



NoisyDiffusion: Generation of Privacy Preserving Gene Expression Data

Jules Kreuer
Sofiane Ouaari
Prof. Dr. Nico Pfeifer





The Challenge / The Goal

- Generation of Gene Expression Data
- Data from the Cancer Genome Atlas
- Privacy preserving
 - Protect against Membership Inference Attacks



Gene Expression Data

| ID | E1 | E2 | E3 | E4 | E5 | ... | Label |
|----------|--------|--------|--------|--------|--------|-----|-------|
| Cell 1, | 7.08, | 12.14, | 12.56, | 2.50, | 15.30, | ... | A |
| Cell 2, | 9.90, | 12.17, | 7.69, | 4.54, | 12.13, | ... | A |
| Cell 3, | 7.33, | 11.38, | 5.07, | 2.58, | 5.61, | ... | B |
| Cell 4, | 3.24, | 8.43, | 0.73, | 8.44, | 15.67, | ... | A |
| Cell 5, | 9.72, | 12.67, | 0.86, | 7.36, | 13.78, | ... | C |
| Cell 6, | 7.46, | 8.97, | 6.49, | 10.81, | 4.99, | ... | D |
| Cell 7, | 8.82, | 3.02, | 7.35, | 6.50, | 13.30, | ... | D |
| Cell 8, | 3.52, | 13.65, | 14.82, | 0.51, | 8.76, | ... | B |
| Cell 9, | 5.11, | 7.11, | 9.73, | 11.46, | 3.51, | ... | C |
| Cell 10, | 12.59, | 15.21, | 0.01, | 6.92, | 12.61, | ... | D |
| Cell 11, | 10.91, | 8.12, | 14.86, | 7.47, | 6.32, | ... | A |
| Cell 12, | 9.46, | 8.13, | 6.82, | 2.39, | 11.90, | ... | D |
| Cell 13, | 3.26, | 15.61, | 3.84, | 13.96, | 10.40, | ... | C |
| Cell 14, | 2.76, | 14.22, | 11.48, | 3.72, | 15.07, | ... | B |
| Cell 15, | 7.22, | 5.54, | 0.04, | 5.9, | 3.84, | ... | B |
| ... | ... | ... | ... | ... | ... | ... | ... |

- Floats
- No order
- No reference scale



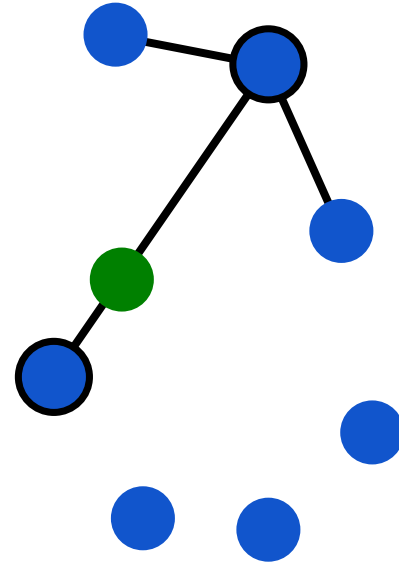
Dataset

- TCGA BRCA <1,089 individuals x 978 genes>
 - Five subtypes
 - Unbalanced distribution
- TCGA COMBINED <4,323 individuals x 978 genes>
 - 10 cancer types, including Breast, Colorectal, Esophagus, Kidney, Liver, Lung, Ovarian, Pancreatic, Prostate, and Skin
 - Unbalanced distribution



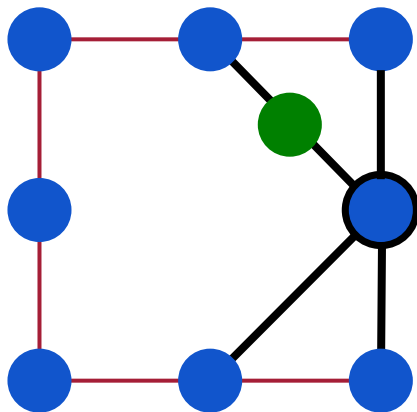
SMOTE

- Synthetic Minority Over-Sampling Technique
- Pick any sample x_i
- Pick one of its k-nearest neighbours x_{zi}
- $x_{\text{new}} = x_i + \lambda(x_{zi} - x_i)$ with random $\lambda \in (0, 1)$
- Can be problematic for certain data distributions.





