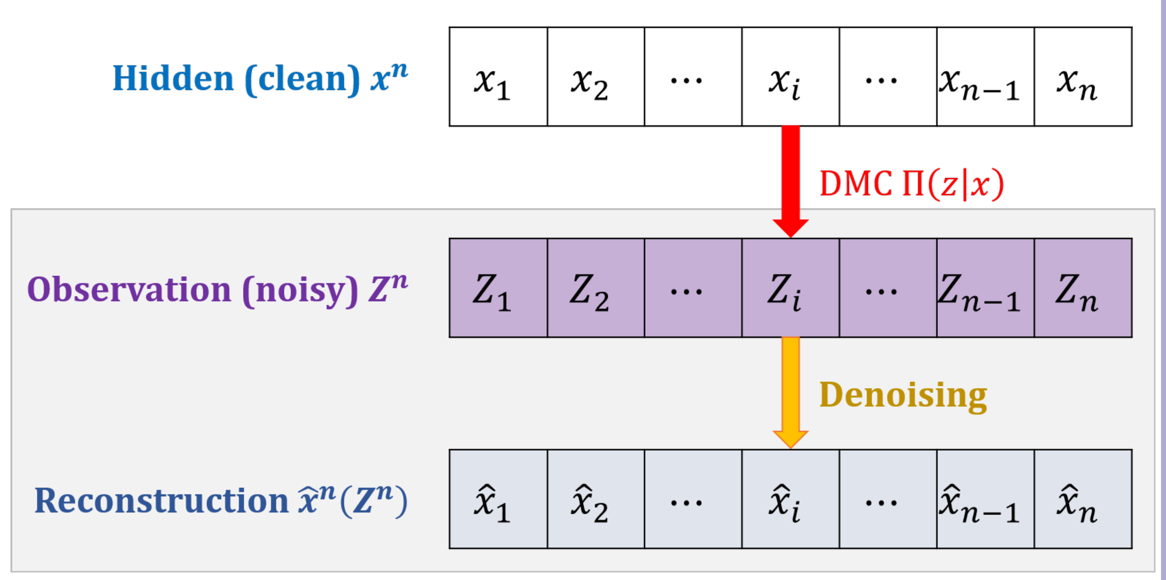
CONDITIONAL DISTRIBUTION LEARNING WITH NEURAL NETWORKS AND ITS APPLICATION TO UNIVERSAL IMAGE DENOISING

Jongha "Jon" Ryu   Young-Han Kim
{jongharyu,yhk}@ucsd.edu
Dept. of Electrical and Computer Engineering, University of California, San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering

Discrete denoising

- Discrete alphabets $\mathcal{X}, \mathcal{Z}, \hat{\mathcal{X}}$
- Known DMC $\Pi(z|x)$ with inverse channel $\Pi^\dagger(x|z)$

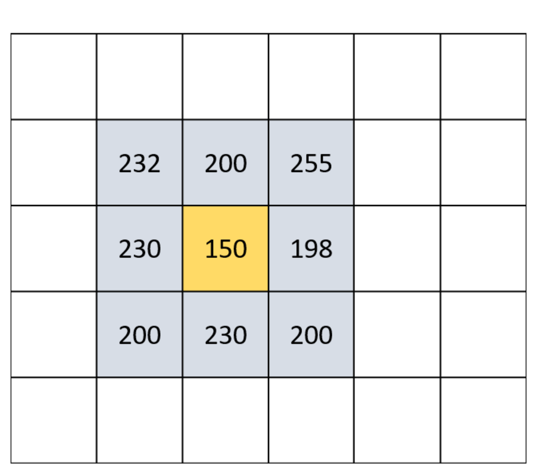


- Loss function $\Lambda\colon \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$
- **Goal**: Based on noisy observation Z^n, reconstruct a clean sequence $\hat{x}^n(Z^n)$ which minimizes the cumulative loss

$$\sum_{i=1}^{n} \Lambda(X_i, \hat{x}_i(Z^n))$$

Context-based methods

- Process each symbol based on its *context*
  - Context = *neighboring* symbols/pixels
  - a.k.a. sliding window algorithms

