Carnegie Mellon University

ML MACHINE LEARNING DEPARTMENT

Generalization and Learnability under Differential Privacy and its Variants

Yu-Xiang Wang

Based on joint works with
Jing Lei and Steve Fienberg

Bottleneck of machine learning?

Privacy law

Food safety

Understand how brain works

Better treatment

Better education

Medical diagnosis

Clean energy

movie recommendation

Fraud detection

Increasing privacy awareness

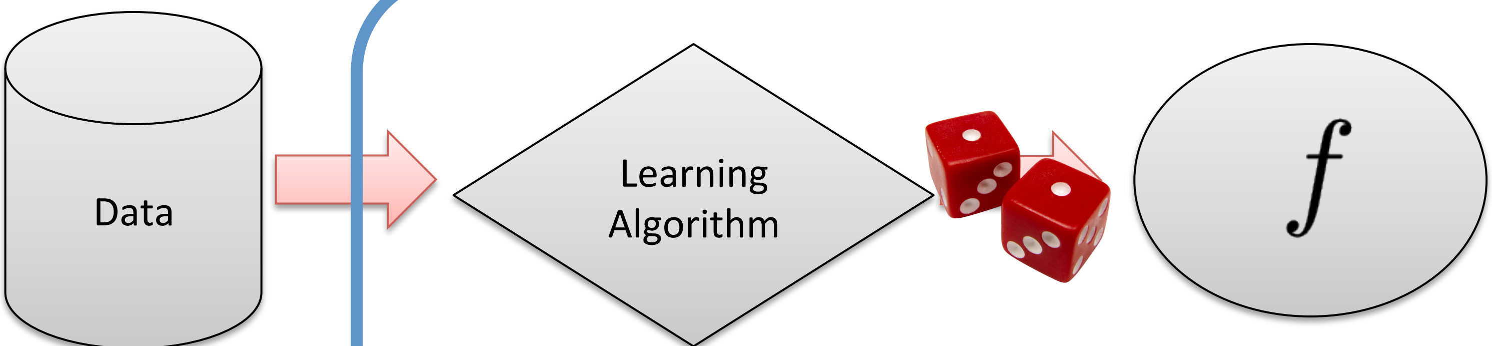
The second Netflix Prize cancelled to settle a lawsuit

NYC’s Taxi data set breached

Vijay Pandurangan posts:

Differentially private machine learning

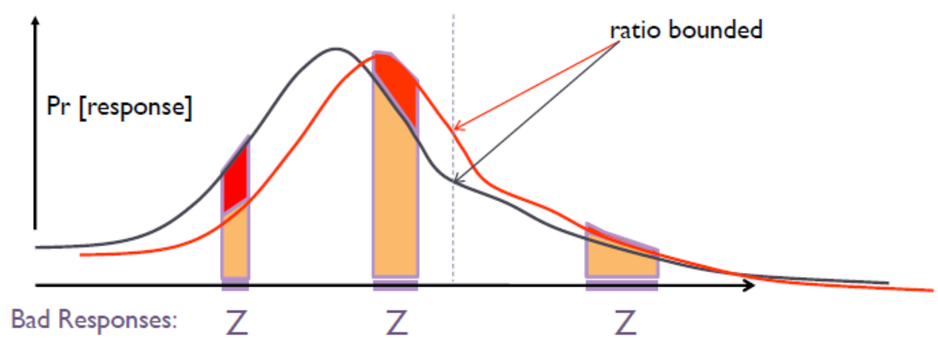
Randomized algorithm



Feature-label pairs
Unlabeled features
Feature points

Support Vector Machine
K-means clustering
Kernel density estimation

Classifier
K cluster centers
Estimated density function



Example: Recommendation System

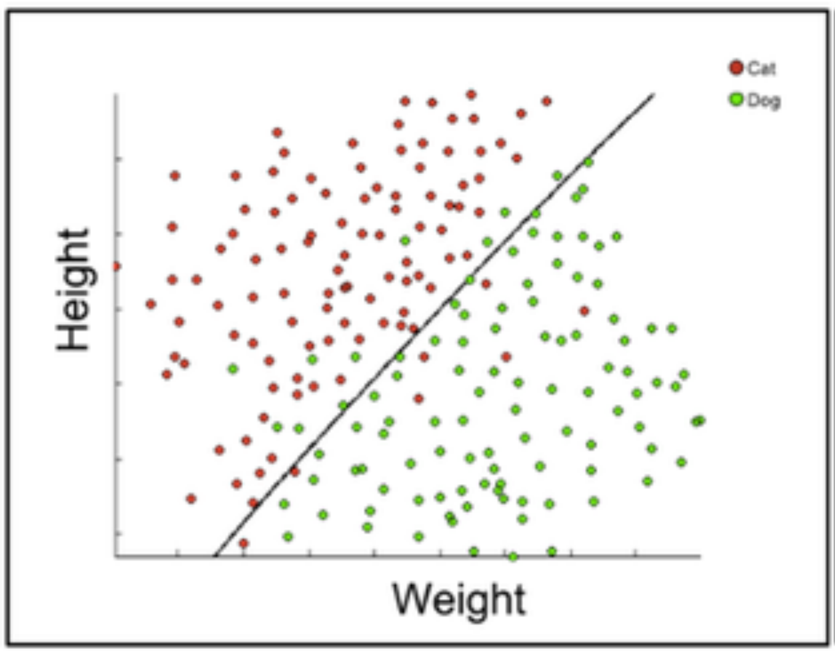
- Model based collaborative filtering.
  - Learning: $f \leftarrow A(Z)$ uses all user data.
  - Prediction: $y_i \leftarrow f(z_i)$ uses f and his own data.

- Setting:
  - Trust the service provider. Netflix is not an adversary.
  - Other users might be adversaries.

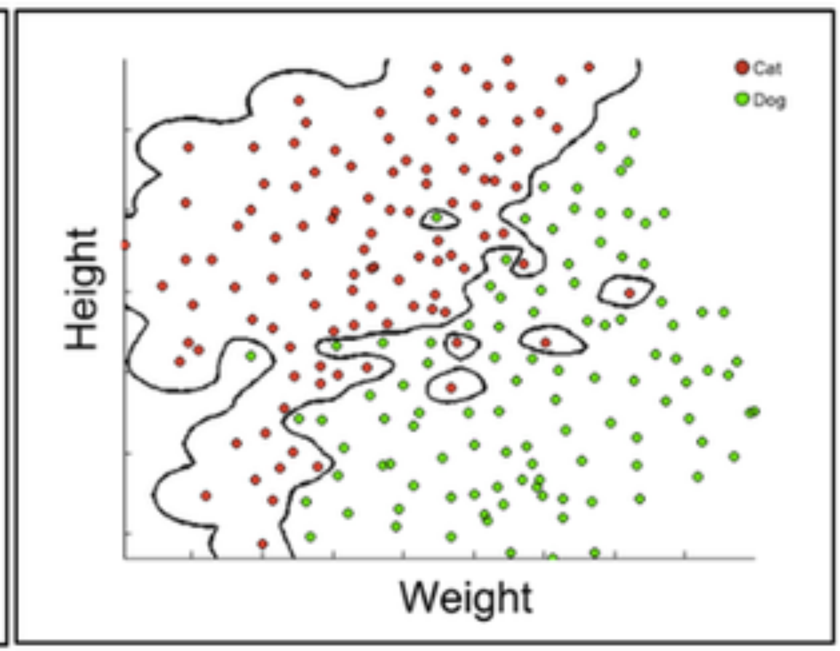
- If A is private, prediction is "post-processing".