SMOTE Problematic Distribution



— True distribution

- Problematic for very small datasets if k too large



Main Inspiration

Differentially Private Diffusion Models

Tim Dockhorn^{1,2,3} Tianshi Cao^{1,3,4} Arash Vahdat¹ Karsten Kreis¹

¹ NVIDIA ² University of Waterloo ³ Vector Institute ⁴ University of Toronto

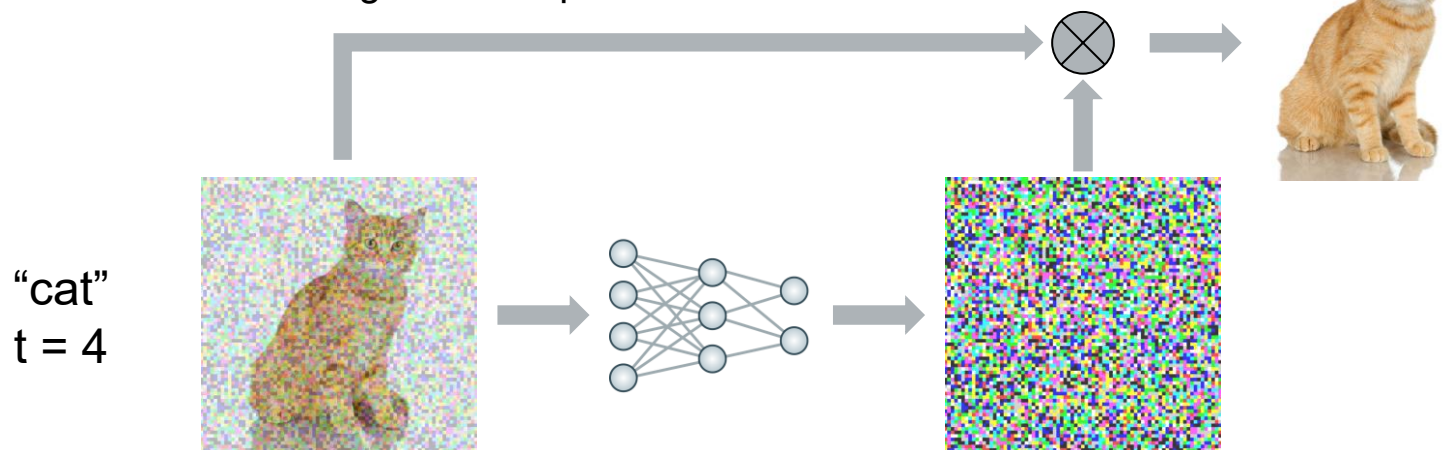
TMLR 2023

Tim Dockhorn, Tianshi Cao, Arash Vahdat, & Karsten Kreis (2023). Differentially Private Diffusion Models. *Transactions on Machine Learning Research*.



