PAC Privacy: Automatic Privacy Measurement and Control of Data Processing

Hanshen Xiao

Srini Devadas

Massachusetts Institute of Technology

MIT CSAIL

Information leakage model

Information leakage from arbitrary disclosure can be described as the following model:

$$X \rightarrow M(X) \in \mathbb{R}^d$$

- X: data/messages/files

- $M(X)$, the information disclosed
  - Statistics (mean/median) of a sensitive dataset X
  - Neural networks learned from samples X
  - Side channel info: traffic/memory patterns X

What is privacy?

In words, adversary cannot guess (recover) your secret correctly (approximately correctly)

- For what kind of adversary and with what kind of power
  - Computation restriction
  - Prior knowledge
- Mathematical quantification of the inference hardness
  - Impossibility of what kind of inference task
  - Measurement of the hardness

Classic Security Definitions
Data-independent privacy/security

- Shannon Perfect Secrecy (statistical indistinguishability): for any possible inputs X and X', the distributions of $M(X)$ and $M(X')$ are identical
- Computational Indistinguishability: for any possible inputs X and X', the distributions of $M(X)$ and $M(X')$ are indistinguishable for a computationally-bounded adversary
- Differential Privacy (DP): for any two adjacent datasets X and X', the divergence of distributions under some divergence function $D_\alpha(\mathrm{P}_{M(X)} \parallel \mathrm{P}_{M(X')})$ is bounded

A high-level picture of PAC Privacy:
Instance-based privacy

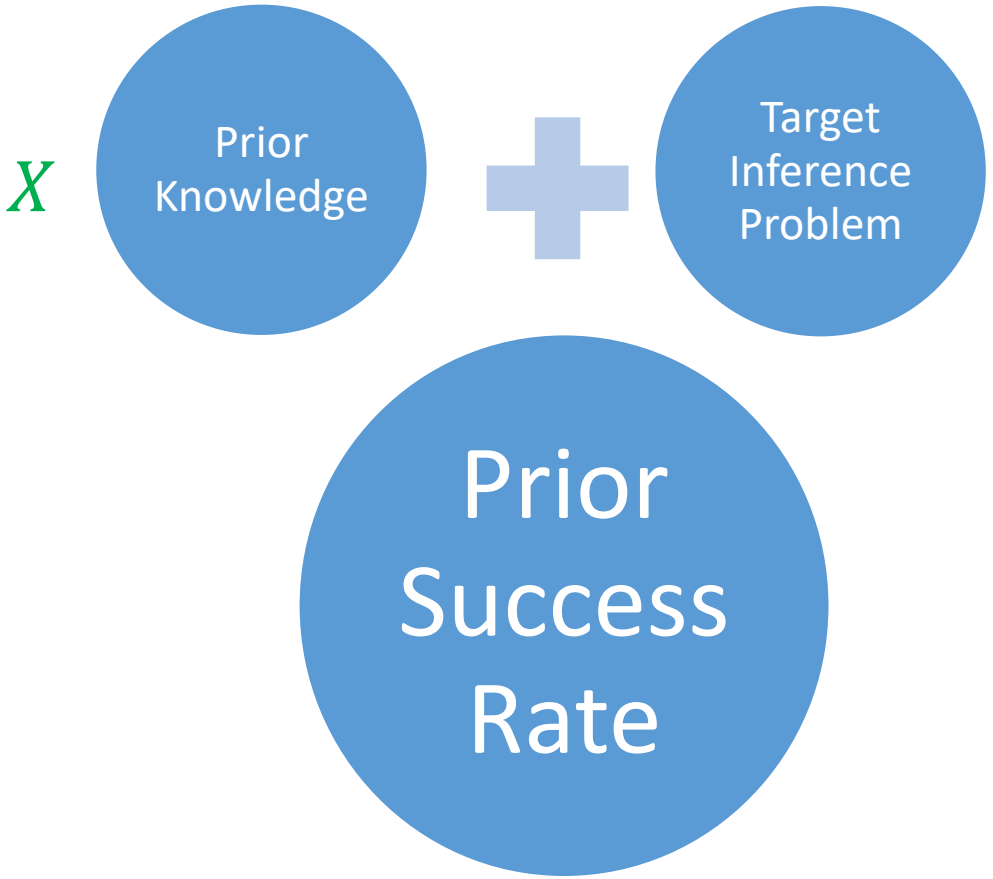
A bridge between semantic privacy interpretation and mathematical quantification