Synergy between learning and privacy

Underfitting
f is parsimonious
Private information compressed.



Overfitting
f memorizes the dataset
Knowing f breaches privacy



Plan today

- Revisiting “What can be learned privately?”
  - Vapnik’s General Learning Setting
  - Characterizing private learnability
- To what extent can we weaken DP?
  - But still preserve the basic property that “privacy => generalization”.
  - A characterization of on-avg generalization.

Notations

- Data domain $\mathcal{Z}$ Or $\mathcal{X} \times \mathcal{Y}$
- Hypothesis class $\mathcal{H}$
- Loss Function : $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$
- Task: find $h \in \mathcal{H}$ with low risk.

PAC Learning vs. General Learning Setting

(Agnostic) PAC Learning: Binary classification.

- Logistic Regression
- Stochastic convex optimization
- Generalized Linear model
- Linear Regression
- RKHS Learning Kernel SVM
- K-means clustering
- Multiclass classification
- Density estimation
- Matrix factorization
- Recommender system

General Learning Setting

- Reinforcement Learning
- Online learning

Problems in general learning setting

An illustration of problems in the General Learning setting.

Problem	Hypothesis class $\mathcal{H}$	$\mathcal{Z}$ or $\mathcal{X} \times \mathcal{Y}$	Loss function ℓ
PAC Learning	$\mathcal{H} \subset \{f : \{0,1\}^d \to \{0,1\}\}$	$\{0,1\}^d \times \{0,1\}$	$1(h(x) \neq y)$
Regression	$\mathcal{H} \subset \{f : [0,1]^d \to \mathbb{R}\}$	$[0,1]^d \times \mathbb{R}$	$\|h(x) - y\|^2$
Density Estimation	Bounded distributions on $\mathcal{Z}$	$\mathcal{Z} \subset \mathbb{R}^d$	$-\log(h(z))$
K-means Clustering	$\{S \subset \mathbb{R}^d : \|S\| = k\}$	$\mathcal{Z} \subset \mathbb{R}^d$	$\min_{c \in h} \\|c - z\\|^2$
RKHS classification	Bounded RKHS	RKHS$\times\{0,1\}$	$\max\{0, 1 - y\langle x, h\rangle\}$
RKHS regression	Bounded RKHS	RKHS$\times\mathbb{R}$	$\|\langle x, h\rangle - y\|^2$
Sparse PCA	Rank-r projection matrices	$\mathbb{R}^d$	$\\|hz - z\\|^2 + \lambda\\|h\\|_1$
Robust PCA	All subspaces in $\mathbb{R}^d$	$\mathbb{R}^d$	$\\|\mathcal{P}_h(z) - z\\|_1 + \lambda \mathrm{rank}(h)$
Matrix Completion	All subspaces in $\mathbb{R}^d$	$\mathbb{R}^d \times \{1,0\}^d$	$\min_{b \in h} \\|y \circ (b - x)\\|^2 + \lambda \mathrm{rank}(h)$
Dictionary Learning	All dictionaries $\in \mathbb{R}^{d \times r}$	$\mathbb{R}^d$	$\min_{b \in \mathbb{R}^r} \\|hb - z\\|^2 + \lambda\\|b\\|_1$
Non-negative MF	All dictionaries $\in \mathbb{R}_+^{d \times r}$	$\mathbb{R}^d$	$\min_{b \in \mathbb{R}_+^r} \\|hb - z\\|^2$
Subspace Clustering	A set of k rank-r subspaces	$\mathbb{R}^d$	$\min_{b \in h} \\|\mathcal{P}_b(z) - z\\|^2$
Topic models (LDA)	$\{\mathbb{P}(\text{word}\|\text{topic})\}$	Documents	$-\max_{b \in \{\mathbb{P}(\text{Topic})\}} \sum_{w \in z} \log \mathbb{P}_{b,h}(w)$

Learnability and Private Learnability

A **learning algorithm**: $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ is consistent for distribution $\mathcal{D}$ if

$$\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z^n)}\ell(h,z) \to \min_{h\in\mathcal{H}} \mathbb{E}_z \ell(h,z).$$

Definition 1 (Learnability) *A learning problem $(\mathcal{Z}, \mathcal{H}, \ell)$ is learnable if there exists an algorithm $\mathcal{A}$ and rate $\xi(n)$, such that $\mathcal{A}$ is consistent with rate $\xi(n)$ for any distribution $\mathcal{D}$ defined on $\mathcal{Z}$. [Note: this is agnostic, distribution-free learning!]*

Private Learnability: Learnable by an ϵ-DP algorithm, for an $\epsilon < \infty$

Defining Stability and AERM

- Stability: Any adjacent Z and Z', the difference in the expected risk $\rightarrow$ 0 as n $\rightarrow$ ∞.

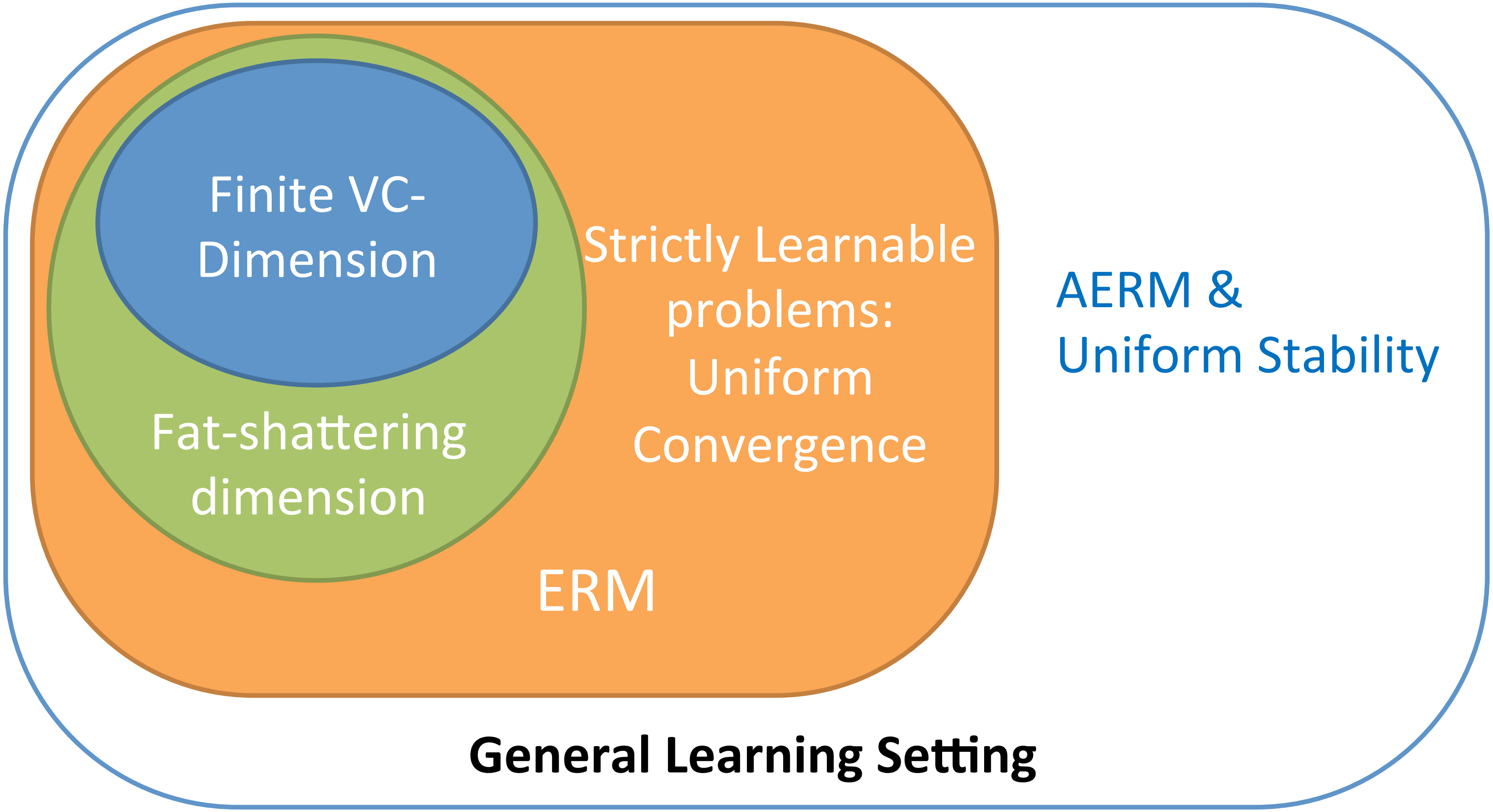
- AERM: the estimate minimizes empirical risk as n $\rightarrow$ ∞.