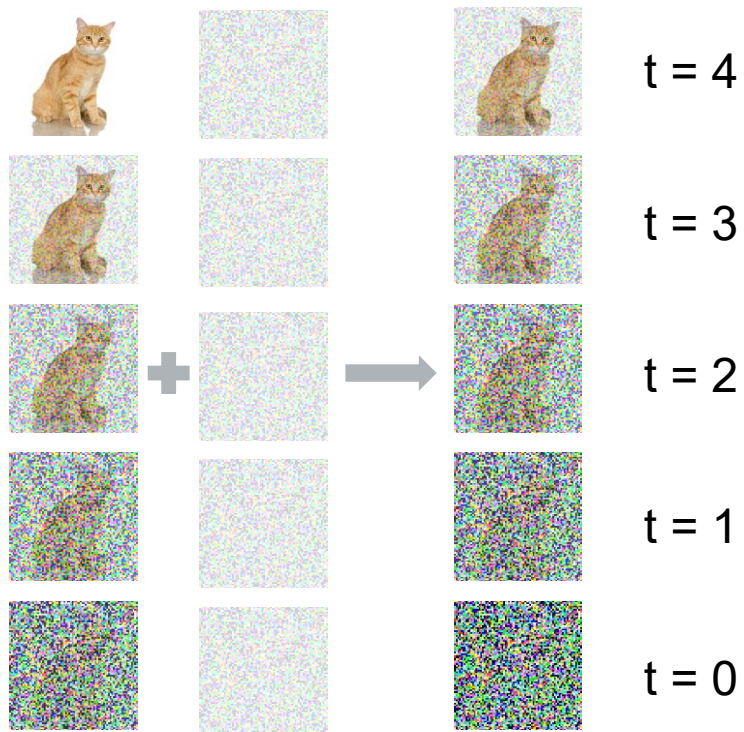
Diffusion Models

- Repeated Denoising
- Guidance
 - Embedding of image description
 - Denoising “time” step





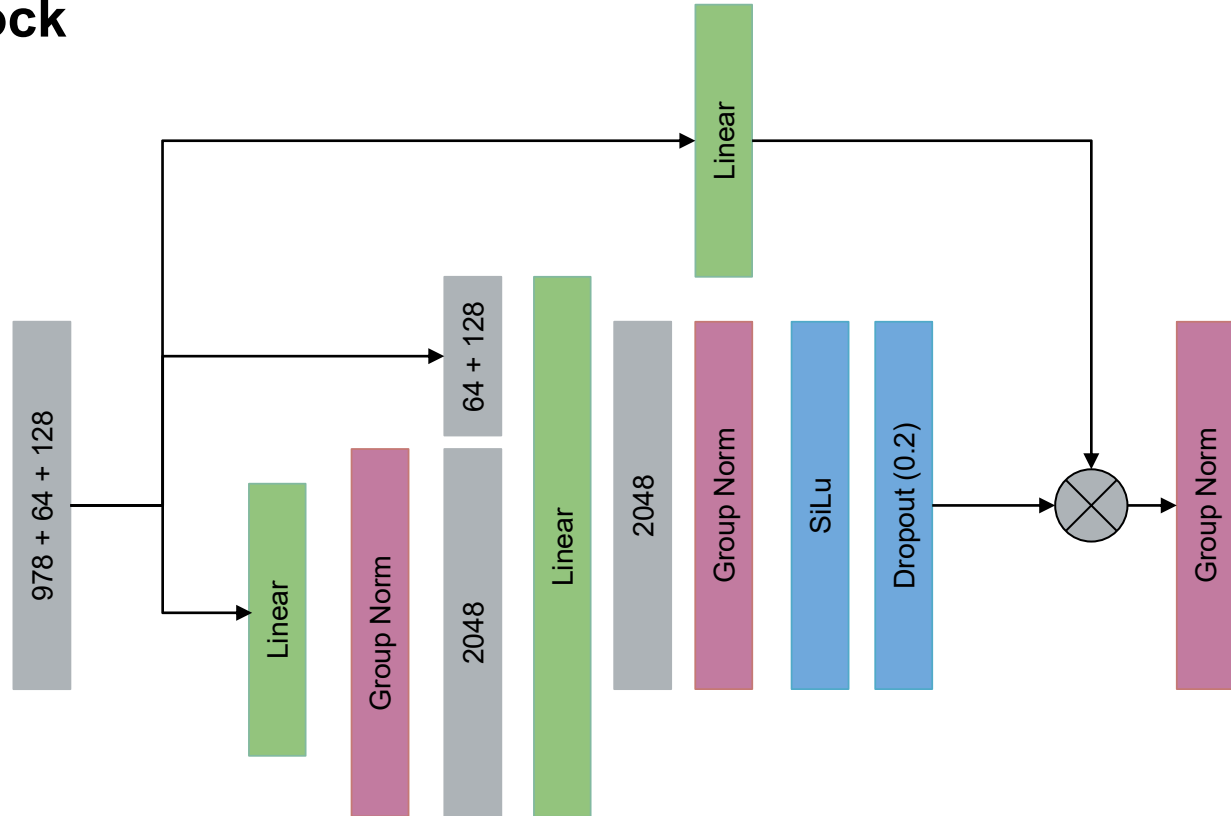
Diffusion Model: Data Generation





ResLinear Block

- Time
- Label
- Gene Expression





Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \dots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N} \sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t , noise scale σ , group size L , gradient norm bound C .