A high-level picture of PAC Privacy:
Instance-based privacy

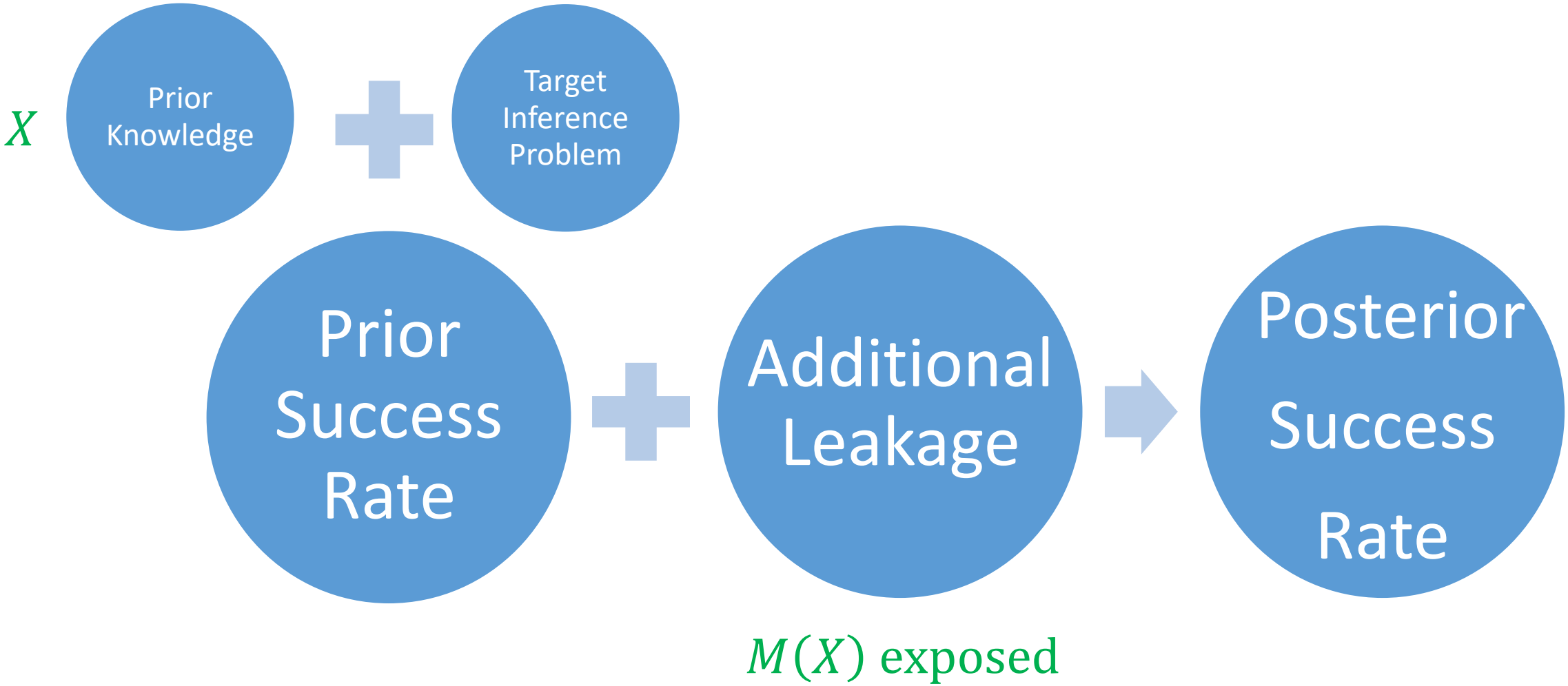
A bridge between semantic privacy interpretation and mathematical quantification

1. Determine adversary's prior knowledge on X
   - The adversary knows all public parameter setups
   - The adversary knows your secret images are about pets or portraits
2. Set an adversarial inference task of interest and a lower bound of failure rate
   - Adversary cannot guess one bit of X correctly with probability more than ¾
   - Adversary cannot recover any single sample of X with error in l_2 norm smaller than 1 with probability more than ½
3. Provide a privacy-preserving scheme on the objective processing function M

Workflow to quantify inference hardness



Workflow to quantify inference hardness



Mathematical preparation to quantify inference hardness

- Adversary's guessing $\tilde{X}$ on sensitive data X
- A success criterion ρ: $\rho(\tilde{X}, X) = 1$ iff the adversary produces satisfied inference
- Optimal prior success rate: $1 - \delta_o^\rho$ or minimal prior failure rate δ_o^ρ, determined by ρ and prior knowledge
- Posterior success rate $1 - \delta$: the adversary can return satisfied $\tilde{X}$, such that $\rho(\tilde{X}, X) = 1$, after observing the release $M(X)$ with probability $1 - \delta$
- Posterior advantage is $\delta_o^\rho - \delta$

PAC Privacy:

Instance-based privacy

We borrow the idea of PAC learning and describe the attack as a learning problem.