What is known in non-private setting?

- PAC Learning (Binary Classification)
  - finite VC-dimension ⇔ Learnability (BEHW–89)
  - Achieved by ERM.

- General Learning Setting
  - Strict Learnable by ERM ⇔ Uniform Convergence (Vapnik–95)
  - ∃ a problem learnable, but ERM fails. (SSSS-10)
  - AERM + Stability ⇔ Learnability (SSSS-10)

What is known in non-private setting?



What is known about private learnability?

- PAC Learning (on discrete domain):
  - SQ = Private SQ (BDMN-08)
  - PAC = Private PAC (KLNRS-08)
  - sample complexity on realizable setting (BNS-13).

- DP extensions of specific problems, or classes of problems.
  - (CMS-11, KST-12, BST-14) and many more.

What is known about private learnability?

SQ = PSQ. (``SuLQ'', Blum et. al. 05)

PAC = PPAC
"What can we learn privately?"

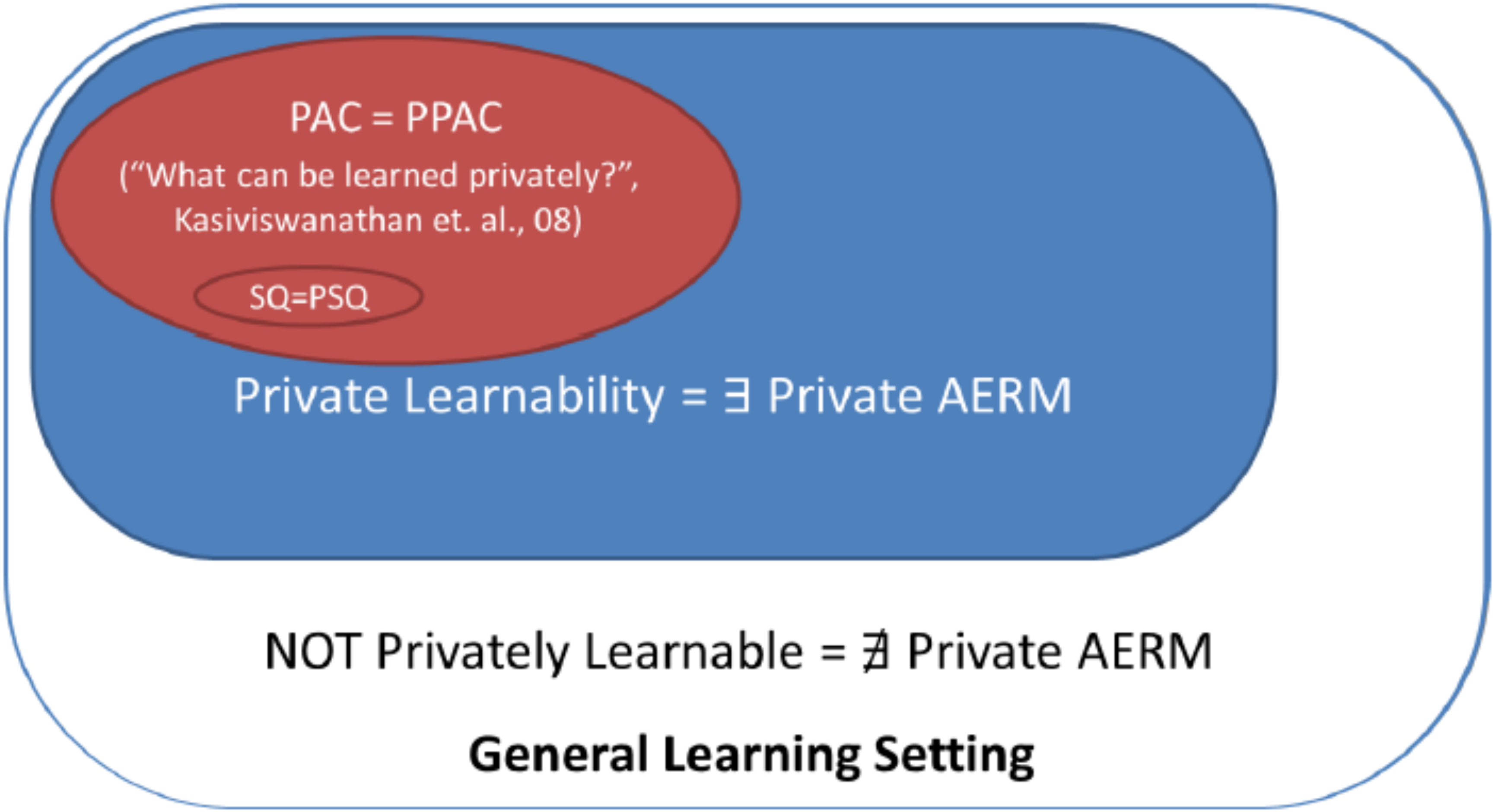
SQ

- Logistic Regression
- Generalized Linear model
- Stochastic convex optimization
- Linear Regression
- RKHS Learning Kernel SVM
- K-means clustering
- Multiclass classification
- Density estimation
- Matrix factorization
- Recommender system

