right\|_1 \le \frac{M\cdot\delta}{2}.$$

Proof: By the definition of $\hat{\mathbf{P}}^\delta$, it is clear that

$$\left\|\hat{\mathbf{P}}^\delta\left(X_0|y_{-k}^k\right) - \hat{\mathbf{P}}\left(X_0|y_{-k}^k\right)\right\|_\infty \le \frac{\delta}{2}.$$

Therefore, $\left\|\hat{\mathbf{P}}^\delta\left(X_0|y_{-k}^k\right) - \hat{\mathbf{P}}\left(X_0|y_{-k}^k\right)\right\|_1 = \sum_{a=0}^{M-1}\left|\hat{p}^\delta\left(a|y_{-k}^k\right) - \hat{p}(a|y_{-k}^k)\right| \le \sum_{a=0}^{M-1}\frac{\delta}{2} \le \frac{M\delta}{2}$. ■

From Lemma 3, we can expect that for the sufficiently small δ, performance of the denoisers using $\hat{\mathbf{P}}^\delta$ and $\hat{\mathbf{P}}$ respectively, for computing the Bayes response will be close to each other.

Lemma 4 *Consider $\hat{\mathbf{P}}(X_0|Y_{-k}^k)$ and ϵ^* defined in Lemma 2 and the performance target $D_{x^n}^k$ defined in Eq. (15) in the paper. Then, we have*

$$\left|D_{x^n}^k - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\hat{\mathbf{P}}(X_0|Y_{-k}^k)\right)\right]\right| \le \Lambda_{\max}\cdot\epsilon^*,$$

in which $P_{x^n}^k\otimes\mathcal{C}$ stands for the joint distribution on (X_0, Y_{-k}^k) defined by the empirical distribution $P_{x^n}^k(u_{-k}^k) = \frac{1}{n-2k}\mathbf{r}[x^n, u_{-k}^k]$ with $\mathbf{r}[x^n,u_{-k}^k] = |\{k+1\le i\le n-k : x_{i-k}^{i+k} = u_{-k}^k\}|$ and the channel density $\mathcal{C}$.

Proof: First, we identify that

$$\begin{aligned}
&\left|D_{x^n}^k - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\hat{\mathbf{P}}(X_0|Y_{-k}^k)\right)\right]\right| \\
=&\left|\mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\mathbf{P}(X_0|Y_{-k}^k)\right) - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}[U\left(\hat{\mathbf{P}}(X_0|Y_{-k}^k)\right)\right]\right| &(7)\\
=&\int_{\mathbb{R}^{2k+1}}\mathbb{E}\left[\mathbf{\Lambda}\left(X,\mathcal{B}(\hat{\mathbf{P}}(X_0|y_{-k}^k))\right) - \mathbf{\Lambda}\left(X,\mathcal{B}(\mathbf{P}(X_0|y_{-k}^k))\right)\Big|y_{-k}^k\right]\cdot p\left(y_{-k}^k\right)dy_{-k}^k, &(8)
\end{aligned}$$

where the $\mathbb{E}(\cdot)$ in (8) stands for the conditional expectation with respect to $\mathbf{P}(X_0|y_{-k}^k)$, which is the posterior distribution induced from $P_{x^n}^k\otimes\mathcal{C}$. Now, the following inequality holds for each y_{-k}^k:

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{\Lambda}\left(X,\mathcal{B}(\hat{\mathbf{P}}(X_0|y_{-k}^k))\right) - \mathbf{\Lambda}\left(X,\mathcal{B}(\mathbf{P}(X_0|y_{-k}^k))\right)\Big|y_{-k}^k\right] &(9)\\
=&\sum_{a=0}^{M-1}\mathbf{P}(X_0=a|y_{-k}^k)\cdot\left[\mathbf{\Lambda}\left(a,\mathcal{B}(\hat{\mathbf{P}}(X_0|y_{-k}^k))\right) - \mathbf{\Lambda}\left(a,\mathcal{B}(\mathbf{P}(X_0|y_{-k}^k))\right)\right] &(10)\\
\le&\sum_{a=0}^{M-1}\left(\mathbf{P}(X_0=a|y_{-k}^k) - \hat{\mathbf{P}}(X_0=a|y_{-k}^k)\right)\cdot\left[\mathbf{\Lambda}\left(a,\mathcal{B}(\hat{\mathbf{P}}(X_0|y_{-k}^k))\right) - \mathbf{\Lambda}\left(a,\mathcal{B}(\mathbf{P}(X_0|y_{-k}^k))\right)\right] &(11)\\
\le&\sum_{a=0}^{M-1}\left|\left(\mathbf{P}(X_0=a|y_{-k}^k) - \hat{\mathbf{P}}(X_0=a|y_{-k}^k)\right)\right|\cdot\Lambda_{\max} = \Lambda_{\max}\cdot\|\mathbf{P}(X_0|y_{-k}^k) - \hat{\mathbf{P}}(X_0|y_{-k}^k)\|_1, &(12)
\end{aligned}$$