Definition 1 [(δ, ρ, D) PAC Privacy]. For a data processing mechanism M, given some data distribution D, and a measurement/criterion $\rho(\cdot, \cdot)$, we say M satisfies (δ, ρ, D)-PAC Privacy if the following experiment is impossible:

A user generates data X from distribution D and sends $M(X)$ to an adversary. The adversary who knows D and M is asked to return an estimation $\tilde{X}$ on X such that with probability at least $(1-\delta)$, $\rho(\tilde{X}, X) = 1$

Mutual Information and Entropy

- Mutual information is extensively studied in information theory
  - For two random variables x and w in some joint distribution, the mutual information $MI(x;w)$ is defined as
$$MI(x;w) = D_{KL}(P_{x,w} \parallel P_X \otimes P_w)$$
i.e., the KL-divergence between the joint distribution of (x, w) and the product of the marginal distributions of x and w, respectively

- Equivalently, mutual information can also be expressed by entropy:
$$MI(x;w) = \mathrm{H}(\mathrm{x}) - \mathrm{H}(\mathrm{x}|\mathrm{w}) = \mathrm{H}(\mathrm{w}) - \mathrm{H}(\mathrm{w}|\mathrm{x})$$

Bounding Posterior Advantage

The Bridge: Posterior advantage can be captured by f-divergence

Theorem: For any processing function $M : X^* \to Y$, and any f-divergence,

$$\Delta_f \delta = D_{\mathrm{f}}\left(1_\delta \parallel 1_{\delta_o^\rho}\right) = \inf_{P_W} D_{\mathrm{f}}\left(P_{(X,M(X))} \parallel P_X \otimes P_w\right),$$

for any random variable $\mathrm{w} \in Y$

When we select D_{f} to be KL-divergence and $P_w = P_{M(X)}$, then

$$\Delta_{KL}\delta \leq MI(X; M(X))$$

R. H. S. does not have ρ

Noise is not necessary for PAC Privacy

When data is of sufficient entropy and the processing has a closed form

Noise is not necessary for PAC Privacy

When data is of sufficient entropy and the processing has a closed form

Example 1 [Mean estimation of Gaussian data]: Suppose a sensitive data $x \sim \mathrm{N}(0,1)$ and other $(n-1)$ i.i.d. samples $x_1, x_2, \ldots, x_{n-1} \sim \mathrm{N}(0,1)$ are used to produce a mean estimation

$$M(x) = \frac{1}{n} \cdot \left(\sum_{i=1}^{n-1} x_i + x\right)$$

Then,

$$MI\big(x; M(x)\big) = H\big(M(x)\big) - H(M(x)|x) = 0.5 \log\left(1 + \frac{1}{n-1}\right)$$

Can use this bound to show privacy

Differentially-Private (Input-Independent) Mean Estimation

$M(X) = Mean(X)$, where each of n records is a scalar in $[0, 1]$

Global Sensitivity of $M = 1/n$

Laplace Mechanism:

Output $M(X) + Z$, where $Z \sim 1/(n\varepsilon)\ \mathrm{Lap}(0, 1)$

No noise → no privacy, $\varepsilon = \infty$

Automatic Privacy Analysis

Automatic Privacy Analysis

- We have reduced the privacy proof to control $MI(X; M(X))$
- Noise B, especially continuous noise, can help us derive tractable upper bound of $MI(X; M(X) + B)$ in the general case
- What we want:
  - The processing mechanism M can be a black-box: no algorithmic analysis is needed
  - Automatic privatization protocol: when $MI(X; M(X))$ is not sufficiently small, we can automatically generate a scheme to perturb $M(X)$ until it produces satisfied security parameters
  - In particular, if we only focus on the posterior advantage, the data distribution/generation can also be black-box

