Bangladesh University of Engineering and Technology
Department of Electrical and Electronic Engineering
EEE 400 (January 2024) - Final Examination
Presentation Date: 16th March 2024

Implementation and Performance Analysis of DPMix Algorithm
A Random Mixing Technique for Synthesizing Differentially Private Datasets

Presented by:

Sabir Mahmud

Student ID: 1906032

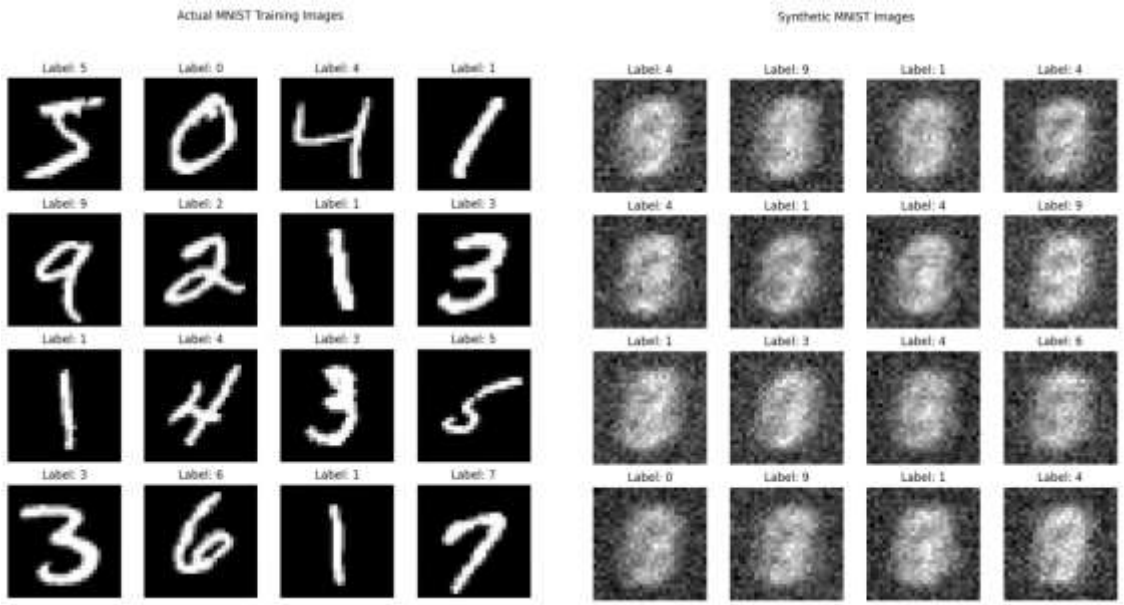
Examination Committee:

1. Dr. Hafiz Imtiaz (Supervisor)
2. Dr. Mohammed Imamul Hassan (Member)
3. Dr. Mohammad Ariful Haque (Member)

Outline

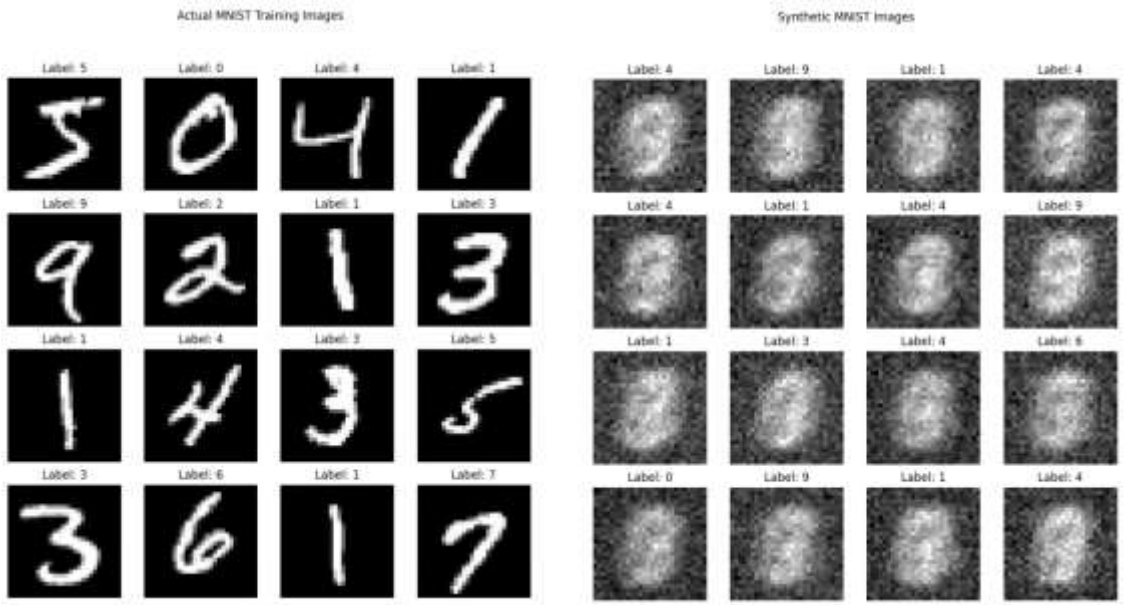
- Abstract
- Literature Review
- Implementation
- Analysis
- Results
- Conclusion
- Acknowledgement and References

EEE 400: Implementation and Performance Analysis of DPMix Algorithm



Outline

EEE 400: Implementation and Performance Analysis of DPMix Algorithm



Why We Need Privacy?

Privacy protects human dignity, ensures safety and supports self-determination. Therefore, protecting privacy is essential for us as humans.

We need privacy to:

1. **Protect personal Information**
   Our financial, medical, and personal data should not be exposed to unauthorized parties.
2. **Prevent Identity theft**
   Without privacy, criminals can misuse our personal details for fraud and scams.
3. **Ensure Freedom & Security**
   Privacy allows us to express our opinions without fear of surveillance or manipulation.
4. **Maintain Trust**
   When our privacy is protected, we feel safe and trust the people and organizations around us.