in which (11) follows from the definition of the Bayes response. Therefore,

$$\begin{aligned}
(8)\le&\Lambda_{\max}\cdot\int_{\mathbb{R}^{2k+1}}\|\mathbf{P}(X_0|y_{-k}^k) - \hat{\mathbf{P}}(X_0|y_{-k}^k)\|_1\cdot p\left(y_{-k}^k\right)dy_{-k}^k = \Lambda_{\max}\cdot\mathbb{E}\|\mathbf{P}(X_0|Y_{-k}^k) - \hat{\mathbf{P}}(X_0|Y_{-k}^k)\|_1 &(13)\\
\le&\Lambda_{\max}\cdot\epsilon^*, &(14)
\end{aligned}$$

in which (14) follows from Lemma 2. Note that difference between two expected loss of denoiser based on the Bayes response is bounded with the difference between two probability vectors. ■

Lemma 5 *Consider $\hat{\mathbf{P}}(X_0|Y_{-k}^k)$ and ϵ^* defined above and define $\hat{\mathbf{P}}^\delta = \mathcal{R}_\delta(\hat{\mathbf{P}})$ as in Lemma 3. Then,*

$$\left|\mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\hat{\mathbf{P}}(X_0|Y_{-k}^k)\right) - U\left(\hat{\mathbf{P}}^\delta\left(X_0|Y_{-k}^k\right)\right)\right]\right| \le 2\Lambda_{\max}\cdot\left(\epsilon^* + \frac{M\cdot\delta}{4}\right).$$

Proof: We have the following chain of inequalities:

$$\begin{aligned}
&\left|\mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\hat{\mathbf{P}}\left(X_0|Y_{-k}^k\right)\right)\right] - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\hat{\mathbf{P}}^\delta\left(X_0|Y_{-k}^k\right)\right)\right]\right| \\
\le&\left|\mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\hat{\mathbf{P}}\left(X_0|Y_{-k}^k\right)\right)\right] - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\mathbf{P}\left(X_0|Y_{-k}^k\right)\right)\right]\right| \\
&+\left|\mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\mathbf{P}\left(X_0|Y_{-k}^k\right)\right)\right] - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[U\left(\hat{\mathbf{P}}^\delta\left(X_0|Y_{-k}^k\right)\right)\right]\right| &(15)\\
\le&\,\Lambda_{\max}\cdot\mathbb{E}\|\mathbf{P}(X_0|Y_{-k}^k) - \hat{\mathbf{P}}(X_0|Y_{-k}^k)\|_1 + \Lambda_{\max}\cdot\mathbb{E}\|\mathbf{P}(X_0|Y_{-k}^k) - \hat{\mathbf{P}}^\delta(X_0|Y_{-k}^k)\|_1 &(16)\\
\le&\,2\Lambda_{\max}\cdot\mathbb{E}\|\mathbf{P}(X_0|Y_{-k}^k) - \hat{\mathbf{P}}(X_0|Y_{-k}^k)\|_1 + \Lambda_{\max}\cdot\mathbb{E}\|\hat{\mathbf{P}}(X_0|Y_{-k}^k) - \hat{\mathbf{P}}^\delta(X_0|Y_{-k}^k)\|_1 &(17)\\
\le&\,2\Lambda_{\max}\cdot\mathbb{E}\|\mathbf{P}(X_0|Y_{-k}^k) - \hat{\mathbf{P}}(X_0|Y_{-k}^k)\|_1 + \frac{\Lambda_{\max}\cdot M\cdot\delta}{2} &(18)\\
\le&\,2\Lambda_{\max}\cdot\epsilon^* + \frac{\Lambda_{\max}\cdot M\cdot\delta}{2}, &(19)
\end{aligned}$$