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C})$

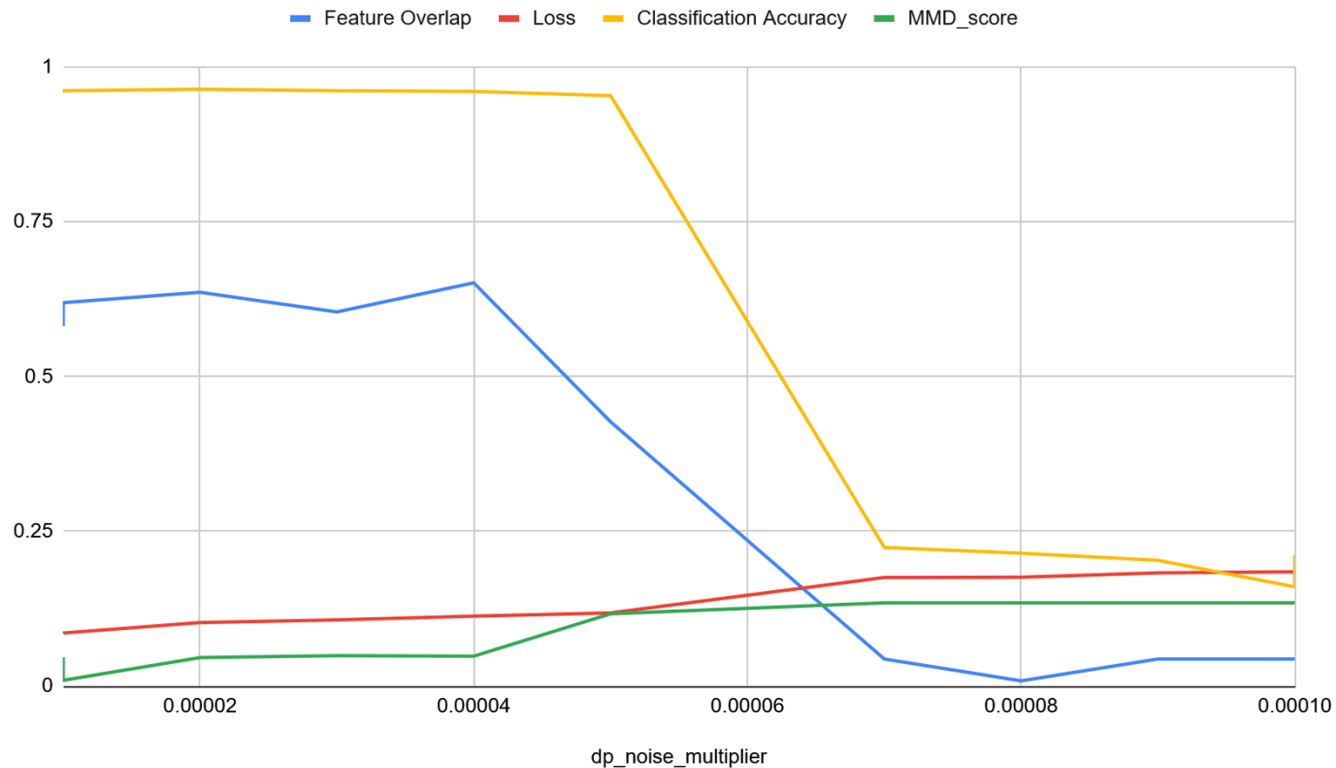
Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L} (\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}))$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ϵ, δ) using a privacy accounting method.