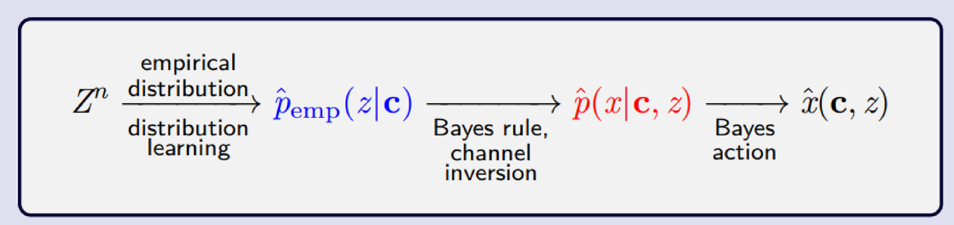


- The underlying source can be parsed into IID sources with respect to the context model
- **Common two-stage algorithm**
  - **(Step 1: Conditional distribution learning)** For each context, learn the conditional distribution of the symbol given the context
  - **(Step 2: Bayes action)** Given a context and a symbol, take the Bayes optimal action with respect to the conditional distribution from **(Step 1)**
- (+) Clean divide-and-conquer
- (−) Sparse context problem: inherent tradeoff in choosing the context order k
  - Small k: data structure may not be captured
  - Large k: too few samples per context

DUDE algorithm

- Discrete Universal DEnoiser (Weissman et al. 2005)
- $\mathcal{C}$: a set of contexts
- Hyperparameter: window size k
- Construct a denoiser $\hat{X} = \hat{x}(Z^n)\colon \mathcal{C} \times \mathcal{Z} \to \hat{\mathcal{X}}$ based on data Z^n

DUDE algorithm runs in *two passes*



1. Find the *empirical distribution*

$$q_{\mathrm{emp}}(z|\mathbf{c}) \triangleq \frac{|\{j\colon \mathbf{c}_j = \mathbf{c}, z_j = z\}|}{|\{j\colon \mathbf{c}_j = \mathbf{c}\}|}$$

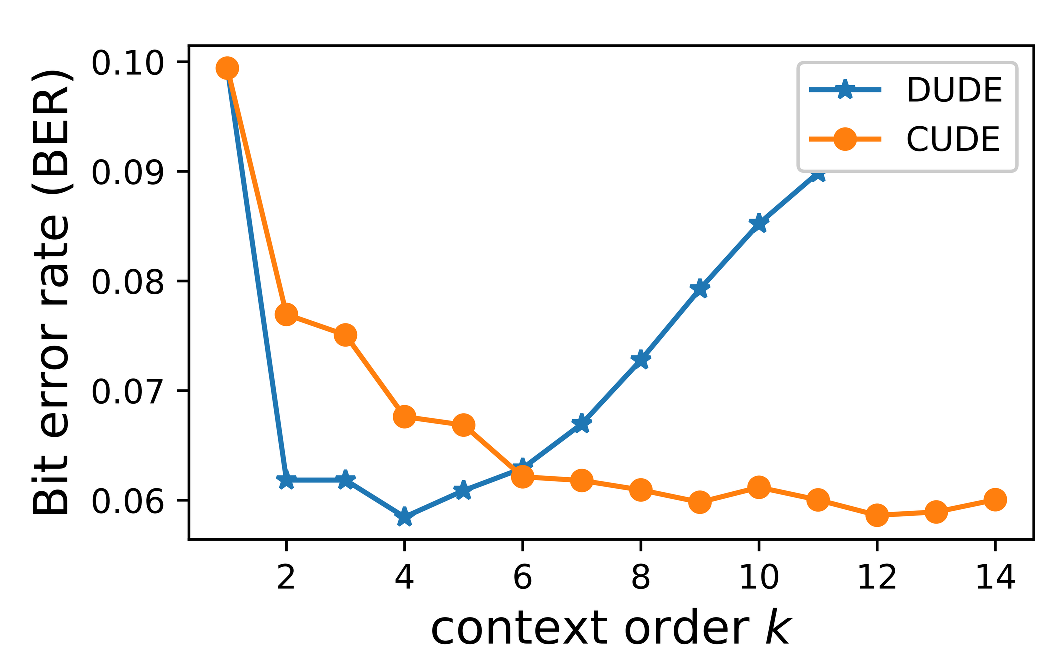
by counting the number of occurrences of noisy symbols for each context $\mathbf{c}$