Privacy in a Data-Driven World

- In modern world, many government and non-government organizations need our private and personal data to carry out their activities
- For example, academic, medical and military sectors might need some data from the population that we might consider personal
- We may feel insecure about giving up such private data, even though it is essential for their research works
- To solve such problems, a special security system was developed called ‘Differential Privacy’

Brief Overview of Differential Privacy

- A privacy preserving technique that conceals individual data points in a dataset by adding controlled random noise.
- Differential privacy promises that the outcome of a survey will stay the same whether an individual participates or not in it
- ε,δ-differential privacy ensures that a change to one entry in a database only creates a small change in the probability distribution of the output
- In 2019 a new privacy preserving algorithm named 'DPMix' was proposed by Lee et al. in their research paper titled with 'Synthesizing Differentially Private Datasets using Random Mixing'
- DPMix is a random mixing technique where a new synthetic dataset is formed by mixing ℓ randomly chosen data points from the original dataset and then perturbing them with an additive noise.

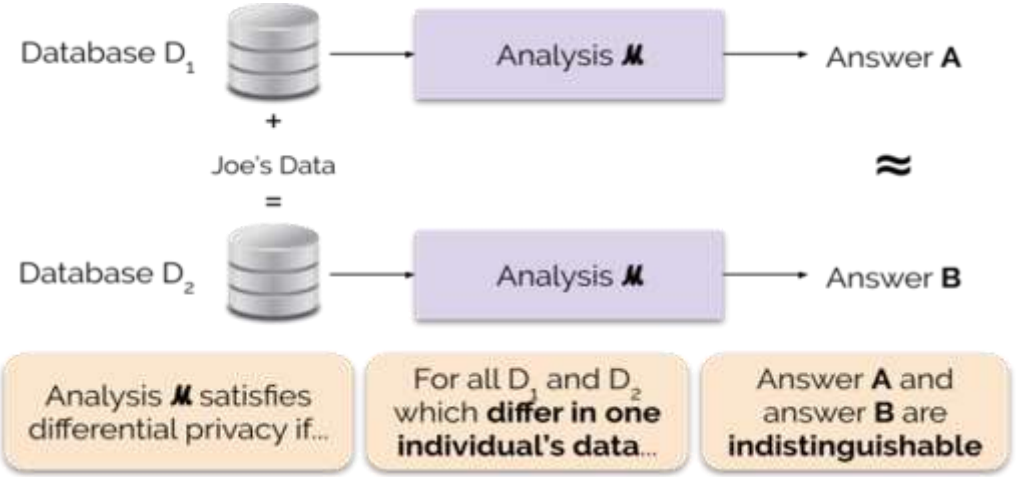


Fig: An informal definition of differential privacy

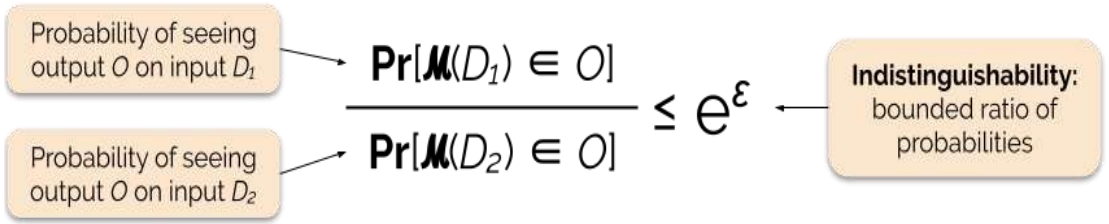
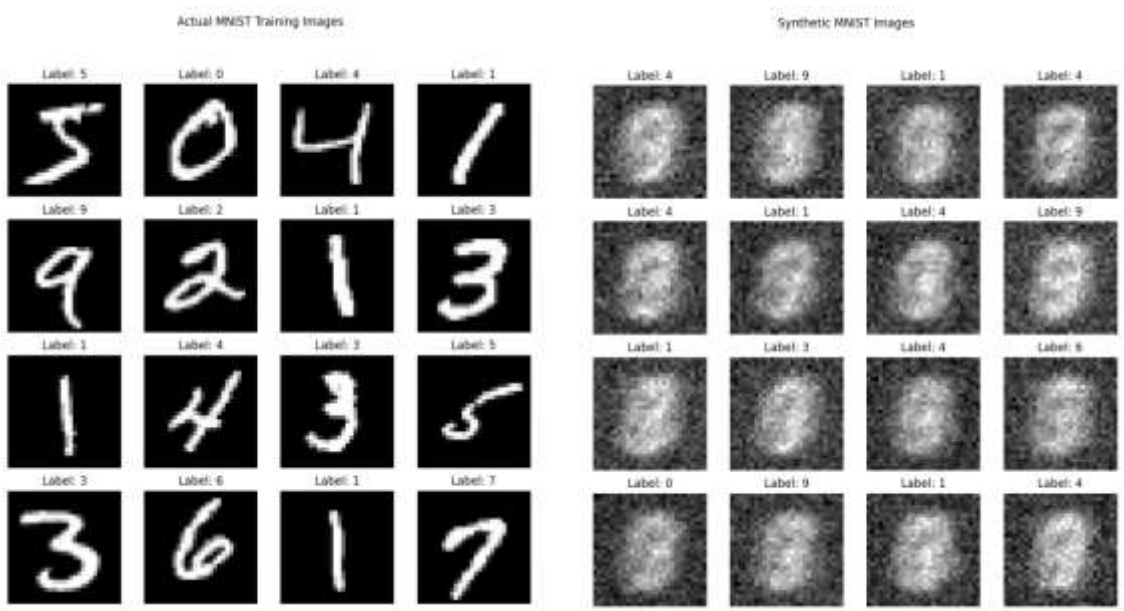