in which (15) follows from the triangular inequality, (16) follows from (13) and replacing $\hat{\mathbf{P}}$ with $\hat{\mathbf{P}}^\delta$ in (7), (17) follows from applying the triangular inequality once more, (18) follows from Lemma 3, and (19) follows from (14). Note that probability vectors $\mathbf{P}, \hat{\mathbf{P}}$ in Lemma 4 replaced $\hat{\mathbf{P}}, \hat{\mathbf{P}}^\delta$ in Lemma 5 respectively. ■

Lemma 6 *For every $n \ge 1$, $x^n \in \mathcal{A}^n$, $\epsilon > 0$ and measurable $g_k : \mathbb{R}^{2k+1} \to \mathcal{A}$,*

$$\Pr\left(\left|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\mathbf{\Lambda}\left(x_i, g_k\left(Y_{i-k}^{i+k}\right)\right) - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[\mathbf{\Lambda}\left(X_0, g_k\left(Y_{-k}^k\right)\right)\right]\right| > \epsilon\right) \le 2(2k+1)\exp\left(-\frac{2\left(n-2k\right)}{(2k+1)}\epsilon^2\cdot\frac{1}{\Lambda_{\max}^2}\right).$$

Proof: We have the following:

$$\begin{aligned}
&\Pr\left(\left|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\mathbf{\Lambda}\left(x_i, g_k\left(Y_{i-k}^{i+k}\right)\right) - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[\mathbf{\Lambda}\left(X_0, g_k\left(Y_{-k}^k\right)\right)\right]\right| > \epsilon\right) \\
=&\,2\cdot\Pr\left(\frac{1}{n-2k}\sum_{m=0}^{2k}\sum_{\substack{i\in\{k+1,\dots,n-k\},\\ \lceil(i-m)/(2k+1)\rceil=(i-m)/(2k+1)}}\mathbf{\Lambda}\left(x_i, g_k\left(Y_{i-k}^{i+k}\right)\right) - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[\mathbf{\Lambda}\left(X_0, g_k\left(Y_{-k}^k\right)\right)\right] > \epsilon\right) \\
\le&\,2(2k+1)\cdot\Pr\left(\frac{2k+1}{n-2k}\sum_{\substack{i\in\{k+1,\dots,n-k\},\\ \lceil i/(2k+1)\rceil=i/(2k+1)}}\mathbf{\Lambda}\left(x_i, g_k\left(Y_{i-k}^{i+k}\right)\right) - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[\mathbf{\Lambda}\left(X_0, g_k\left(Y_{-k}^k\right)\right)\right] > \epsilon\right) &(20)\\
\le&\,2(2k+1)\exp\left(-\frac{2\left(n-2k\right)}{(2k+1)}\epsilon^2\cdot\frac{1}{\Lambda_{\max}^2}\right). &(21)
\end{aligned}$$

Note that if $|i-j| > 2k$, $\mathbf{\Lambda}(x_i, g_k(Y_{i-k}^{i+k}))$ is independent from $\mathbf{\Lambda}(x_j, g_k(Y_{j-k}^{j+k}))$. (20) follows from the union bound, and (21) follows from the fact that $\mathbf{\Lambda}(x_i, g_k(Y_{i-k}^{i+k})) - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}[\mathbf{\Lambda}(X_0, g_k(Y_{-k}^k))]$ is a zero-mean, bounded, independent random variable, and the Hoeffding's inequality. Thus, for every kth-order sliding window denoiser, difference between empirical loss and expected loss is vanishing with high probability. ■