2. Find the Bayes optimal denoiser under $q(x|\mathbf{c}, z)$, by Bayes rule and channel inversion

- (+) Easy to implement, low complexity
- (+) Theoretical guarantee
  - **Universality**: for any underlying stationary process, DUDE asymptotically attains the Bayes optimal performance, provided that k grows properly with n
- (−) Sparse context problem
  - For grayscale images, $|\mathcal{X}| = |\mathcal{Z}| = 256$
  - Even for $k = 1$, $|\mathcal{C}| = |\mathcal{Z}|^{4k(k+1)} = 256^8 = 2^{64}$

Example

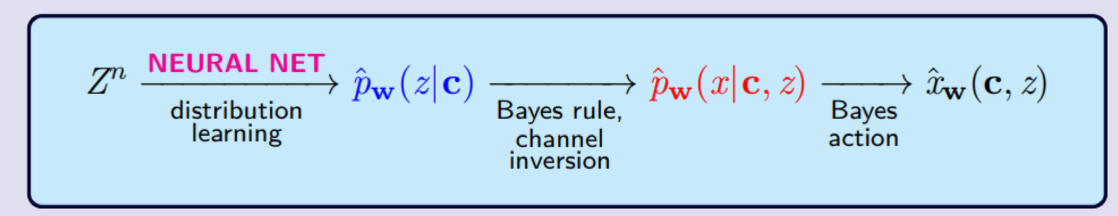
- **Source**: Binary symmetric 1st order Markov sequence with flip probability 0.1
- **Channel**: BSC(p) with $p = 0.1$



Proposed CUDE algorithm

- Context-aggregated Universal DEnoiser
- **Key idea**: Two-stage algorithm with conditional distribution aggregated among similar contexts by a neural network

CUDE algorithm also runs in *two passes*



1. Train a neural network $\mathbf{c} \mapsto q_{\mathbf{w}}(z|\mathbf{c})$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \ln \frac{1}{q_{\mathbf{w}}(z_i|\mathbf{c}_i)}$$
$$= \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{\mathbf{c}} q_{\mathrm{emp}}(\mathbf{c}) D(q_{\mathrm{emp}}(z|\mathbf{c}) \| q_{\mathbf{w}}(z|\mathbf{c}))$$

2. Find the Bayes optimal denoiser under $q_{\mathbf{w}}(x|\mathbf{c}, z)$ by Bayes rule and channel inversion

- Learns $q_{\mathbf{w}}(z|\mathbf{c})$ that minimizes the relative entropy $D(q_{\mathrm{emp}}(z|\mathbf{c}) \| q_{\mathbf{w}}(z|\mathbf{c}))$ for all $\mathbf{c}$ collectively
- **Intuition**: context aggregation via neural networks
  - $\mathbf{c} \mapsto q_{\mathbf{w}}(z|\mathbf{c})$ is continuous
  - A neural network has *finite* capacity

CUDE vs. Neural DUDE

- **Neural DUDE** (Moon et al. 2016) trains a neural network $q_{\mathbf{w}} : \mathcal{C} \to \Delta^{|\mathcal{S}|}$ where $\mathcal{S} = \{s : \mathcal{Z} \to \hat{\mathcal{X}}\}$