Fig: A formal definition of ε-differential privacy

Outline

- Abstract
- Literature Review
- Implementation
- Analysis
- Results
- Conclusion
- Acknowledgement and References

EEE 400: Implementation and Performance Analysis of DPMix Algorithm



Literature Review

1) Differentially Private Data Publishing Algorithms

- **Local Perturbation**
  - A simple data publishing algorithm that corrupts every data point with additive noise
  - Low computational complexity
  - Retains the data domain (e.g., images remain images).
  - With such a synthetic dataset, one may train an efficient deep learning model such as CNNs
  - However, attaining high privacy requires mixing of excessive noise which might reduce utility for deep learning models

- **Random Projection**
  - Reduces dimensionality before adding noise
  - Generates synthetic data in a different domain, limiting domain-specific learning

Unfortunately, the computational complexities of most of the existing algorithms are exponential in the dimensionality of the dataset, so they are not applicable to most deep learning applications that deal with high dimensional data.

Literature Review (Continued)

2) Differentially Private Machine Learning Algorithms

- Rather than privatizing datasets, some works focus on ensuring privacy in learning algorithms
- It's a relaxed version of of the differential private data publishing problem
- However, if an adversary has an access to the input end of the systems, e.g., data storages, privacy is compromised

3) Learning with mixtures

- **Linear Models**
  - Karakus et al. showed that a simple linear model can be trained with noiseless mixtures

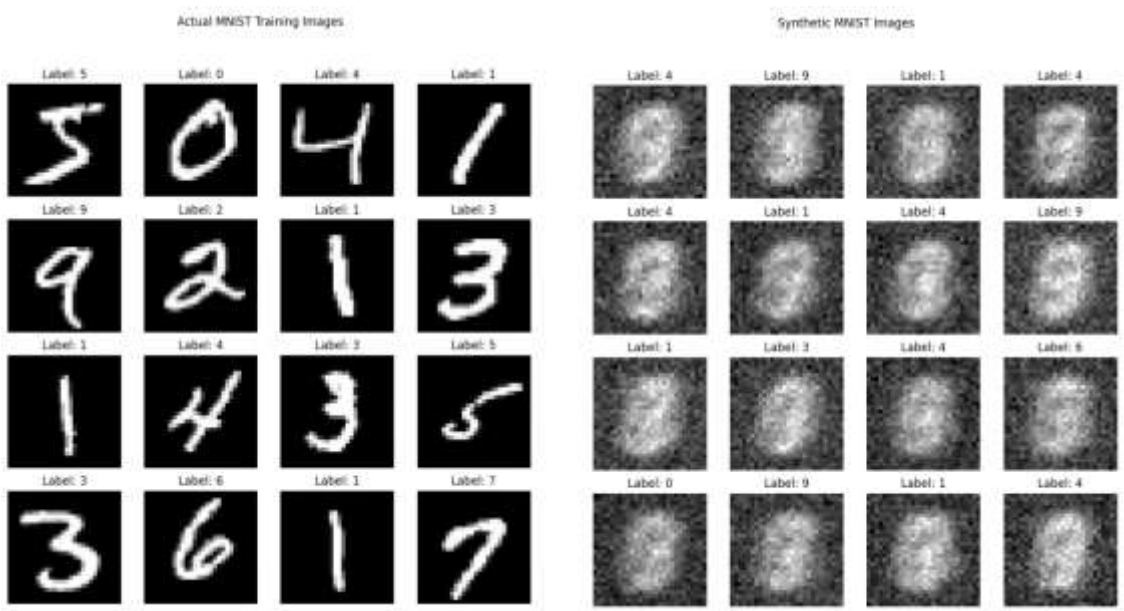
- **Non-linear Models**
  - Tokozume et al. showed that one can train a sound recognition model with mixtures of audio signals
  - A few studies show that one can train image classification models with mixtures of images

While noiseless mixtures of only two or three data points are considered in these studies, we will empirically show that one can train deep neural networks with noisy mixtures of much larger degrees

Outline

- Abstract
- Literature Review
- Implementation
- Analysis
- Results
- Conclusion
- Acknowledgement and References

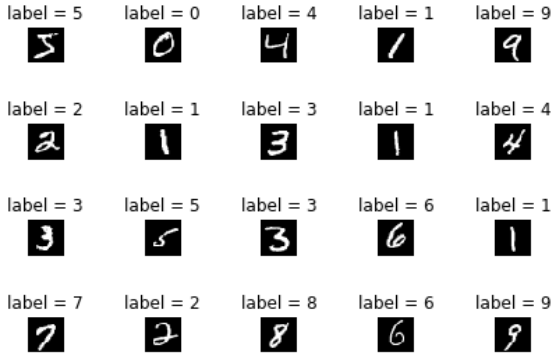
EEE 400: Implementation and Performance Analysis of DPMix Algorithm



Implementation

Dataset:

- MNIST
- A dataset of handwritten digits (0-9) with labels
- 60,000 training samples and 10,000 testing samples
- gray scale images of 28×28 (784 pixels)



DPMix Algorithm:

- A certain number of samples are chosen randomly from the dataset and then mixed together
- We denote the mixture degree by ℓ
- The mixture is then perturbed with additive Gaussian noise

$$X_t^{'} = X^T \times C_t + Q_t$$

$$Y_t^{'} = Y^T \times C_t + R_t$$

- We get a synthetic dataset that is differentially private
- This dataset guarantees strong privacy while ensuring reasonable prediction accuracy for deep learning models

Implementation (Continued)