Lemma 7 *Let $\mathcal{F}_\delta^k$ denote the set of $\mathcal{A}^{2k+1}$-dimensional vecotrs with components in $[0,1]$ that are integer multiples of δ. Note that $\hat{\mathbf{P}}^\delta \in \mathcal{F}_\delta^k$. Also, let $\mathcal{G}_\delta^k = \{\mathcal{B}(\mathbf{P})\}_{\mathbf{P}\in\mathcal{F}_\delta^k}$ be the class of k-th order sliding window denoiser defined by computing the Bayes response with respect to $\mathbf{P} \in \mathcal{F}_\delta^k$. Then, for every $n \ge 1$, $x^n \in \mathcal{A}^n$, $\epsilon > 0$ and $\mathcal{B}(\hat{\mathbf{P}}^\delta) \in \mathcal{G}_\delta^k$,*

$$\Pr\left(\left|L_{\hat{X}_{NN}^\delta}(x^n, Y^n) - \mathbb{E}_{P_{x^n}^k\otimes\mathcal{C}}\left[\mathbf{\Lambda}\left(X_0, \mathcal{B}(\hat{\mathbf{P}}^\delta(X_0|Y_{-k}^k))\right)\right]\right| > \epsilon\right) \le \left[\frac{1}{\delta}+1\right]^M\cdot 2(2k+1)\exp\left(-\frac{2\left(n-2k\right)}{(2k+1)}\epsilon^2\cdot\frac{1}{\Lambda_{\max}^2}\right).$$

Proof: We have

$$\Pr\Big(\Big|L_{\hat{X}^{\delta}_{\text{NN}}}(x^n,Y^n)-\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_0|Y^k_{-k}))\Big)\Big]\Big|>\epsilon\Big)$$

$$=\Pr\Big(\Big|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\mathbf{\Lambda}\Big(x_i,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_i|Y^{i+k}_{i-k}))\Big)-\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_0|Y^k_{-k}))\Big)\Big]\Big|>\epsilon\Big)$$

$$\leq\Pr\Big(\max_{g^*_k\in\mathcal{G}^k_\delta}\Big|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\mathbf{\Lambda}\Big(x_i,g^*_k\left(Y^{i+k}_{i-k}\right)\Big)-\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,g^*_k\left(Y^{i+k}_{i-k}\right)\Big)\Big]\Big|>\epsilon\Big) \tag{22}$$

$$\leq\Big|\mathcal{G}^k_\delta\Big|\cdot 2(2k+1)\exp\Big(-\frac{2\,(n-2k)}{(2k+1)}\epsilon^2\cdot\frac{1}{\Lambda^2_{\max}}\Big) \tag{23}$$

$$\leq\Big[\frac{1}{\delta}+1\Big]^M\cdot 2(2k+1)\exp\Big(-\frac{2\,(n-2k)}{(2k+1)}\epsilon^2\cdot\frac{1}{\Lambda^2_{\max}}\Big), \tag{24}$$

in which (22) follows from considering the uniform convergence, (23) follows from the union bound, and (24) follows from the crude upper bound on the cardinality $|\mathcal{G}^k_\delta|$. Note that the window size k in the superscript of upper bound for the cardinality ($[\frac{1}{\delta}+1]^M$) is removed compared to that of Gen-DUDE ($[\frac{1}{\delta}+1]^{M^{2k+1}}$). The distinction between them follows from difference in modeling where Gen-CUDE tries to directly model the marginal posterior distribution with neural network rather than the joint posterior of $(2k+1)$-tuple. ■

Now, we prove our main theorm.

Theorem 1 *Consider ϵ^* in Lemma 2. Then, for all $k,n\geq 1$, $\delta>0$, and $\epsilon>\Lambda_{\max}\cdot(3\epsilon^*+\frac{M\cdot\delta}{2})$, and for all x^n,*

$$\Pr\Big(|L_{\hat{X}^{n,\delta}_{NN}}(x^n,Y^n)-D^k_{x^n}|>\epsilon\Big)\leq C_1(k,\delta,M)\exp\Big(-\frac{2\,(n-2k)}{(2k+1)}C_2(\epsilon,\epsilon^*,\Lambda_{\max},M,\delta)\Big),$$

in which $C_1(k,\delta,M)\triangleq 2(2k+1)[\frac{1}{\delta}+1]^M$ and $C_2(\epsilon,\epsilon^*,\Lambda_{\max},M,\delta)\triangleq(\epsilon-\Lambda_{\max}\cdot(3\epsilon^*+\frac{M\cdot\delta}{2}))^2\cdot\frac{1}{\Lambda^2_{\max}}$.