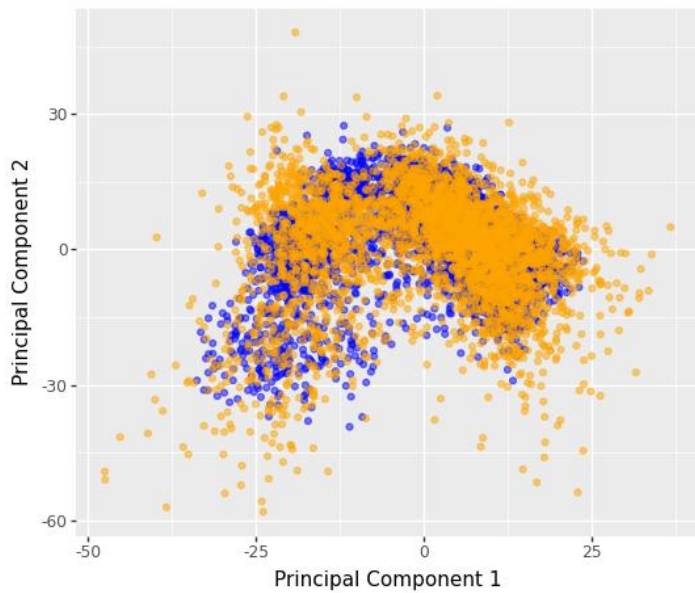
TCGA Combined; Batch size: 64; Epoch: 200; SMOTE: 3000; β_{start} : 0.001; β_{end} : 0.02; γ : 0.001; λ : 0.001



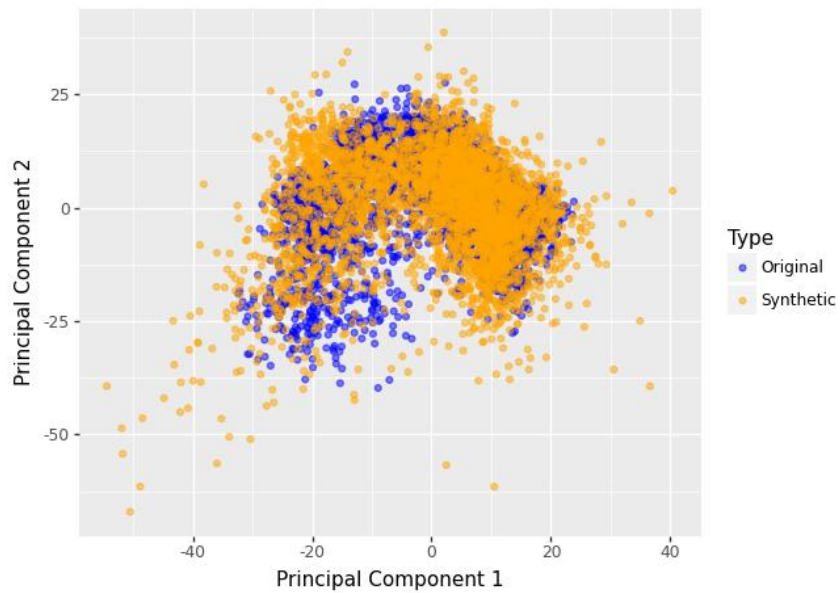
| Method | Accuracy | AUPR | Overlapping features | MMD score | Discriminative score |
|--|----------|--------|----------------------|-----------|----------------------|
| Baseline (Multivariate) | 97.52% | 99.20% | 64 | 0.0092 | 57.49% |
| NoisyDiffusion | 96.92% | 98.82% | 57.4 | 0.005 | 69.94% |
| Non-negative matrix factorization distorted input for CVAE | 95.40% | 96.74% | 46.8 | 0.0518 | 81.35% |
| Synthetic RNA-seq Data Generation with Private-PGM ($\epsilon = 10$) | 93.92% | 96.56% | 47.6 | 0.0055 | 63.57% |
| Synthetic RNA-seq Data Generation with Private-PGM | 92.88% | 96.71% | 44.2 | 0.0061 | 66.66% |
| Class-conditional diffusion model with differential privacy | 5.95% | 12.27% | 11 | 0.1349 | 99.17% |