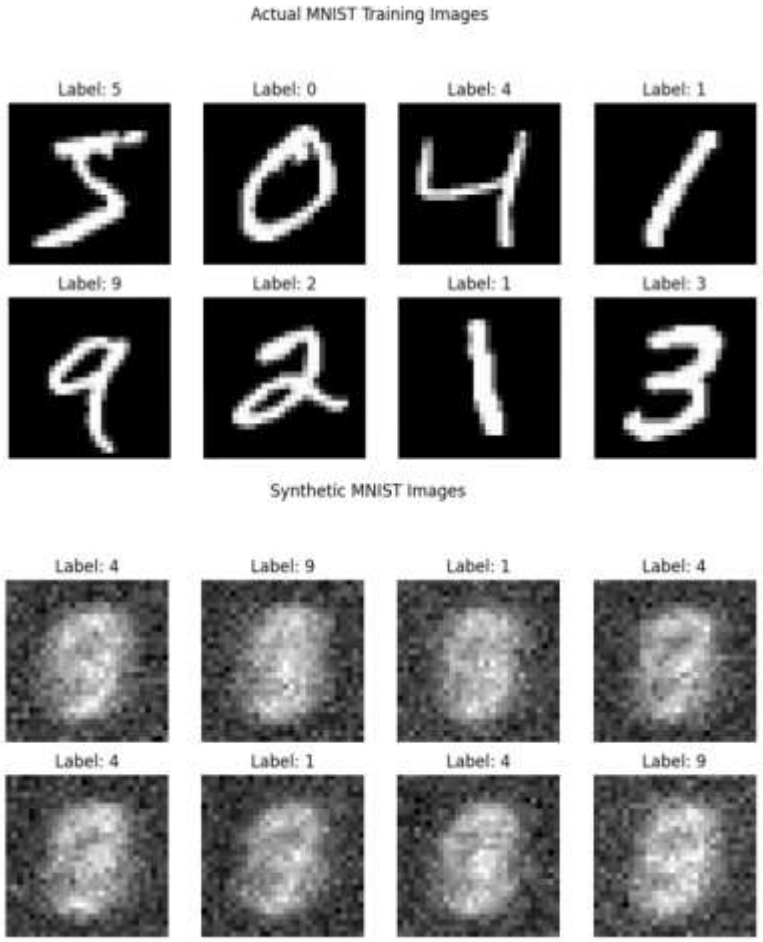


Fig: Samples from the Original & Synthetic Datasets

CNN model trained with the Synthetic Dataset:

Layer Type	Configuration
Conv2d	32 filters of 5x5, activation ReLU
Max-Pooling	2x2
Conv2d	48 filters of 5x5, activation ReLU
Max-Pooling	2x2
Fully connected	100 units, activation ReLU
Fully connected	100 units, activation ReLU
Fully connected	10 units, activation Softmax

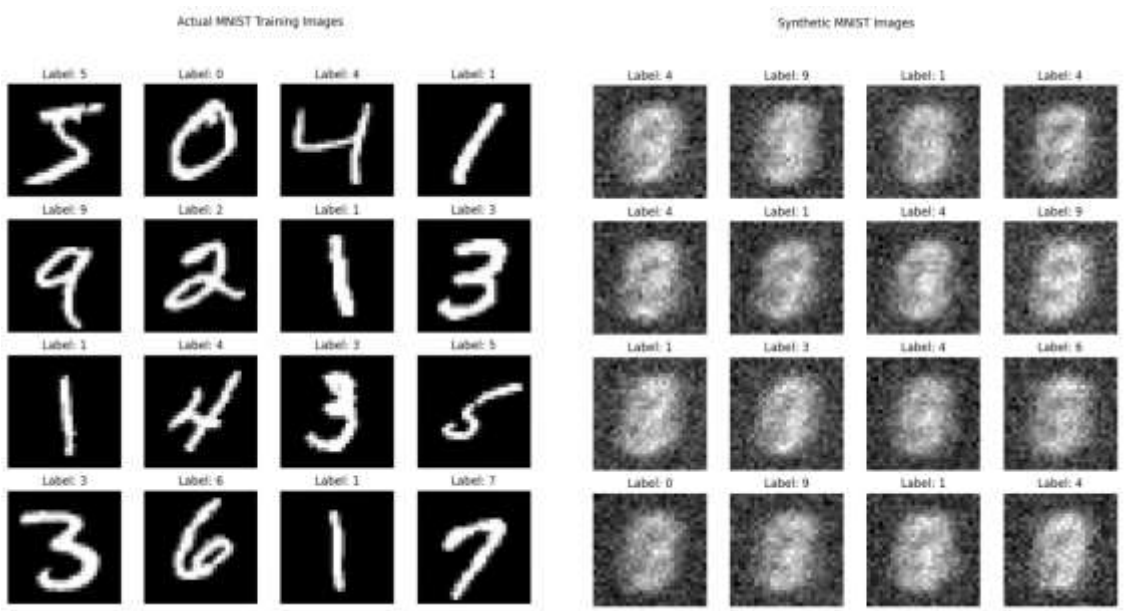
The synthetic dataset is (ε, δ) – differentially private

For, $\varepsilon = 15$ and $\delta = 1/60000$
I got a predictive accuracy of 68.35%

Outline

- Abstract
- Literature Review
- Implementation
- Analysis
- Results
- Conclusion
- Acknowledgement and References