Proof of theorem 1: We utilize all the Lemmas given above to prove the theorem. We have

$$\Pr\Big(\Big|L_{\hat{X}^{\delta}_{\text{NN}}}(x^n,Y^n)-D^k_{x^n}\Big|>\epsilon\Big)$$

$$=\Pr\Big(\Big|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\mathbf{\Lambda}\Big(x_i,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_i|Y^{i+k}_{i-k}))\Big)-\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\mathbf{P}(X_0|Y^k_{-k}))\Big)\Big]\Big|>\epsilon\Big) \tag{25}$$

$$\begin{aligned}\leq\Pr\Big(&\Big|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\mathbf{\Lambda}\Big(x_i,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_i|Y^{i+k}_{i-k}))\Big)-\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_0|Y^k_{-k}))\Big)\Big]\Big|\\&+\Big|\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_0|Y^k_{-k}))\Big)\Big]-\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\hat{\mathbf{P}}(X_0|Y^k_{-k}))\Big)\Big]\Big|\\&+\Big|\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\hat{\mathbf{P}}(X_0|Y^k_{-k}))\Big)\Big]-\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\mathbf{P}(X_0|Y^k_{-k}))\Big)\Big]\Big|>\epsilon\Big)\end{aligned} \tag{26}$$

$$\leq\Pr\Big(\Big|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\mathbf{\Lambda}\Big(x_i,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_i|Y^{i+k}_{i-k}))\Big)-\mathbb{E}_{P^k_{x^n}\otimes\mathcal{C}}\Big[\mathbf{\Lambda}\Big(X_0,\mathcal{B}(\hat{\mathbf{P}}^{\delta}(X_0|Y^k_{-k}))\Big)\Big]\Big|>\epsilon-\Lambda_{\max}\cdot(3\epsilon^*+\frac{M\cdot\delta}{2})\Big) \tag{27}$$

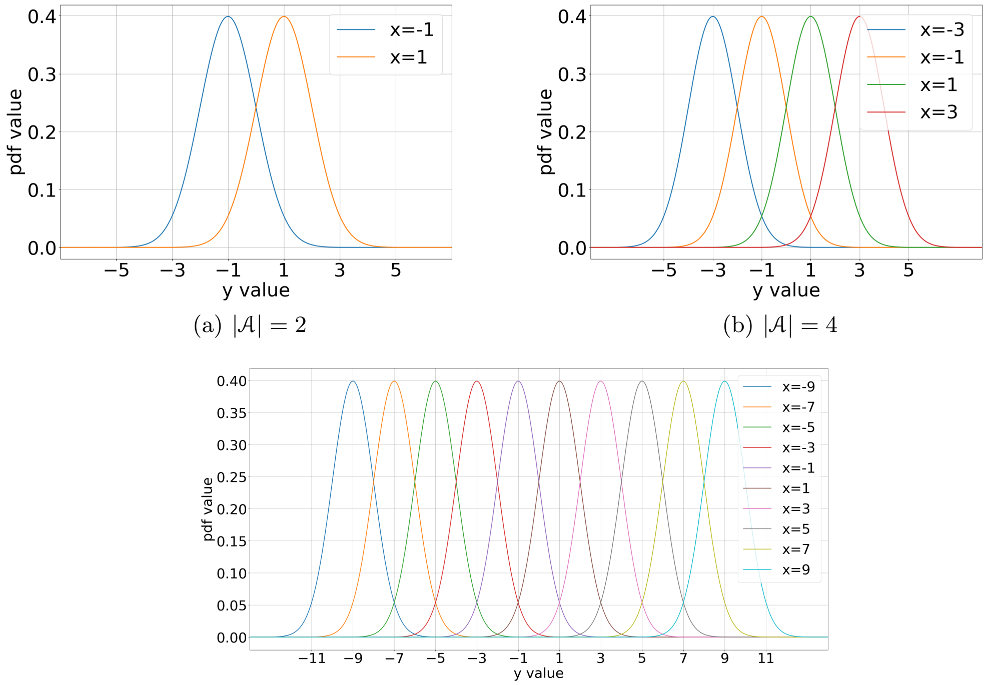
$$\leq\Big[\frac{1}{\delta}+1\Big]^M\cdot 2(2k+1)\exp\Big(-\frac{2\,(n-2k)}{(2k+1)}\cdot\Big(\epsilon-\Lambda_{\max}\cdot(3\epsilon^*+\frac{M\cdot\delta}{2})\Big)^2\cdot\frac{1}{\Lambda^2_{\max}}\Big)) \tag{28}$$