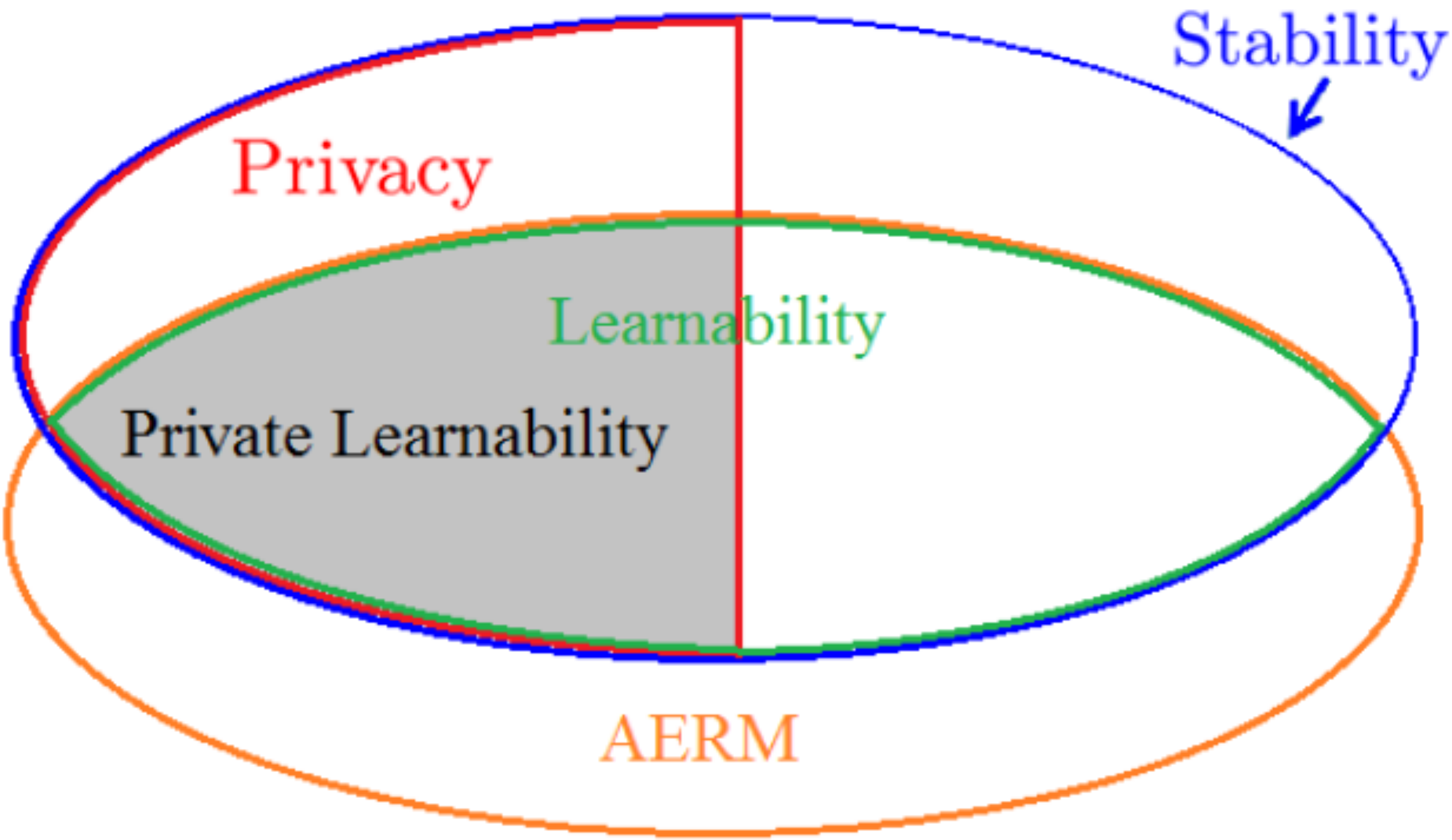
General Learning Setting

- Reinforcement Learning
- Online learning

Our result


PAC = PPAC
("What can be learned privately?", Kasiviswanathan et. al., 08)
SQ=PSQ
Private Learnability = ∃ Private AERM
NOT Privately Learnable = ∄ Private AERM
General Learning Setting

Characterizing Private Learnability



Key ideas of the proof

Subsampling Lemma [BKN-13]:

If A is ϵ-DP on Z of size n.

Then running A on a random subsample of Z with γn data points is $2\gamma\exp(\epsilon)$-DP.

Take $\gamma = 1/\text{sqrt}(n)$

Key ideas of the proof

- Forward direction:
  - Private and AERM
  - Random subsample data (so that Privacy -> 0)
  - **Privacy => Stability => Generalization** (appeared in quite a few recent work, e.g., DFHPRR-14)
- Backward direction:
  - Given a private learning algorithm
  - Construct a new one by random subsampling
  - Show it's **AERM via distribution-free assumption**.