EEE 400: Implementation and Performance Analysis of DPMix Algorithm



Analysis

- After implementing the DPMix algorithm, our target is to observe how the variation in different parameters like:
  - $\boldsymbol{\sigma_x}$ (Additive Noise)
  - $\boldsymbol{\ell}$ (Mixture Degree) and
  - **T** (Total Number of Samples in the Synthetic Dataset)

  affects the accuracy of the predictive model.
- This time, instead of randomly mixing samples from the entire, we mixed only the samples that have same label value.
- i.e. zeros were mixed with zeros, ones were mixed with ones and so on.
- Number of samples for each digit in MNIST training dataset is close to 6000.
- The mixture were perturbed with additive Gaussian Noise having parameters (σ_x , σ_y)
- For convenience we took $\sigma_x = \sigma_y$

Analysis (Continued)

- The synthetic datasets are (ε, δ) – differentially private,

  Where, δ = 1/6000

  And, ε is given by the formula:

$$\varepsilon = \min_{\alpha \in \{2,3,\dots\}} \left(T\varepsilon'_\alpha + \frac{\log(1/\delta)}{\alpha - 1} \right)$$

1. $\varepsilon\alpha'$ Calculation:

$$\varepsilon'_\alpha = \frac{1}{\alpha - 1} \log \left(1 + \left(\frac{l}{n}\right)^2 \binom{\alpha}{2} \min \left(4 \left(\exp\left(\frac{\Delta^2}{l^2}\right) - 1 \right), 2\exp\left(\frac{\Delta^2}{l^2}\right) \right) + 4G(\alpha) \right)$$

2. $G(\alpha)$ Calculation:

$$G(\alpha) = \sum_{j=3}^{\alpha} \binom{l}{n}^j \binom{\alpha}{j} \sqrt{B(2\lfloor j/2 \rfloor) \cdot B(2\lfloor j/2 \rfloor)}$$

3. Bernoulli Polynomial Calculation:

$$B(l) = \sum_{i=0}^{l} (-1)^i \binom{l}{i} \exp\left(\frac{i(i-1)\Delta^2}{2l^2}\right)$$

4. Sensitivity Calculation:

$$\Delta^2 = \frac{d_x}{\sigma_x^2} + \frac{d_y}{\sigma_y^2}$$

CNN model:

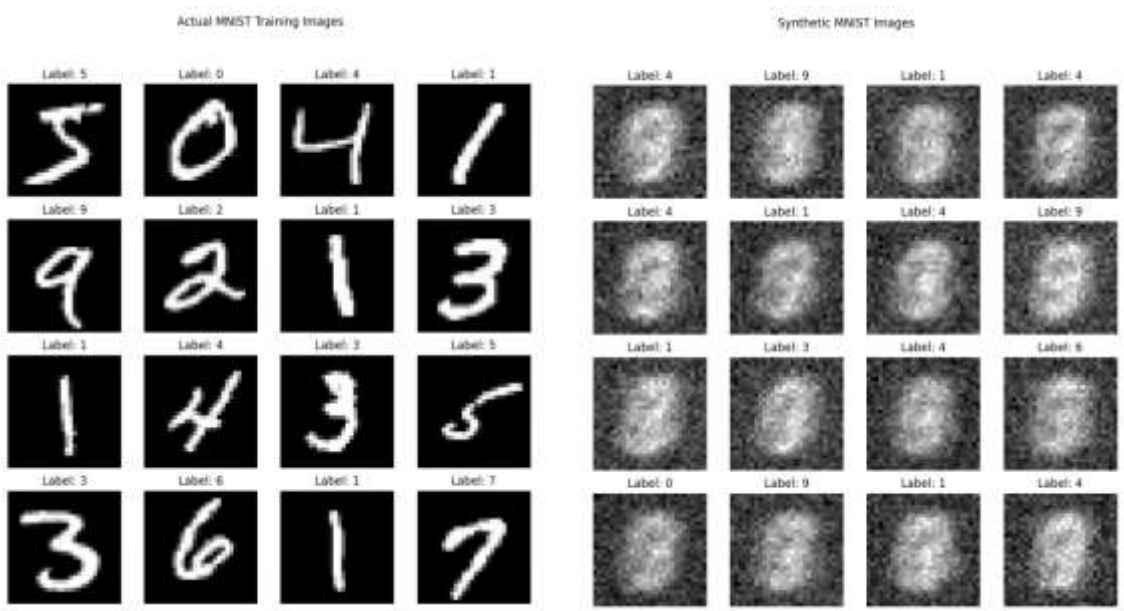
Layer Type	Configuration
Conv2d	32 filters of 5x5, activation ReLU
Max-Pooling	2x2
Conv2d	48 filters of 5x5, activation ReLU
Max-Pooling	2x2
Fully connected	100 units, activation ReLU
Fully connected	100 units, activation ReLU
Fully connected	10 units, activation Softmax

The same CNN model was trained using the Synthetic Datasets with variation in parameters σ_x, ℓ and T