Theorem: Automatic Analysis

$d \times d$ Covariance Matrices

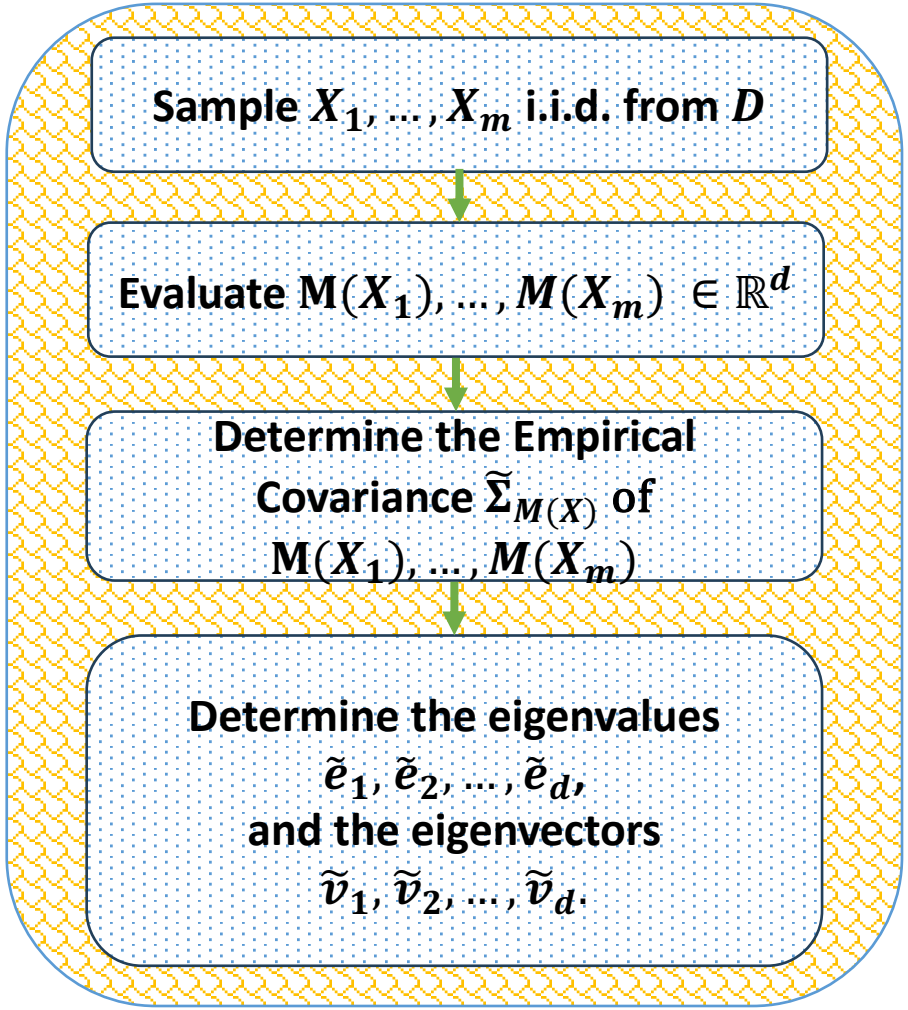
Theorem: For an arbitrary deterministic mechanism M and a Gaussian noise $B \sim \mathrm{N}(0, \Sigma_B)$

$$MI(X; M(X) + B) \leq \frac{1}{2} \log \det\left(I_d + \Sigma_{M(X)} \Sigma_B^{-1}\right)$$