PCA of Original and Synthetic Data



PCA of Original and Synthetic Data





| Method | MC Leaks AUC | GAN MIA AUC | MC MIA TPR FPR=0.01 | GAN MIA TPR FPR=0.01 | MC MIA TPR@FPR=0.1 |
|---|-----------------|----------------|------------------------|-------------------------|-----------------------|
| Baseline (Multivariate) | 51.44% | 51.70% | 1.20% | 1.30% | 10.10% |
| NoisyDiffusion | 50.77% | 51.21% | 1.13% | 1.19% | 10.24% |
| Non-negative matrix factorization distorted input for CVAE | 50.51% | 50.91% | 1.02% | 1.10% | 10.19% |
| Synthetic RNA-seq Data Generation with Private-PGM ($\epsilon = 10$) | 50.30% | 50.13% | 1.04% | 0.97% | 10.55% |
| Synthetic RNA-seq Data Generation with Private-PGM | 50.16% | 50.16% | 1.04% | 0.94% | 10.40% |
| Class-conditional diffusion model with differential privacy | 50.48% | 50.00% | 50.18% | 100.00% | 50.18% |



Wrapping up

- SMOTE effectively oversamples gene expression data.
 - Diffusion Models with DP-SGD represent TCGA datasets well.
 - Challenge-tuned hyperparameters will differ in real applications.
-
- Thank you!
 - Thanks to elsa and the european / german taxpayers!



More information on our group:
Uni Tübingen / Pfeiferlab



Metrics

- **MC MIA AUC (Monte Carlo Membership Inference Attack AUC)**
 - Based on the Area Under the ROC Curve (AUC).
 - A score of 0.5 is a random guess; 1.0 is a perfect attack.
 - "Monte Carlo" refers to the use of repeated random sampling for estimation.
- **GAN-leaks MIA AUC**
 - Similar to MC MIA AUC but uses a specific attack method tailored for Generative Adversarial Networks (GANs).
- **MC MIA PR AUC (Monte Carlo Membership Inference Attack PR AUC)**
 - Uses the area under the Precision-Recall (PR) curve instead of the ROC curve.
 - More informative for imbalanced datasets, which are common in these attack scenarios.
 - A higher score indicates a more successful attack.
- **GAN-leaks MIA PR AUC**
 - The Precision-Recall AUC version of the specific GAN-leaks attack methodology.
- **MC / GAN-leaks MIA TPR@FPR=0.01**
 - Measures the True Positive Rate (TPR), or the percentage of correctly identified training members.
 - Calculated at a fixed, low False Positive Rate (FPR) of 1%.
 - A high TPR value indicates a significant privacy leak.
- **MC / GAN-leaks MIA TPR@FPR=0.1**
 - False Positive Rate fixed at a less strict 10%.
 - Shows attack performance at a different trade-off level between true and false positives.



```

t = torch.randint(0, self.num_timesteps, (x.shape[0],), dtype=torch.long) # Sample a random timestep
noisy_x, target_noise = self.get_noise_schedule(t, x)

predicted_noise = model(noisy_x, t.float(), labels) # Forward pass

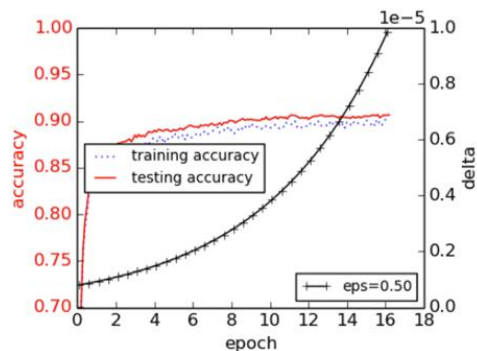
loss = F.mse_loss(predicted_noise, target_noise)
optimizer.zero_grad() # Backward pass
loss.backward()

torch.nn.utils.clip_grad_norm_(model.parameters(), max_grad_norm) # First clip gradients

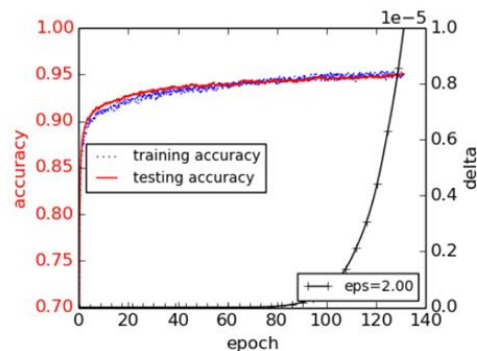
for param in model.parameters(): # Add noise proportional to the gradient norm
    if param.requires_grad and param.grad is not None:
        noise = torch.normal(
            mean=0,
            std=dp_noise_multiplier * max_grad_norm,
            size=param.grad.shape)
        param.grad += noise

optimizer.step() # Update weights

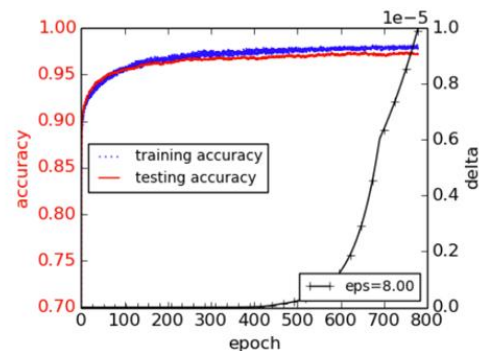
```