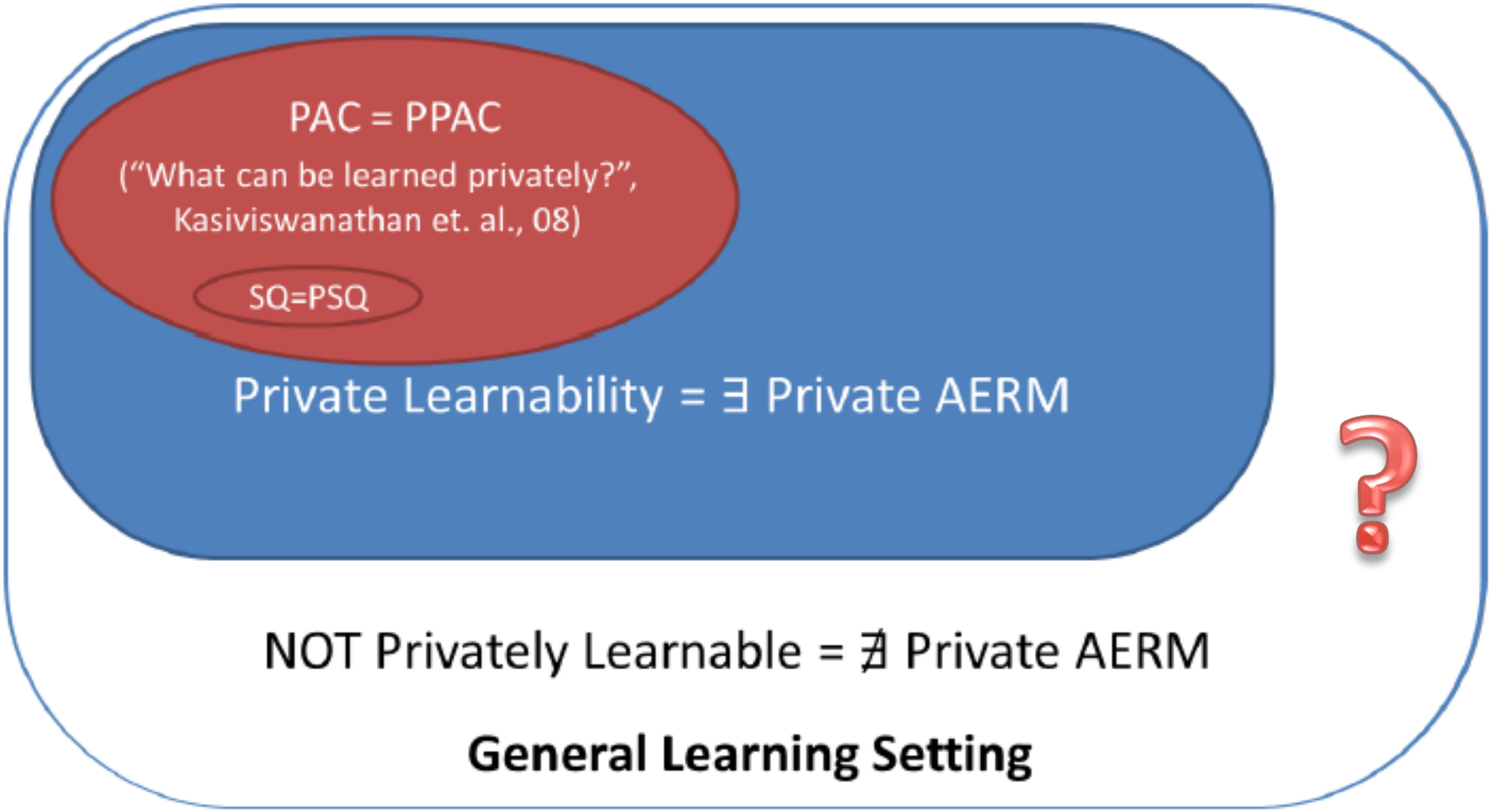
Implications

- The task reduces to finding a private ERM

- A generic procedure that produces a learning algorithm for all privately learnable problems:

$$\underset{\substack{(\mathcal{A},\epsilon):\\ \mathcal{A}:\mathcal{Z}^n\to\mathcal{H},\\ \mathcal{A}\text{ is }\epsilon\text{-DP}}}{\operatorname{argmin}} \left[\epsilon + \sup_{Z\in\mathcal{Z}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \inf_{h\in\mathcal{H}}\hat{R}(h,Z)\right)\right]$$

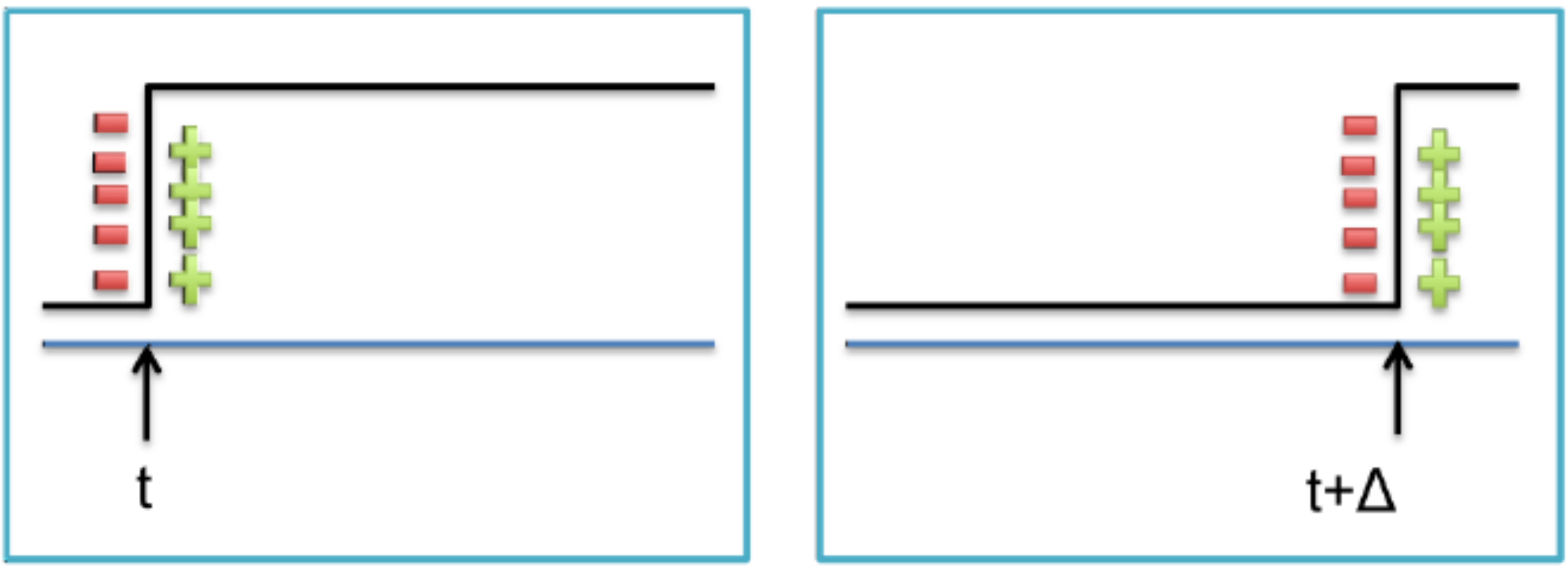
When can we learn but not privately learn?



Example in Chaudhuri and Hsu, 2011.

The difficulty of private classification in continuous domain (Chaudhuri and Hsu)

$$\frac{p(\mathcal{A}(Z))}{p(\mathcal{A}(Z'))} \leq \exp(n\epsilon).$$



Also, implicitly implied by sample complexity lower bound in [BKN13]

How to fix this?

- Lipschitz loss function, e.g., hinge loss.

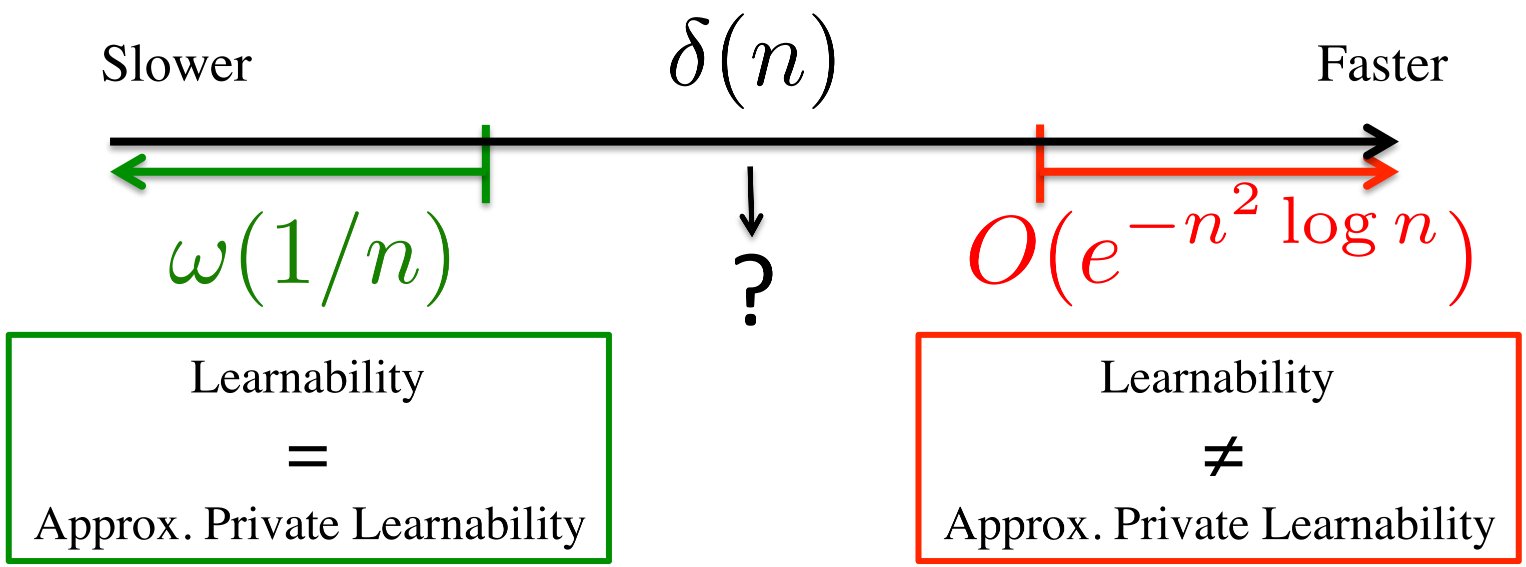
- Drop the distribution-free requirement.
  - Private $\mathcal{D}$-learnability.

