Understanding Disclosure Risk in Differential Privacy with Applications to Noise Calibration and Auditing

52nd International Conference on Very Large Data Bases

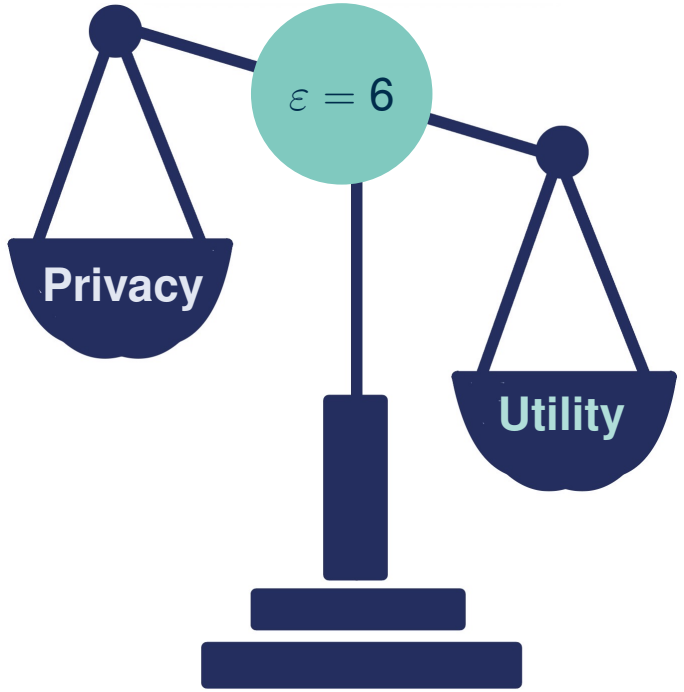
Patricia Guerra-Balboa, Annika Sauer, Héber H. Arcolezi, Thorsten Strufe

PRIVACY AND SECURITY

KASTEL

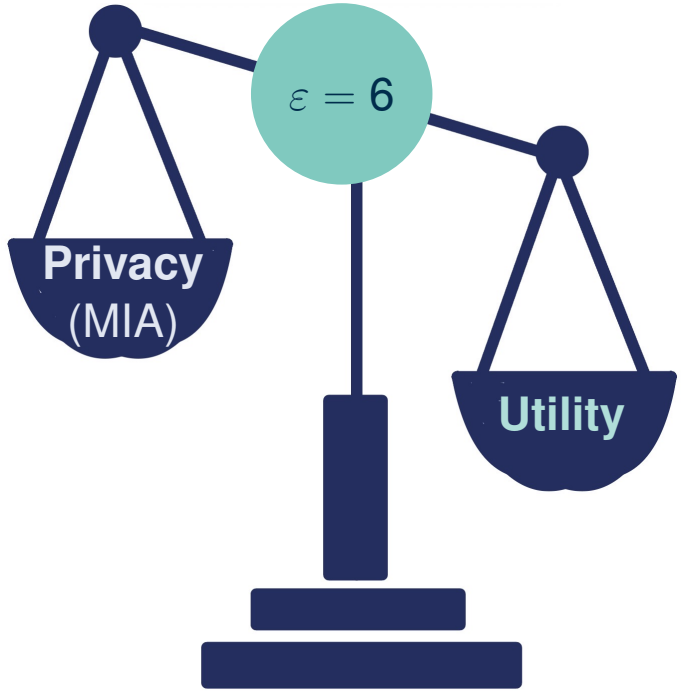
Motivation

- ε is a **key parameter** that governs the **privacy–utility trade-off** in DP.



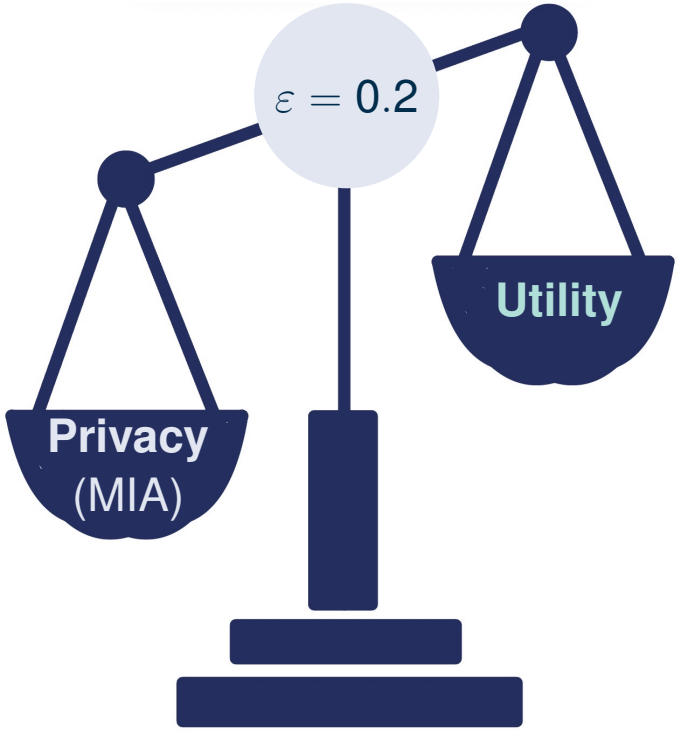
Motivation

- ε is a **key parameter** that governs the **privacy–utility trade-off** in DP.
- ✓ ε is formally connected to **Membership Inference Attacks (MIAs)**.