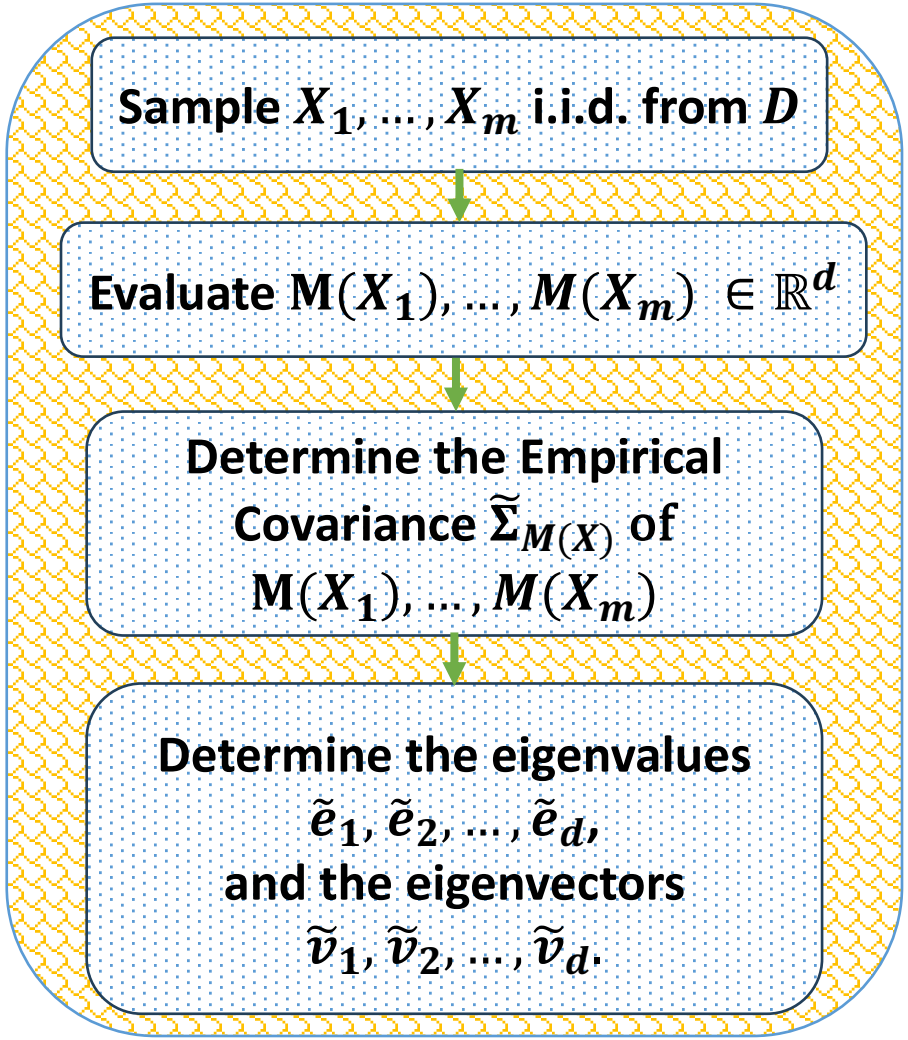
Let the eigenvalues of $\Sigma_{M(X)}$ be $(\lambda_1, \ldots, \lambda_d)$, then there exists some Σ_B such that $\mathrm{E}[\|B\|] \leq (\sum_{j=1}^{d} \sqrt{\lambda_j})^2$ and $MI(X; M(X) + B) \leq \frac{1}{2}$

- The noise B fits the geometry of the distribution of $M(X)$
- The magnitude of noise B is not explicitly dependent on the dimension: when $\sum_{j=1}^{d} \sqrt{\lambda_j} = \mathrm{O}(1)$, we only need to add constant noise

Main Algorithm (I): Learning from your data and the processing



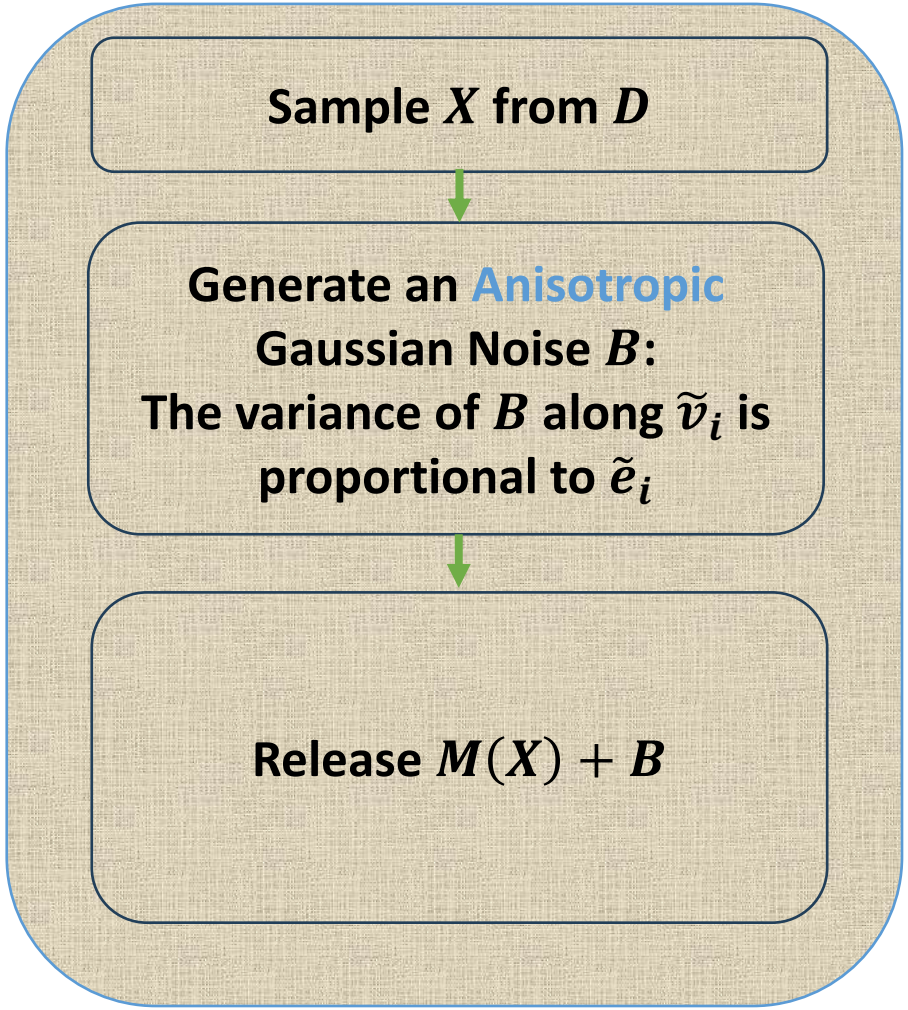
Main Algorithm (I): Learning from your data and the processing



