Efficient Context-based Algorithms for Sequential Data

Jongha (Jon) Ryu

Department of ECE
University of California San Diego

June 28, 2018

Sequential Data and Context-based Algorithms

- **Sequential data**: text, image, genome sequence, stock prices, ...
- **Tasks**: estimation, denoising, classification, compression, prediction, ...

Sequential Data and Context-based Algorithms

- **Sequential data**: text, image, genome sequence, stock prices, ...
- **Tasks**: estimation, denoising, classification, compression, prediction, ...
- **Context-based algorithms** (a.k.a. sliding window algorithms): make a *decision* based on *context*

Sequential Data and Context-based Algorithms

- **Sequential data**: text, image, genome sequence, stock prices, ...
- **Tasks**: estimation, denoising, classification, compression, prediction, ...
- **Context-based algorithms** (a.k.a. sliding window algorithms): make a *decision* based on *context*

Sequential Data and Context-based Algorithms

- **Sequential data**: text, image, genome sequence, stock prices, ...
- **Tasks**: estimation, denoising, classification, compression, prediction, ...
- **Context-based algorithms** (a.k.a. sliding window algorithms): make a *decision* based on *context*

Sequential Data and Context-based Algorithms

- **Sequential data**: text, image, genome sequence, stock prices, ...
- **Tasks**: estimation, denoising, classification, compression, prediction, ...
- **Context-based algorithms** (a.k.a. sliding window algorithms): make a *decision* based on *context*

Sequential Data and Context-based Algorithms

- **Sequential data**: text, image, genome sequence, stock prices, ...
- **Tasks**: estimation, denoising, classification, compression, prediction, ...
- **Context-based algorithms** (a.k.a. sliding window algorithms): make a *decision* based on *context*

Sequential Data and Context-based Algorithms

- **Sequential data**: text, image, genome sequence, stock prices, ...
- **Tasks**: estimation, denoising, classification, compression, prediction, ...
- **Context-based algorithms** (a.k.a. sliding window algorithms): make a *decision* based on *context*

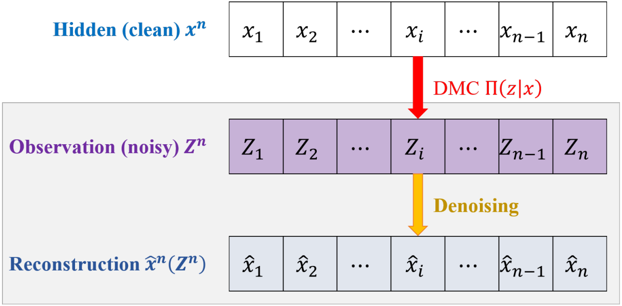
Sequential Data and Context-based Algorithms

- **Sequential data**: text, image, genome sequence, stock prices, ...
- **Tasks**: estimation, denoising, classification, compression, prediction, ...
- **Context-based algorithms** (a.k.a. sliding window algorithms): make a *decision* based on *context*

- Common two-stage approach
  1. Learn conditional distributions from data
  2. Take Bayes optimal actions
- Quick-and-clean divide-and-conquer approach
  1. Decompose a complex problem into disjoint memoryless problems
  2. Plug-in optimal strategy for each subproblem
- Sparse context problem
  - To capture a higher order dependence, need large contexts
  - But only limited # of samples $\longrightarrow$ poor performance!
- **Q: How can we resolve the sparse context problem?**

Problem Setting: 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 Z^n, reconstruct a clean $\hat{x}^n(Z^n)$ which minimizes

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

Problem Setting: Context Model

- Concretely, let's use balanced contexts of size k (hyperparameter)
  - For 1D data, $\mathbf{C}_i \triangleq (Z_{i-k}^{i-1}, Z_{i+1}^{i+k}) \in \mathcal{C} \triangleq \mathcal{C}^{(k)} \cong \mathcal{Z}^{2k}$

	A	G	A	C	T	C	G		

  - For 2D data

	232	200	255		
	230	150	198		
	200	230	200		

- In general, any valid context model (hyperparameter) would work

Traditional Approach (1): DUDE

- **D**iscrete **U**niversal **DE**noiser [Weissman et al., 2005]
- **DUDE** runs in *two passes*

$$Z^n \xrightarrow[\text{distribution learning}]{\text{empirical distribution}} \hat{p}_{\mathrm{emp}}(z|\mathbf{c}) \xrightarrow[\text{Bayes rule, channel inversion}]{} \hat{p}(x|\mathbf{c}, z) \xrightarrow[\text{Bayes action}]{} \hat{x}(\mathbf{c}, z)$$

1. Find the conditional empirical distribution

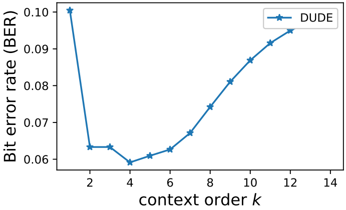
$$\hat{p}_{\mathrm{emp}}(z|\mathbf{c}) \triangleq \frac{|\{j\colon \mathbf{c}_j = \mathbf{c}, z_j = z\}|}{|\{j\colon \mathbf{c}_j = \mathbf{c}\}|}$$