(1) Large noise



(2) Medium noise



(3) Small noise

Figure 3: Results on the accuracy for different noise levels on the MNIST dataset. In all the experiments, the network uses 60 dimension PCA projection, 1,000 hidden units, and is trained using lot size 600 and clipping threshold 4. The noise levels (σ, σ_p) for training the neural network and for PCA projection are set at $(8, 16)$, $(4, 7)$, and $(2, 4)$, respectively, for the three experiments.

| Date | Method | Accuracy (real) | Accuracy (synthetic) | AUPR (real) | AUPR (synthetic) | Number of overlapping important features | MMD score | Discriminative score | Distance to the closest (real) | Distance to the closest (synthetic) | MC MIA AUC | GAN-leaks MIA AUC |
|------------|---|-----------------|----------------------|-------------|------------------|--|-----------|----------------------|--------------------------------|-------------------------------------|------------|-------------------|
| 2025-03-15 | Class-conditional diffusion model with differential privacy | 87.33% | 9.91% | 87.14% | 24.44% | 1.8 | 0.2697 | 99.80% | 24.0301 | 9,887,871.1730 | 49.37% | 50.00% |
| 2025-03-15 | NoisyDiffusion | 87.05% | 77.50% | 85.70% | 75.96% | 18.2 | 0.0109 | 60.39% | 24.0183 | 25.1569 | 52.33% | 53.95% |
| 2025-03-14 | Synthetic RNA-seq Data Generation with Private-PGM | 86.23% | 81.54% | 87.10% | 76.30% | 15.6 | 0.0086 | 86.38% | 24.0422 | 27.2186 | 50.16% | 50.52% |
| 2025-03-08 | Non-negative matrix factorization distorted input for CVAE | 85.67% | 82.09% | 86.00% | 74.57% | 15.4 | 0.0228 | 83.43% | 24.0493 | 26.4565 | 50.62% | 51.34% |
| 2025-03-16 | Baseline (Multivariate) | 86.41% | 82.09% | 85.77% | 83.42% | 20.6 | 0.0166 | 54.36% | 24.0532 | 28.3770 | 52.33% | 52.79% |
| 2025-03-16 | Synthetic RNA-seq Data Generation with Private-PGM (e = 10) | 86.23% | 82.92% | 87.10% | 77.58% | 15 | 0.0074 | 77.69% | 24.0422 | 27.0686 | 50.42% | 50.18% |

TCGA BRCA



Die Farben der Universität Tübingen