$$=C_1(k,\delta,M)\exp\Big(-\frac{2\,(n-2k)}{(2k+1)}C_2(\epsilon,\epsilon^*,\Lambda_{\max},M,\delta)\Big), \tag{29}$$

where (25) follows from the definition of $L_{\hat{X}^{\delta}_{\text{NN}}}(x^n,Y^n)$ and $D^k_{x^n}$, (26) follows from triangle inequality, (27) follows from applying Lemma 5 and Lemma 4, and (28) follows from Lemma 7. Thus, we proved the theorem. ■

Appendix B Noise Channel Densities

Here, we show the noisy channel density $\{f_x(y)\}_{x\in\mathcal{O}}$ used for the experiments in Section 5.2 and Section 5.3 of the paper. Figure 3 shows the channel densities for the synthetic data experiments in Section 5.2, and Figure 4 shows the channel densities for the 454 and Ion Torrent data experiments in Section 5.3.

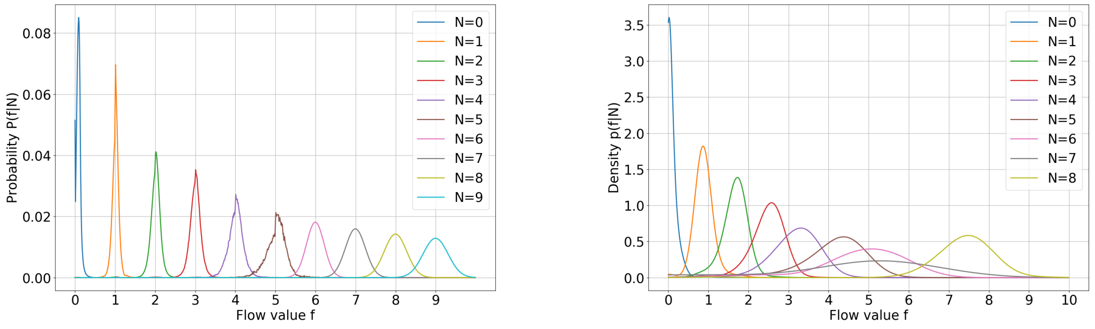


(a) $|\mathcal{A}| = 2$

(b) $|\mathcal{A}| = 4$

(c) $|\mathcal{A}| = 10$

Figure 3: Noisy channel densities used for the synthetic data experiments.



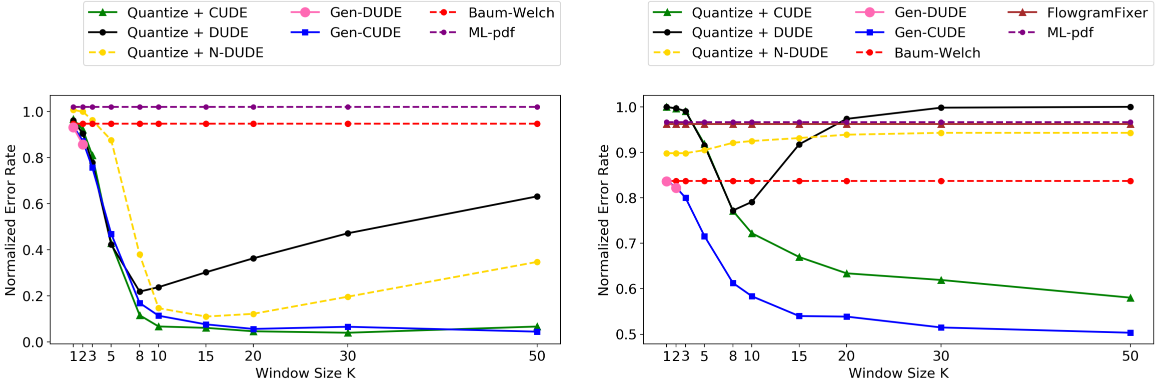
(a) 454 Pyrosequencing

(b) Ion Torrent

Figure 4: Probability ensities of the flowgram-values for the homopolymer lengths in each DNA sequencer. For Ion Torrent, we estimated channel density using Gaussian kernel density estimation with bandwidth=0.6 on the separated holdout dataset.