"Fake" data: $X_1, \dots, X_m$

Eigenvalues and Eigenvectors: The power of output distribution along each direction in $\mathbb{R}^d$

Main Algorithm (II): Confident Release



Necessary noise along each direction $\tilde{v}_i$ proportional to the distribution power of $M(X)$ along $\tilde{v}_i$

Examples

Supervised Learning Toy Example

End-to-end privacy analysis:

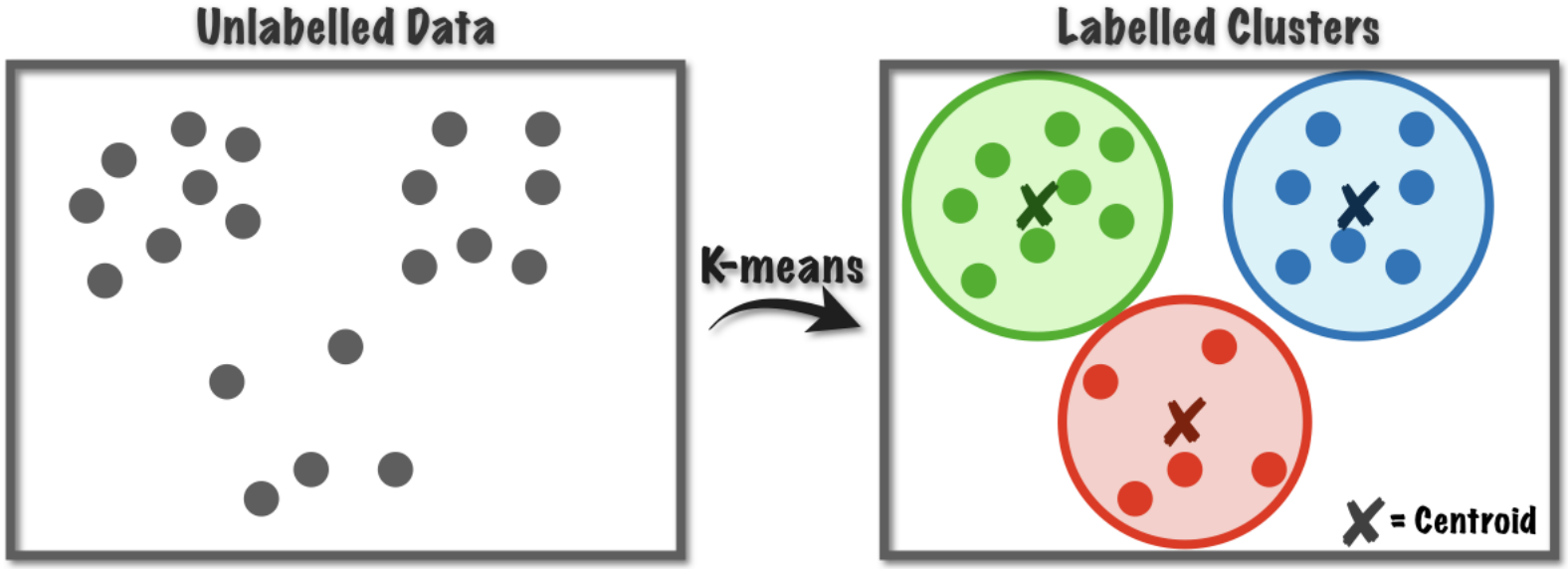
Black-box deep learning algorithm

- Train a three-layer fully-connected neural network on the MNIST dataset, which contains 70, 000 28×28 handwritten-digit images
- Data generation X by randomly sampling 35, 000 samples out of the entire data set
- **Strong interpretation:** even if the adversary knows the universe, he cannot identify/recover your sensitive data used for a trained model

Supervised Learning Toy Example - 2

- Small noise but strong privacy for the entire set
  - An independent Gaussian noise $\mathrm{E}[\|B\|] = 3.7$ is sufficient to ensure *MI(X; M(X) + B) ≤ 1*
- Non-privately, the trained-out neural network achieves 94.8% classification accuracy
- Under the perturbation to ensure PAC privacy, we achieve an accuracy of 93.5%

K Means Clustering



K Means Clustering

Given a set of observations $S = \{x_1, \dots, x_n\}$, we aim to partition S into K subsets, $S_1, S_2, \dots, S_k$, whose means are $\mu_1, \dots, \mu_k$ such that

$$\arg_S \min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \quad (*)$$

- Output of K centroids;