Outline

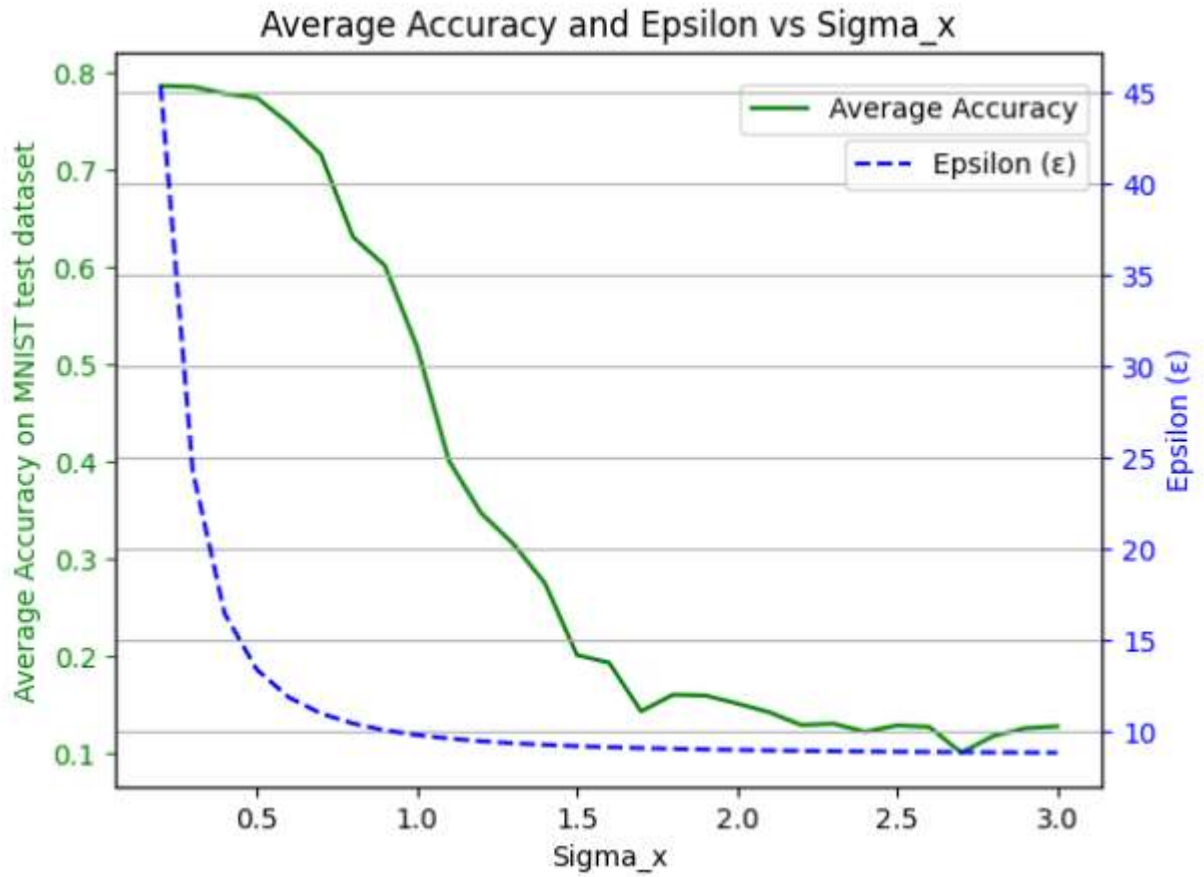
- Abstract
- Literature Review
- Implementation
- Analysis
- **Results**
- Conclusion
- Acknowledgement and References

EEE 400: Implementation and Performance Analysis of DPMix Algorithm



Results

1. Accuracy Vs Noise (σ_x)

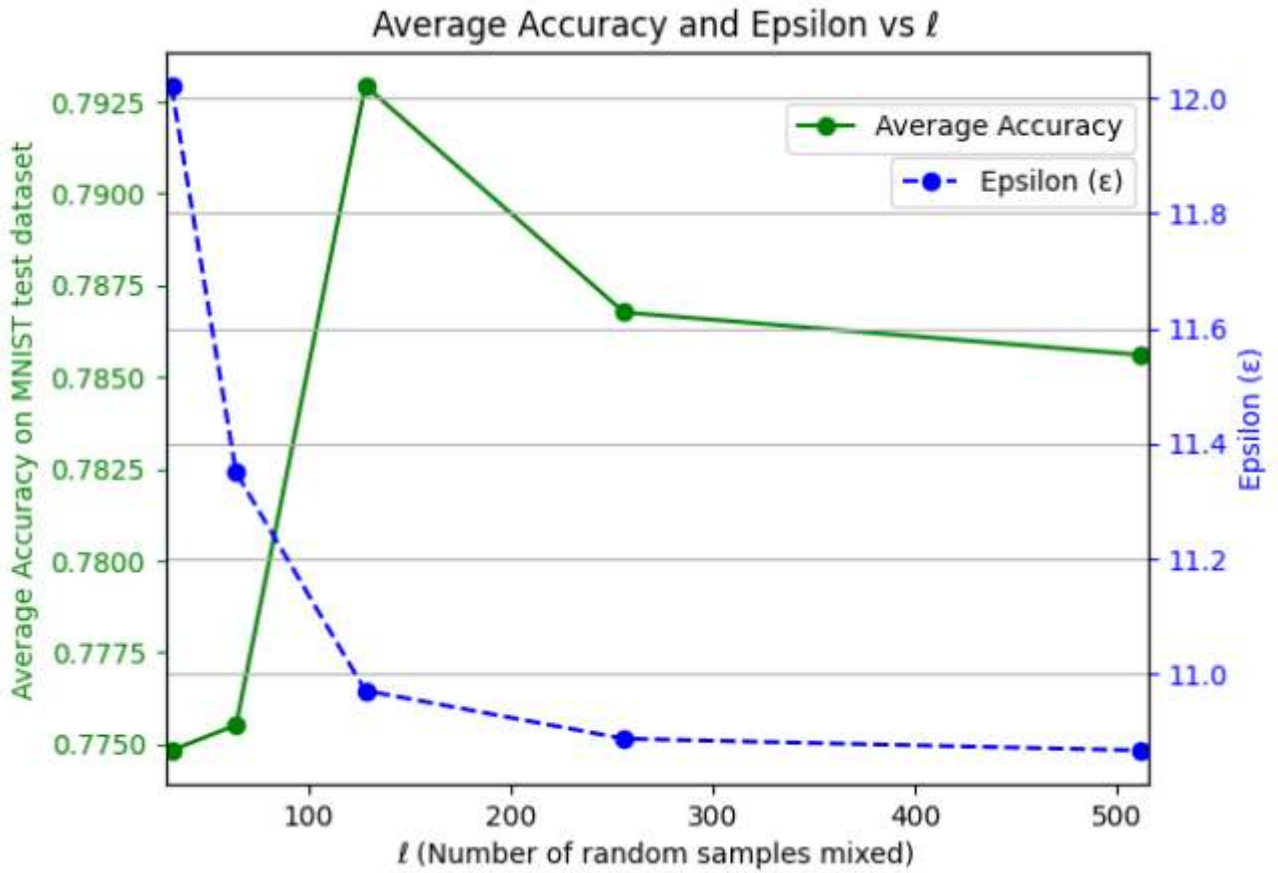


- ✓ From the plot it is evident that, Accuracy decreases as the variance of Noise increases
- ✓ Higher additive noise decreases the privacy parameter ε
- ✓ Lower ε = Strong privacy
- ✓ Higher ε = Weak privacy