(ε,δ)-private learnability

- Extend subsampling lemma and stability lemma to (ε,δ)-DP.

- Results:
  - If we require $\delta = o(1/n)$,
  - Or if we require $\delta = o(1/poly(n))$,
  - Approx. Private AERM = Approx. Private Learnability.

Are all learnable problems (ε,δ)-privately learnable?



Story so far

- In general learning setting:
  - Private ERM learns all learnable problems.
  - Many problems are not privately learnable.
  - (ε,δ)-DP does not seem to solve the problem.

- Even if a problem is privately learnable…
  - might not be practical.

Practical frustrations with DP

- Need to add too much noise/ruin inferences.
  - Resulting in poor utility.
  - E.g., Contingency Table (Fienberg et. al. 2010), GWAS data (Yu et. al., 15), etc.

- Need a lot of tricks/hacks to work
  - E.g., "clipping" "rescaling" as in the Netflix data.

- Worst-case guarantee
  - Protects the worst possible data set.
  - Sensitive to outliers.

Same randomization, many interpretations

- How small needs ε be?

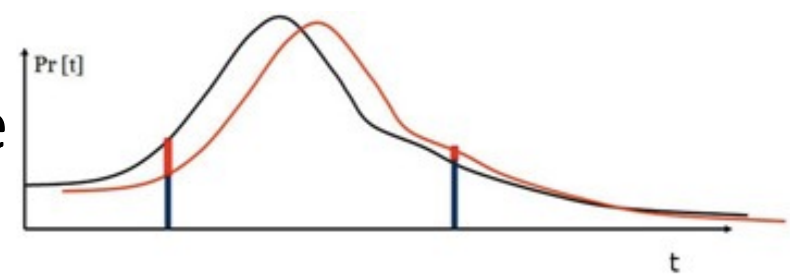
  – ε-DP = 100
  – ε-Personal DP for each person. ε< 0.2 for 95% of them.
  – On avg privacy: ε=0.1

Liu, W., and Smola. "Fast differentially private matrix factorization." RecSys'15.

Weakening the privacy definition

- "A" outputs two distributions from Z and Z'.

- Any privacy definition should require the two distributions to be close.
  - ε-DP ⇔ ε-Max-Divergence



- Use weaker distance measure?

Divergence privacy and f-divergence

- First seen in Barber & Duchi (2014).

$$D_f(P \parallel Q) \equiv \int_\Omega f\left(\frac{dP}{dQ}\right) dQ.$$

- With P,Q being A(Z), A(Z')

- When f = x log x, this becomes **KL-divergence**.

$$D_{KL}(P\|Q) = \int_\Omega \frac{dP}{dQ} \log \frac{dP}{dQ} dQ$$

On-Average KL-Privacy

- Differential Privacy:

$$\sup_{Z,Z':d(Z,Z')\leq 1} \sup_{h\in\mathcal{H}} \log \frac{p_{h\sim\mathcal{A}(Z)}(h)}{p_{h\sim\mathcal{A}(Z')}(h)} \leq \epsilon$$

- On-Average KL-Privacy:

$$\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log \frac{p_{h\sim\mathcal{A}(Z)}(h)}{p_{h\sim\mathcal{A}([Z_{-1},z])}(h)}\right] \leq \epsilon.$$

On-Average KL-Privacy

- Measures the **average privacy loss** for a particular data generating distribution.
- Unaffected by rare pathological cases.
- Adapt to easy distributions.

Properties of on-average KL-Privacy

- Inherent properties of DP
  - Small group composition
  - Adaptive Composition (caveat:
  - Closed to post-processing
- Does not need bounded loss function!
- When the loss function is bounded, the same algorithm guarantees DP.

Reusable Holdout/Adaptive Data Analysis

- A: learning algorithm output h.

A is ε-DP => A has generalization error bound ε

Definition:
Generalization error = E |Risk - Empirical Risk|.

Dwork et al. "Preserving statistical validity in adaptive data analysis." FOCS'14.

Dwork et al. "The reusable holdout: Preserving validity in adaptive data analysis." Science 349.6248 (2015): 636-638.

Dwork et al. "Generalization in adaptive data analysis and holdout reuse." NIPS (2015).

Characterizing the generalization

$$p(h) \propto \exp(-\mathcal{L}(Z))\pi(h)$$

Theorem: If A is a posterior sampling algorithm for some model:

ε-on-average KL-Privacy ⇔ ε-on-avg-generaling

For each distribution separately

Definition:

On-avg-generalization = | Risk - E Empirical risk|

W., Lei, and Fienberg (2016). "On-Average KL-Privacy and its equivalence to Generalization for Max-Entropy Mechanisms." arXiv:1605.02277.

Why posterior sampling?

- A variational justification:
  - It arises out of a maximum entropy framework.

$$\min_{\mathcal{A}} \underbrace{-H(\mathcal{A}(Z)|Z)}_{\text{An information criterion}} + \underbrace{\mathbb{E}\mathcal{L}(\mathcal{A}(Z), Z)}_{\text{The utility measure}}$$

  - The optimal solution:

$$\mathcal{A}^*(Z) \sim p(h|Z) \propto \exp(-\mathcal{L}(h, Z))$$

Tishby, Pereira & Bialek (2000). The information bottleneck method.
Mir (2012). Information-theoretic foundation of differential privacy.

Why posterior sampling?

$$\underset{\substack{(\mathcal{A},\epsilon):\\ \mathcal{A}:\mathcal{Z}^n\to\mathcal{H},\\ \mathcal{A}\text{ is }\epsilon\text{-DP}}}{\operatorname{argmin}} \left[\epsilon + \sup_{Z\in\mathcal{Z}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \inf_{h\in\mathcal{H}}\hat{R}(h,Z)\right)\right]$$

Softer privacy

Replacing with E

$$\min_{\mathcal{A}} -H(\mathcal{A}(Z)|Z) + \mathbb{E}\mathcal{L}(\mathcal{A}(Z),Z)$$

Why posterior sampling?

- A statistical justification:
  - Near optimal efficiency
  - Asymptotic normality
  - Works even under model misspecification.