- Given a selection of $(\mu_1, \dots, \mu_k)$, to minimize (*), each x_i should be assigned to the closest clustering.

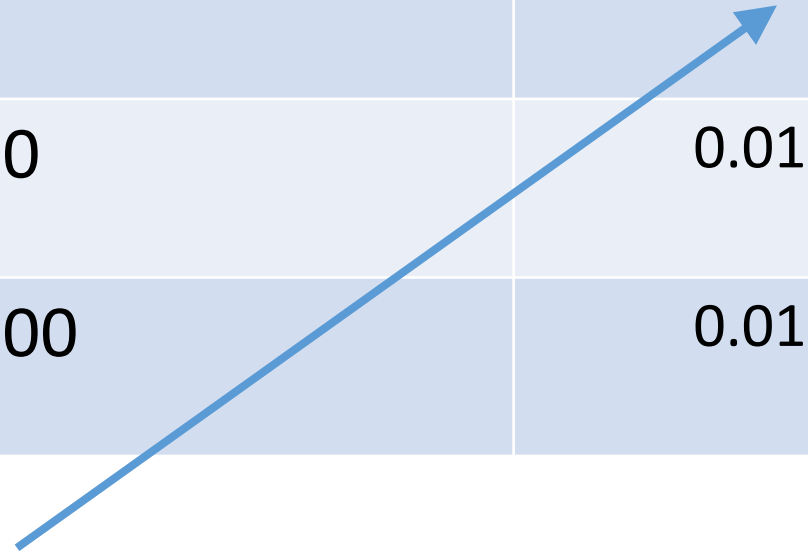
Algorithm Specifics

- Run black-box K-means algorithm on a n-subsampled subset of the entire MNIST dataset
  - Use resultant K centroids to determine clustering for all data points
  - Random initialization in non-convex optimization leads to local minima
  - Different random initializations result in large output variation **even for the same input data**
  - To improve robustness, standard strategy is to average the results over many random trials, so results don't strongly depend on random initialization
- Averaging also reduces variance of results for different subsamplings

How do Parameters Influence Stability? K Means on MNIST

For a fixed $MI(X; M(X))$=1:

# Selected Samples n \# K clusterings	K = 2	K=5	K=10
n = 1,000	0.034	0.117	0.562
n = 5,000	0.013	0.047	0.313
n = 10,000	0.012	0.042	0.250



L2 norm of noise divided by L2 norm of K centroids

Final Observations

- Need to assume private data is from some distribution
  - Conservative strategy: Assume data is public, and subsample to produce private data with entropy
  - Distribution can be arbitrarily complex – just need to sample from it
  - Benefits: O(1) noise, automatic privacy analysis

- PAC Privacy can be used to define notions of algorithm stability, and small mutual information implies low generalization error
  - Stable, private, generalizable machine learning models?