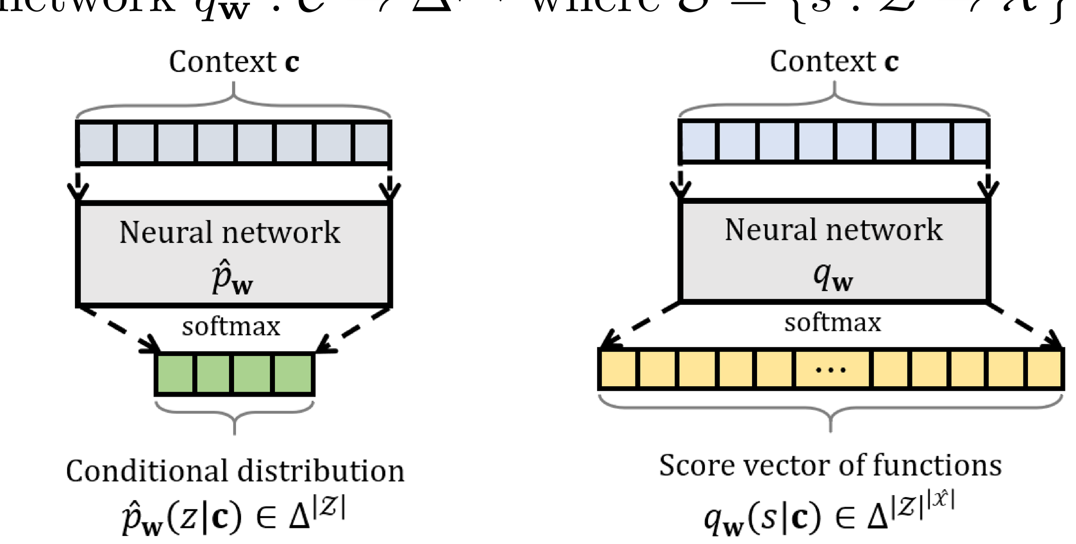


Fig. Comparison of neural networks used in CUDE and Neural DUDE under the context model of order $k = 4$.

- Neural DUDE uses $\arg\max_{s\in\mathcal{S}} q_{\mathbf{w}}(s|\mathbf{c})$ as the best single-symbol denoiser for each $\mathbf{c} \in \mathcal{C}$
- (−) Huge output layer of size $|\mathcal{Z}|^{|\mathcal{X}|}$
- Neural DUDE cannot denoise grayscale images

Experiments

1. Simulation with quaternary images

Noise	Algorithms	Barbara	Boat	Cameraman	Lena
QSC (10%)	DUDE	20.5 (3)	22.0 (2)	24.4 (2)	22.4 (2)
	Neural DUDE	20.7 (26)	21.9 (5)	23.9 (3)	21.9 (27)
	CUDE	**21.5** (36)	**22.6** (11)	**25.2** (10)	**23.1** (6)
QSC (30%)	DUDE	14.7 (3)	16.3 (2)	16.7 (2)	15.7 (3)
	Neural DUDE	16.3 (10)	17.8 (13)	18.7 (16)	17.6 (17)
	CUDE	**16.5** (18)	**18.2** (16)	**19.1** (15)	**17.9** (15)

Table: Comparison of denoising performance in PSNR(dB) for quaternary scaled images corrupted by QSC noise with $\delta = 10\%, 30\%$.