Appendix C Normalized Error Rate Graph for DNA Experiments

Figure 5 shows the denoising performance measured by the Hamming loss. Note the similarity score in the paper is computed after converting the integer-valued denoised sequence (homopolymer length) back to a DNA sequence. We observe the error patterns are similar to those in Figure 2 of the paper.



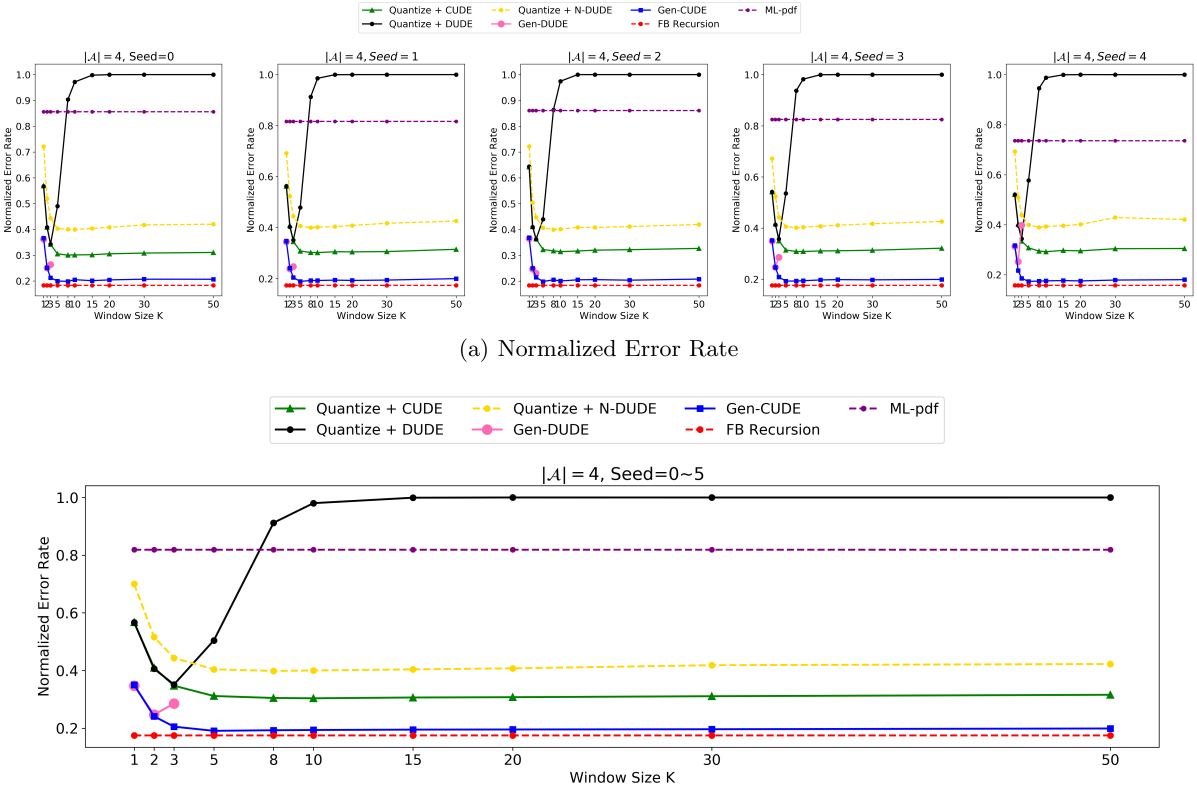
(a) 454 Pyrosequencing

(b) Ion Torrent

Figure 5: Normalized error rate for DNA source data.

Appendix D Error Rate Graph for Randomized Quantizers

We note that the quantizer $Q(\cdot)$ can be freely selected for Gen-CUDE as long as the induced DMC, Π, is invertible. To show the small effect of the quantizer to the final denoising performance, we designed two additional experiments for the $|\mathcal{A}| = 4$ case of Figure 1(a) in the paper. As described in the first paragraph of Section 5.2, the source symbol was encoded as $\{+3, +1, -1, -3\}$ and the decision boundaries of the original $Q(\cdot)$ was $\{-2, 0, +2\}$.