2. Find the Bayes optimal denoiser under $\hat{p}(x|\mathbf{c}, z)$

$$\hat{p}(x|\mathbf{c}, z) = \frac{\Pi(z|x)\hat{p}(x|\mathbf{c})}{\hat{p}_{\mathrm{emp}}(z|\mathbf{c})} = \frac{\Pi(z|x)}{\hat{p}_{\mathrm{emp}}(z|\mathbf{c})} \sum_{z'} \Pi^{\dagger}(x|z')\hat{p}_{\mathrm{emp}}(z'|\mathbf{c})$$

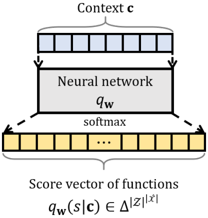
Traditional Approach (1): DUDE

- **D**iscrete **U**niversal **DE**noiser [Weissman et al., 2005]
- Low complexity, universality
  - **Universality**: for any x^n, DUDE asymptotically attains the performance of the best sliding window denoiser of the *same order*
- Sparse context problem
  - For grayscale images, $|\mathcal{X}| = |\mathcal{Z}| = 256$
  - Even for $k = 3$, $|\mathcal{C}^{(k)}| = |\mathcal{Z}|^{2k} = 256^6 = 2^{48}$
  - **Example**
    - Source: binary symmetric 1st order Markov sequence
    - Channel: BSC(p) with $p = 0.1$



Traditional Approach (2): Neural DUDE

- **Neural DUDE** [Moon et al., 2016]
- Introduces a *neural network*
  - Train a *neural network* $q_{\mathbf{w}} : \mathcal{C} \to \Delta^{|\mathcal{S}|}$
  - where $\mathcal{S} := \{s : \mathcal{Z} \to \hat{\mathcal{X}}\}$, a set of all single symbol denoisers



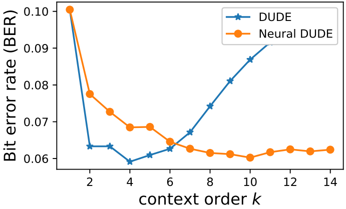
  - After training, for each context $\mathbf{c}$, use

$$s^*(\mathbf{c}) = \arg\max_{s \in \mathcal{S}} q_{\mathbf{w}}(s|\mathbf{c})$$

  - Training data generated based on an unbiased estimator of loss function

Traditional Approach (2): Neural DUDE

- **Neural DUDE** [Moon et al., 2016]
- Outperforms DUDE in practice!



- Huge output layer of size $|\mathcal{Z}|^{|\hat{\mathcal{X}}|}$
  - For $|\mathcal{X}| = |\mathcal{Z}| = |\hat{\mathcal{X}}| = 10$, we have $|\mathcal{Z}|^{|\hat{\mathcal{X}}|} = 10^{10}$

Proposed Method: CUDE

- Context aggregated Universal DEnoiser [Ryu and Kim, 2018]
- CUDE also runs in *two passes*

$$Z^n \xrightarrow[\text{distribution learning}]{\text{NEURAL NET}} \hat{p}_{\mathbf{w}}(z|\mathbf{c}) \xrightarrow[\text{Bayes rule, channel inversion}]{} \hat{p}_{\mathbf{w}}(x|\mathbf{c}, z) \xrightarrow[\text{Bayes action}]{} \hat{x}_{\mathbf{w}}(\mathbf{c}, z)$$

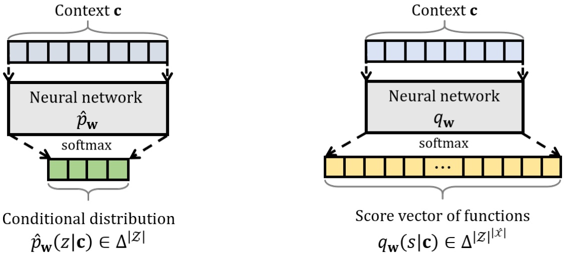
1. Train a neural network $\mathbf{c} \mapsto \hat{p}_{\mathbf{w}}(z|\mathbf{c})$

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

2. Find the Bayes optimal denoiser under $\hat{p}_{\mathbf{w}}(x|\mathbf{c}, z)$

Proposed Method: CUDE

- **C**ontext aggregated **U**niversal **DE**noiser [Ryu and Kim, 2018]
- Outperforms DUDE and Neural DUDE in practice!



(a) CUDE **(b) Neural DUDE**

  - Simple output layer can manage large alphabets
- **Intuition**: context aggregation via neural nets
  1. $\mathbf{c} \mapsto \hat{p}_{\mathrm{w}}(z|\mathbf{c})$ is a continuous mapping
  2. the neural net has *finite* capacity

CUDE: Experiment 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 $\delta = 10\%, 30\%$.