Fixed Values:
$\ell = 128$, T = 12000

Results (Continued)

2. Accuracy Vs Mixture Degree (ℓ)



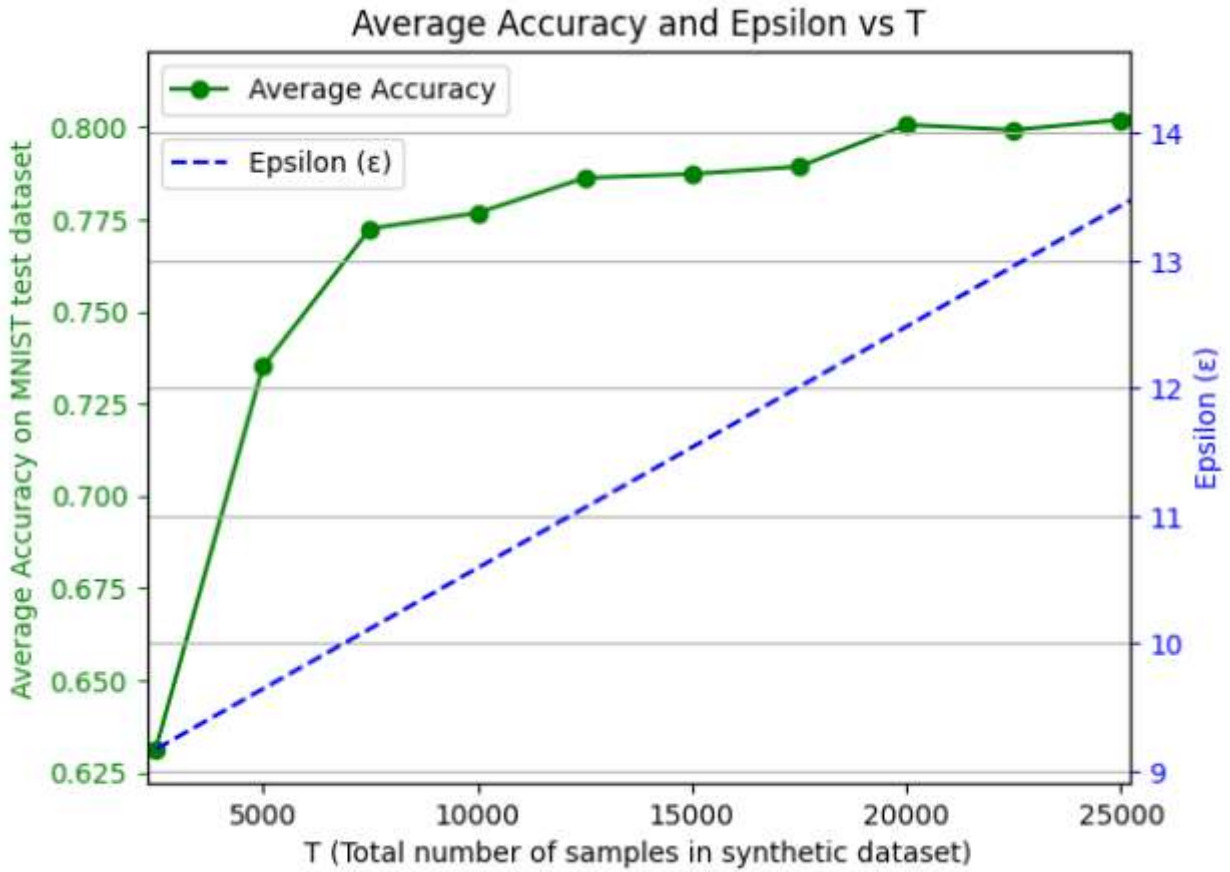
ℓ	Accuracy
32	77.50%
64	77.60%
128	79.30%
256	78.70%
512	78.60%

Fixed Values:
$\sigma_x = 0.7$, T = 12000

- ✓ For linear models increasing ℓ increases prediction accuracy.
- ✓ For nonlinear models there is a sweet spot on ℓ where the accuracy is maximum.
- ✓ In this case $\ell^* = 128$.

Results (Continued)