(a) Normalized Error Rate

(b) Average Error Rate

Figure 6: Error Rate for Five Randomized Quantizers

In Figure 6(a), we show the results of using five randomized quantizers, of which decision boundaries were obtained by uniform sampling from the intervals, $[-3, -1], [-1, +1], [+1, +3]$, respectively. The 5 different resulting quantizers' decision boundaries were the following:

- Seed 0 : $[-1.59, 0.73, 2.09]$
- Seed 1 : $[-1.18, 0.27, 2.46]$
- Seed 2 : $[-2.29, -0.59, 2.49]$
- Seed 3 : $[-1.96, -0.41, 1.12]$
- Seed 4 : $[-1.13, 0.81, 1.61]$.

The five figures in Figure 6(a) show the performance for each quantizer, and Figure 6(b) shows the average error rate of them, which looks quite similar the one shown in Figure 1(a). We can clearly observe that the different quantizers have little effect in the final denoising performance for `Gen-CUDE`. In contrast, we observe that `Gen-DUDE` or `Quantize+DUDE` have more sensitivity to the choice of the quantizer.

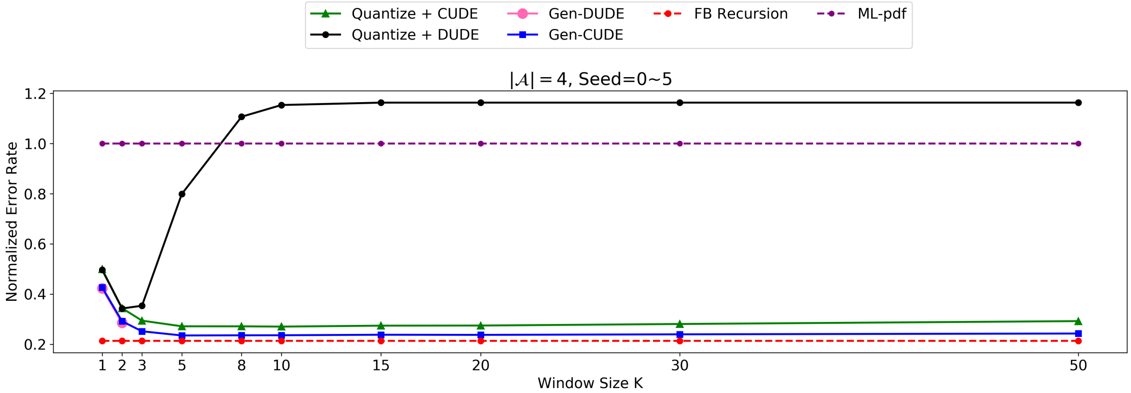


Figure 7: Average Error Rate for Non-square Channel Matrix Case

Furthermore, we note that our Gen-CUDE does not require to have the same number of the quantized symbols as the input symbols, either. In such cases, the $\mathbf{\Pi}^{-1}$ can be simply replaced with a pseudo-inverse as long as $\mathbf{\Pi}$ has full row-rank. Figure 7 is the result of averaging the performances of using five randomized quantizers, of which decision boundaries are randomly selected from the intervals $[-2.7, -2.3]$, $[-1.7, -1.3]$, $[-0.7, -0.3]$, $[0.3, 0.7]$, $[1.3, 1.7]$, $[2.3, 2.7]$, respectively. (Thus, $Q(\cdot)$ has 7 regions.) The used boundaries are as following:

- Seed 0 : $[-2.42, -1.35, -0.48, 0.57, 1.44, 2.36]$
- Seed 1 : $[-2.34, -1.45, -0.41, 0.55, 1.55, 2.32]$
- Seed 2 : $[-2.56, -1.62, -0.4, 0.59, 1.56, 2.55]$
- Seed 3 : $[-2.49, -1.58, -0.68, 0.59, 1.49, 2.51]$
- Seed 4 : $[-2.33, -1.34, -0.58, 0.5, 1.67, 2.64]$.

Again, we see little difference in the performance for `Gen-CUDE` compared to Figure 7 and Figure 1(a) ($|\mathcal{A}| = 4$ case) in the manuscript.