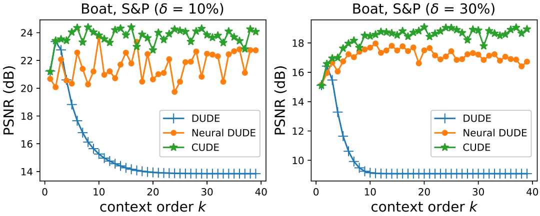
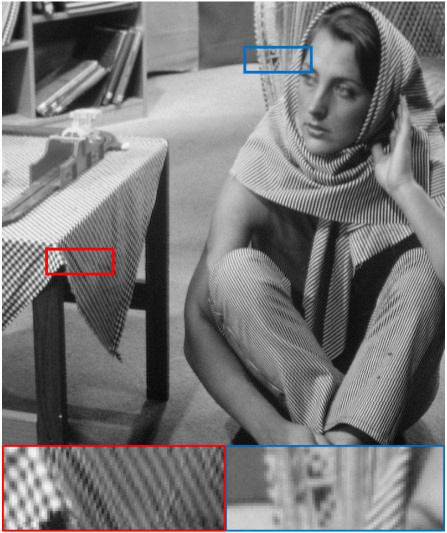


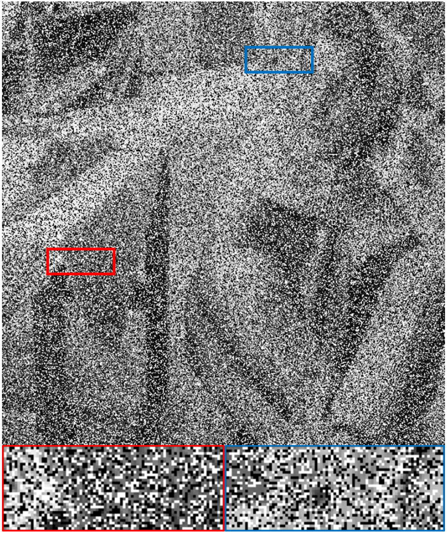
Fig.: PSNR plot for the quaternary boat image corrupted by S&P noise $(\delta = 10\%, 30\%)$ with different context orders.

CUDE: Experiment with Grayscale Images



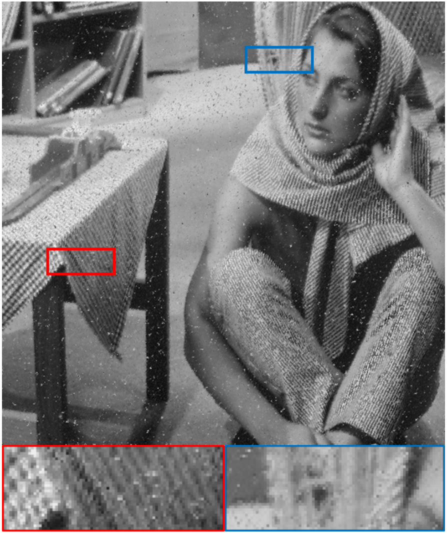
(a) Original

CUDE: Experiment with Grayscale Images



(b) Noisy (S&P(δ) noise, $\delta = 50\%$) 8.3dB

CUDE: Experiment with Grayscale Images

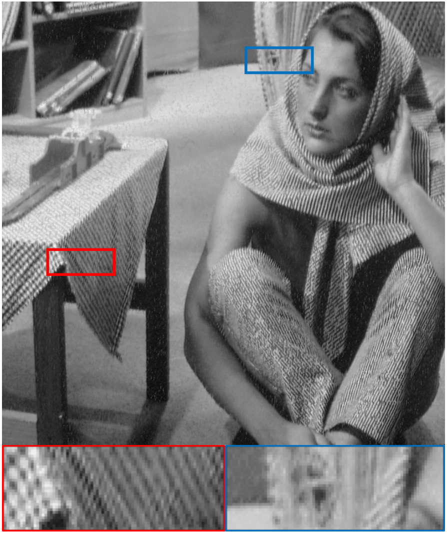


(c) Vanilla CUDE $(k = 1)$ **24.1dB**

CUDE: Experiment with Grayscale Images

(d) IMSM prefiltered image 25.6dB

CUDE: Experiment with Grayscale Images



(e) Iterated CUDE with prefiltering $(k = 15)$ **29.9dB**

Future Directions

- Unstructured noise such as Gaussian
- Performance analysis
  - When is CUDE better than DUDE?
  - Can we quantify the context aggregation effect by neural net?
- Extension to continuous alphabets
  - Conditional *density* estimation via neural network
- Extension to other tasks
  - (Offline) Compression, classification, ...
  - (Online) Prediction, filtering, portfolio selection, ...

[1] Jongha Ryu, Young-Han Kim, **"Conditional Distribution Learning Using Neural Networks and Its Application to Universal Image Denoising"**, accepted to *International Conference on Image Processing (ICIP)*, 2018.

References

Moon, T., Min, S., Lee, B., and Yoon, S. (2016).
Neural universal discrete denoiser.
In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4772–4780. Curran Associates, Inc.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2005).
Universal discrete denoising: known channel.
IEEE Transactions on Information Theory, 51(1):5–28.

Any Questions?