Wang, Fienberg and Smola. "Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo." ICML'15.

Implication to adaptive data analysis

- Dwork et. al. 2015 : Max-Information

k-max-information => k/n-on-average KL-Privacy

For any distribution

- Russo & Zou 2015: Mutual information

$$I(\mathcal{A}(Z), Z) \leq \text{ On-Avg-Gen.} \leq \sigma\sqrt{2I(\mathcal{A}(Z), Z)}$$

For each distribution separately

The first lower bound of this form.

An example on Linear Regression

- Each user is (x,y). We have 100 of them.
- Assume (x,y) are inside [-2,2] × [-1,1]

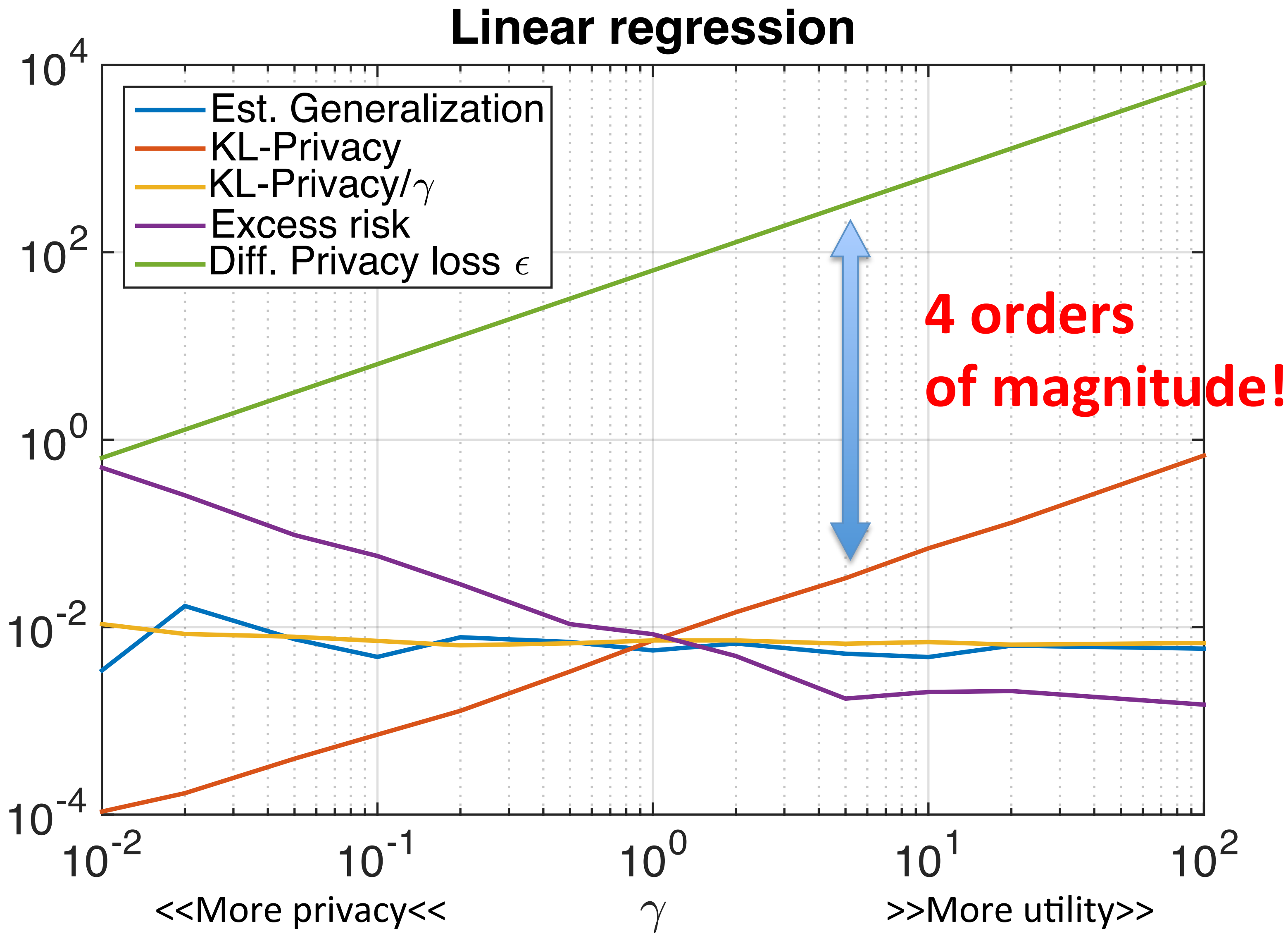
- We want to privately fit a linear regression
  - $y = x\beta_1 + \beta_0$
  - From a bounded space (β_1, β_0) in $[-2,2]^2$

- Loss function is $(y - x\beta_1 - \beta_0)^2$

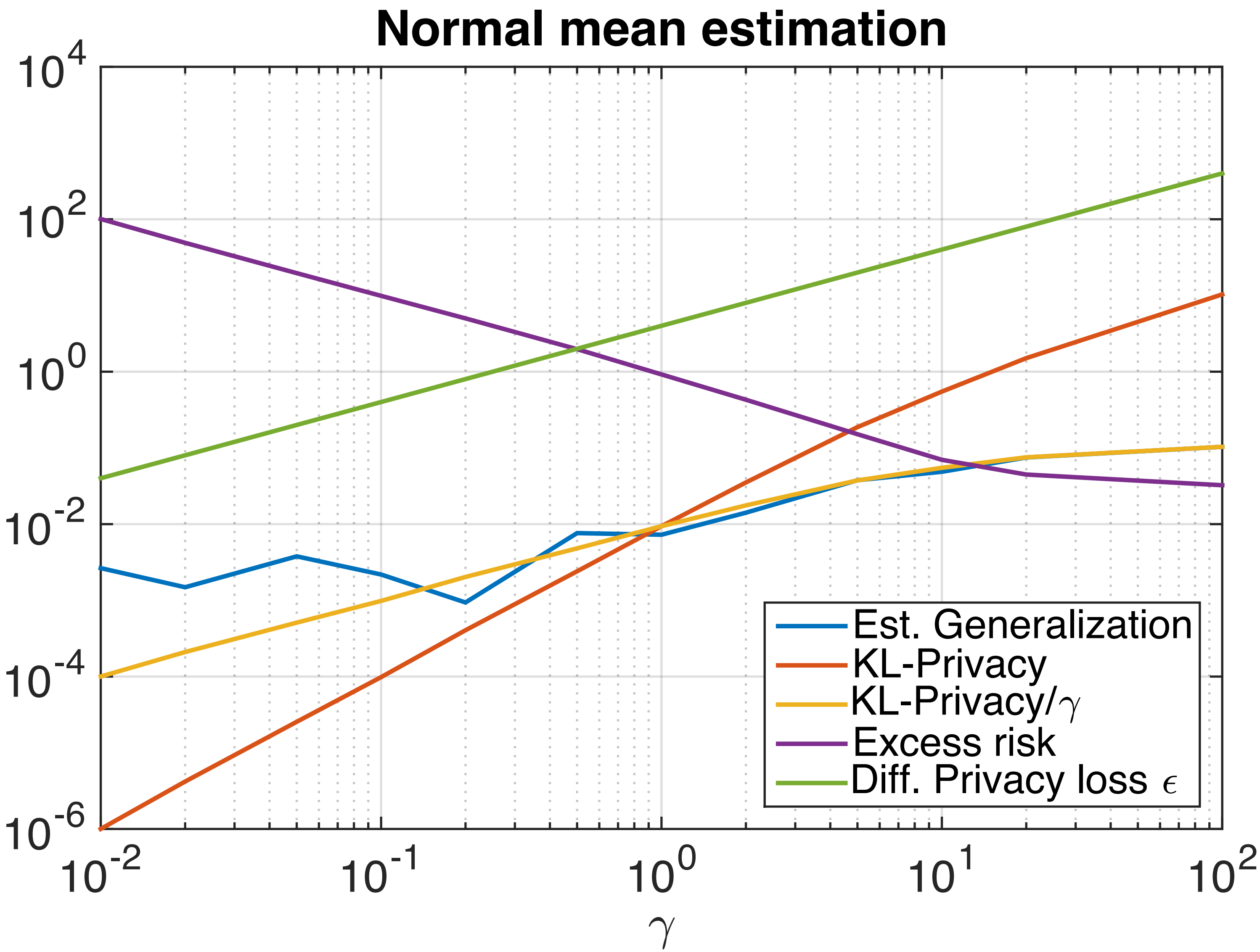
- Privately release through posterior sampling