3. Accuracy Vs Size of Synthetic Dataset (T)



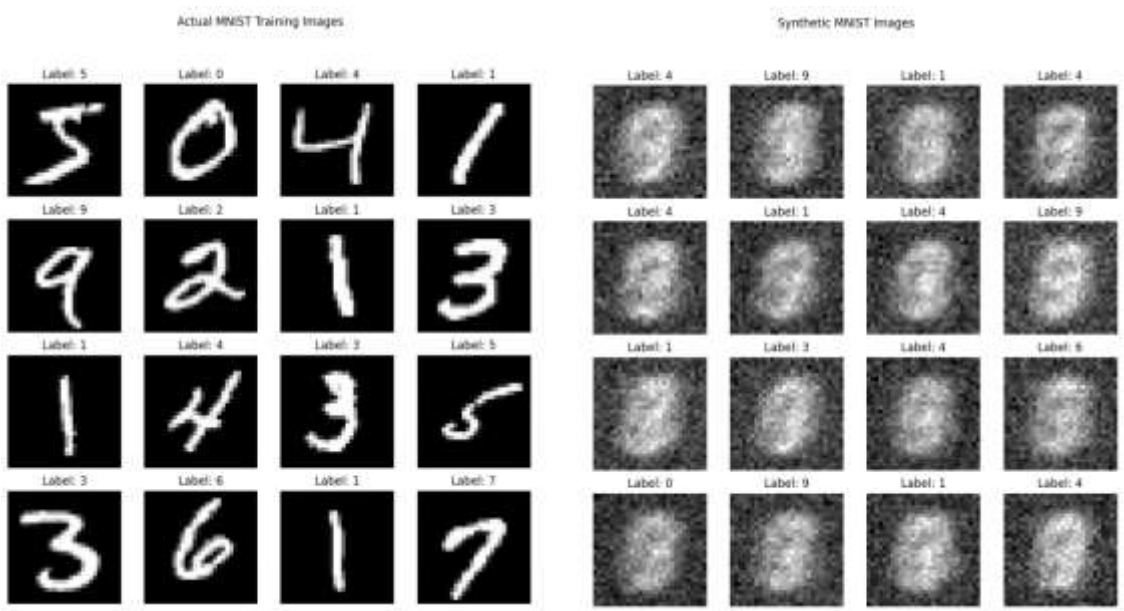
- ✓ Higher number of samples in the synthetic dataset contributes to better prediction accuracy.

Fixed Values:
$\sigma_x = 0.7$, $\ell = 128$

Outline

- Abstarct
- Literature Review
- Implementation
- Analysis
- Results
- Conclusion
- Acknowledgement and References

EEE 400: Implementation and Performance Analysis of DPMix Algorithm



Conclusion

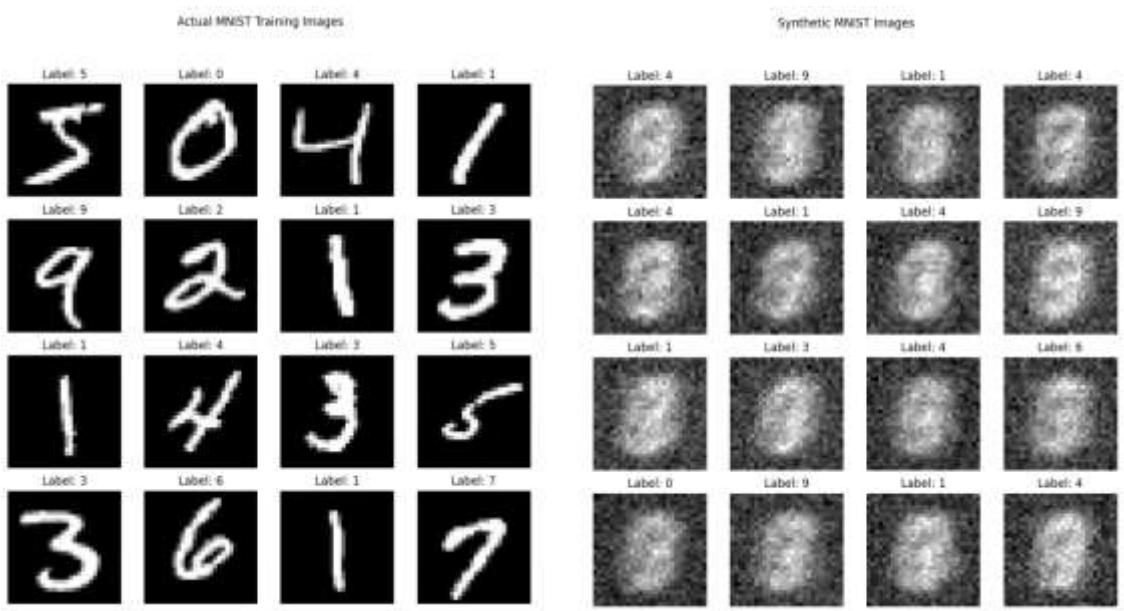
DPMix is a newly introduced algorithm that comes with the following features:

1. **Strong Differential Privacy:**
   DPMix ensures high security for large datasets with minimal additive noise.
2. **Efficient Learning:**
   Enables training of both linear and non-linear models using synthetic data.
3. **Optimized Performance:**
   Careful tuning of mixture degree-ℓ can enhance the deep learning model accuracy.
4. **Innovative Approach:**
   DPMix outperforms most of the existing data manipulation algorithms.
5. **Real-World Impact:**
   Can be one of the best choices for the tech companies, hospitals and other industries for protecting their user data.

Outline

- Abstract
- Literature Review
- Implementation
- Analysis
- Results
- Conclusion
- Acknowledgement and References

EEE 400: Implementation and Performance Analysis of DPMix Algorithm



Acknowledgement and References

ACKNOWLEDGEMENT

I sincerely thank Professor Dr. Hafiz Imtiaz for his invaluable guidance and support he provided me throughout the research work. I am also grateful to my teachers, classmates, and seniors for their encouragement and helpfulness.

Besides, I pay my respect to Kangwook Lee et al. – who came up with the idea of DPMix algorithm in their research paper titled with '*Synthesizing Differentially Private Datasets using Random Mixing*' (2019).

REFERENCES

1. http://csuh.kaist.ac.kr/Conf_Suh_RandomMix_ISIT2019.pdf
2. https://en.wikipedia.org/wiki/Differential_privacy
3. https://youtu.be/gI0wk1CXlsQ?si=Pg0HniyCc9G8Fu71

That's All!

EEE 400: Implementation and Performance Analysis of DPMix Algorithm