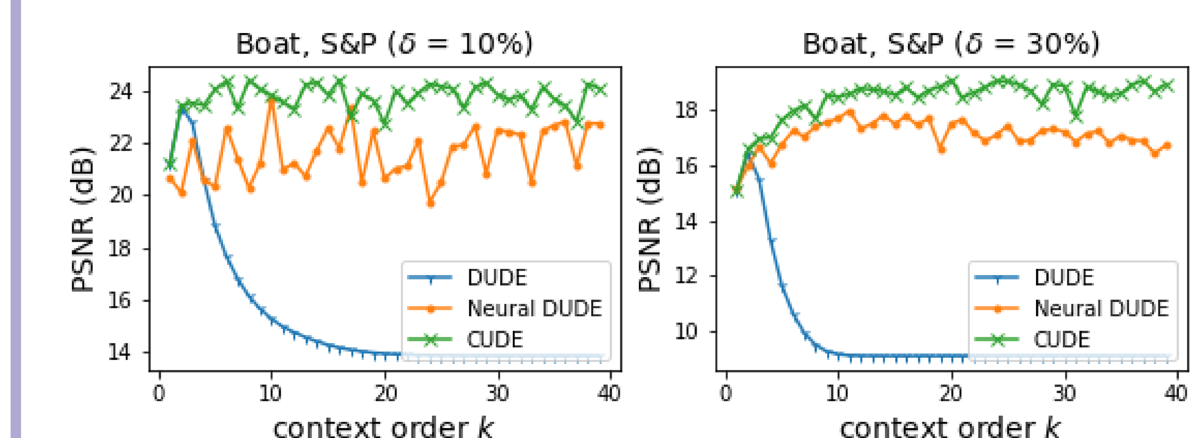
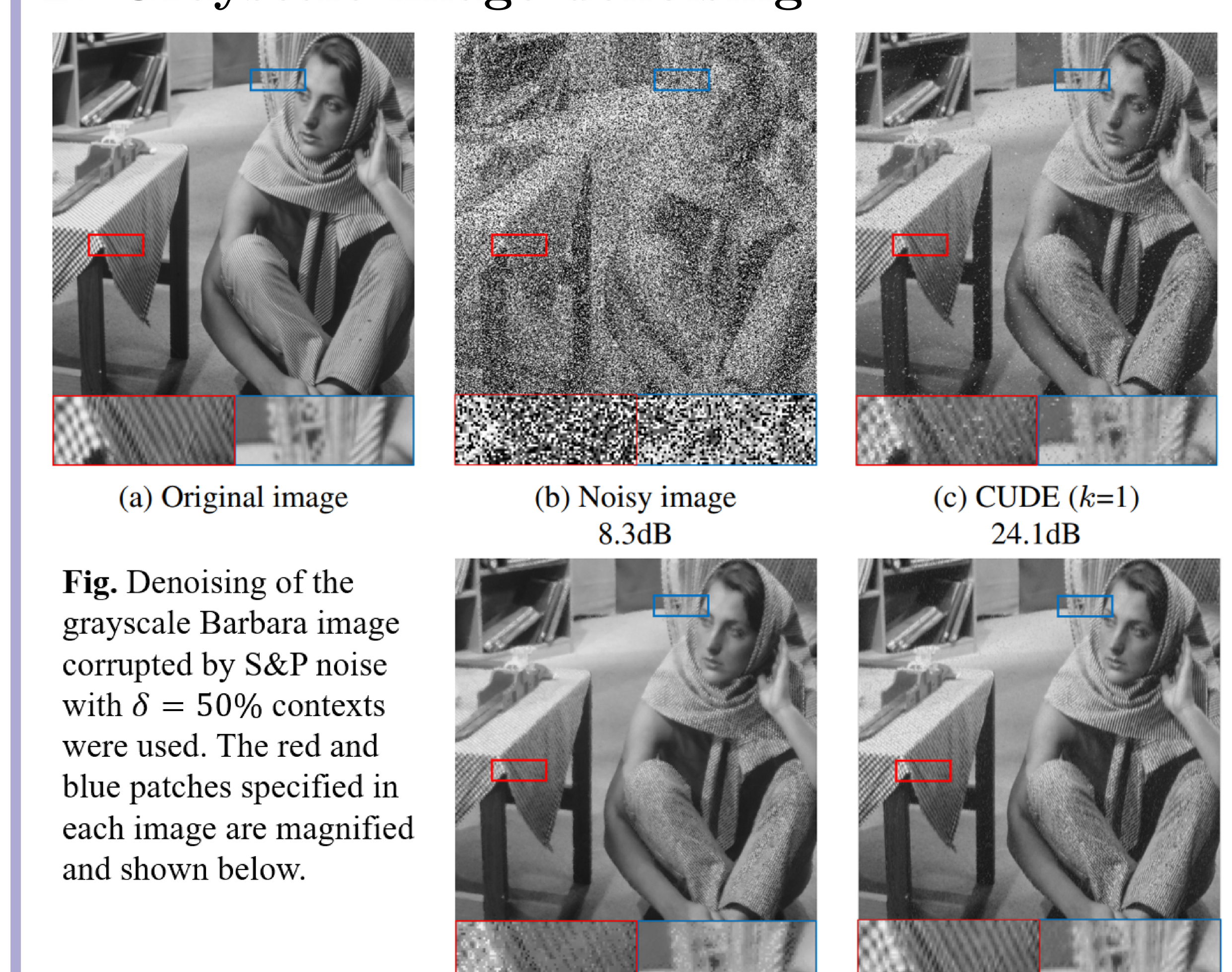


Fig.: PSNR plot for the quaternary boat image corrupted by S&P noise ($\delta = 10\%, 30\%$) with different context orders.

2. Grayscale image denoising



(a) Original image
(b) Noisy image 8.3dB
(c) CUDE (k=1) 24.1dB

Fig. Denoising of the grayscale Barbara image corrupted by S&P noise with $\delta = 50\%$ contexts were used. The red and blue patches specified in each image are magnified and shown below.

(d) IMSM prefiltered image 25.6dB
(e) Iterated CUDE (k=15) **29.9dB**

Future directions

- Continuous alphabets: conditional pdf learning
- Theoretical guarantees
  - When is CUDE better than DUDE?
  - How to quantify the effect of context aggregation?
- Extension to other tasks
  - (Offline) Compression, classification, ...
  - (Online) Prediction, filtering, portfolio selection, ...