$$(\hat{\beta}_0, \hat{\beta}_1) \sim \frac{1}{K} e^{-\gamma \sum_{i=1}^{n} (y_i - \beta_0 - x_i \beta_1)^2}$$

Experiment on Linear Regression



Experiment: normal mean



Summary of On-Avg-KL Privacy

- No changes in algorithm
  - Average case privacy guarantee.
  - Practically meaningful
  - Esp. , when ε is too large.

- Characterizing the on-average generalization
  - Lower bounds of bias in terms of mutual information.

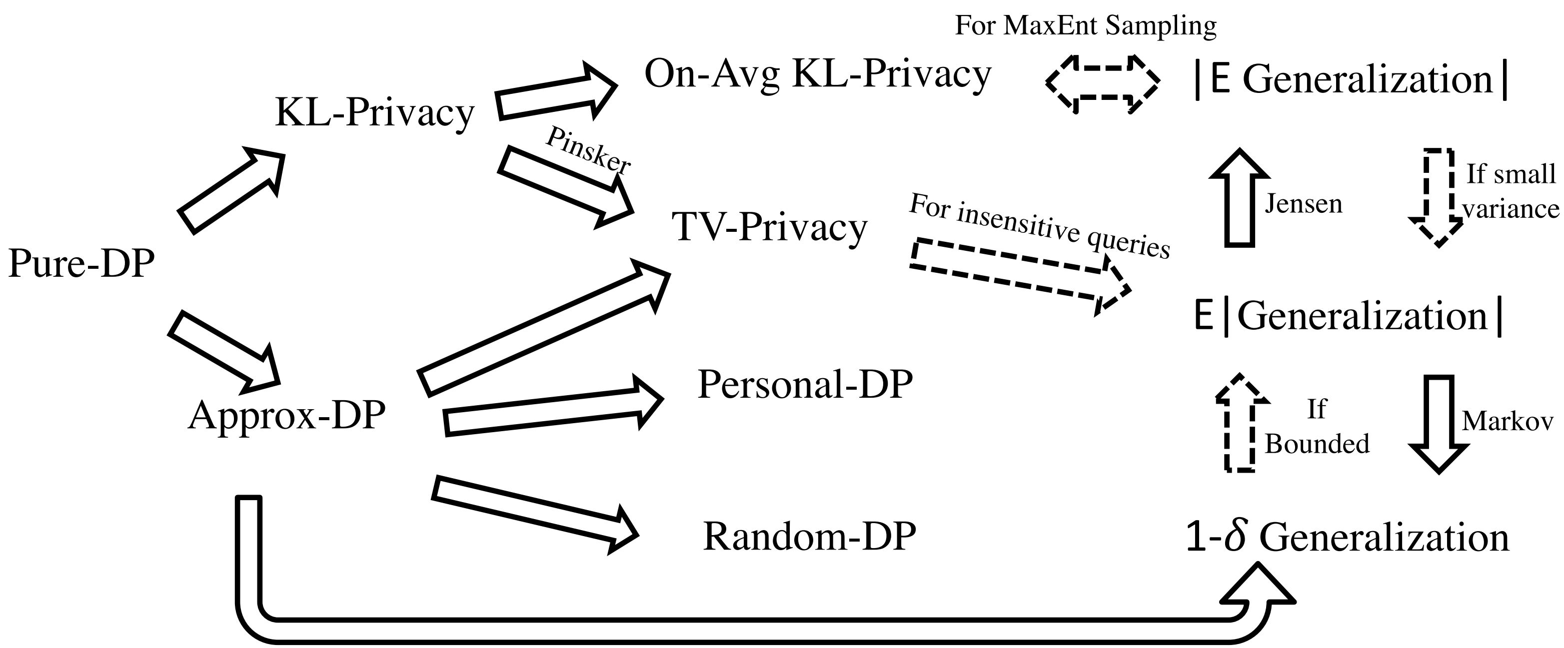
Variants of Differential Privacy

Privacy definition	Z	z	Distance (pseudo)metric
Pure DP	$\sup_{Z\in\mathcal{Z}^n}$	$\sup_{z\in\mathcal{Z}}$	$D_\infty(P\|Q)$
Approx-DP	$\sup_{Z\in\mathcal{Z}^n}$	$\sup_{z\in\mathcal{Z}}$	$D_\infty^\delta(P\|Q)$
Personal-DP	$\sup_{Z\in\mathcal{Z}^n}$	for each z	$D_\infty(P\|Q)$ or $D_\infty^\delta(P\|Q)$
KL-Privacy	$\sup_{Z\in\mathcal{Z}^n}$	$\sup_{z\in\mathcal{Z}}$	$D_{\mathrm{KL}}(P\|Q)$
TV-Privacy	$\sup_{Z\in\mathcal{Z}^n}$	$\sup_{z\in\mathcal{Z}}$	$\|P-Q\|_{TV}$
Rand-Privacy	$1-\delta_1$ any $\mathcal{D}^n$	$1-\delta_1$ any $\mathcal{D}$	$D_\infty^{\delta_2}(P\|Q)$
On-Avg KL-Privacy	$\mathbb{E}_{Z\sim\mathcal{D}^n}$ for each $\mathcal{D}$	$\mathbb{E}_{Z\sim\mathcal{D}}$ for each $\mathcal{D}$	$D_{\mathrm{KL}}(P\|Q)$

Table 1. Summary of different privacy definitions.

For references of these privacy notions, please refer to the paper:
http://arxiv.org/abs/1605.02277

Their connections to generalization



In summary

- Two recent work that investigates the connections of privacy and learning.
- Formalize the equivalence of generalization with some notion of privacy (for MaxEnt mechanisms).
- A practically useful interpretation of DP-algorithms.
- Towards practical privacy protection.