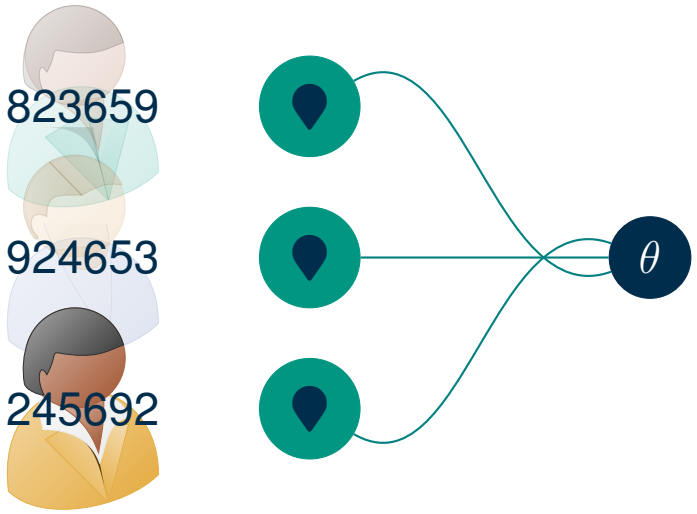


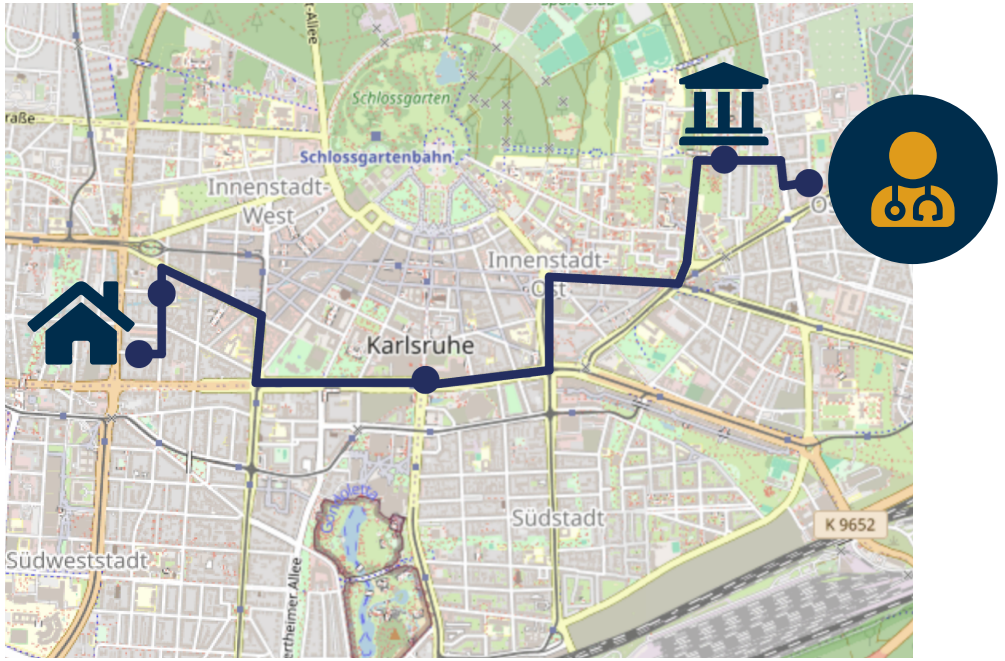
Motivation

- ε is a **key parameter** that governs the **privacy–utility trade-off** in DP.
- ✓ ε is formally connected to **Membership Inference Attacks (MIAs)**.
- ✗ Calibrating noise for MIAs when they are not a concern **unnecessarily reduces utility**.



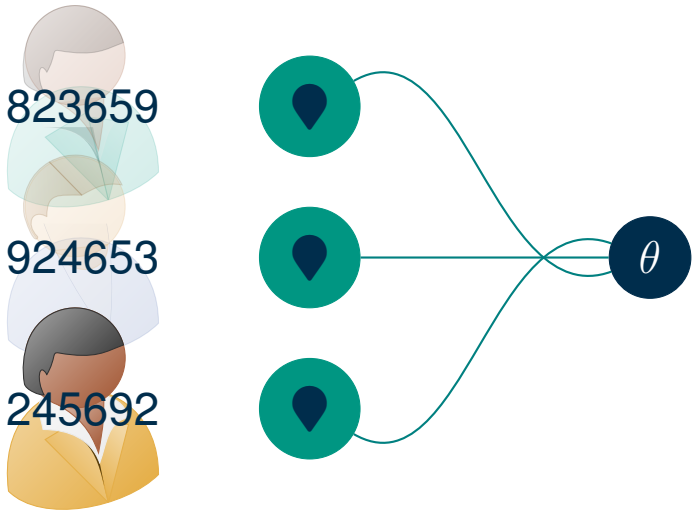
Threat Model: Reconstruction Attacks

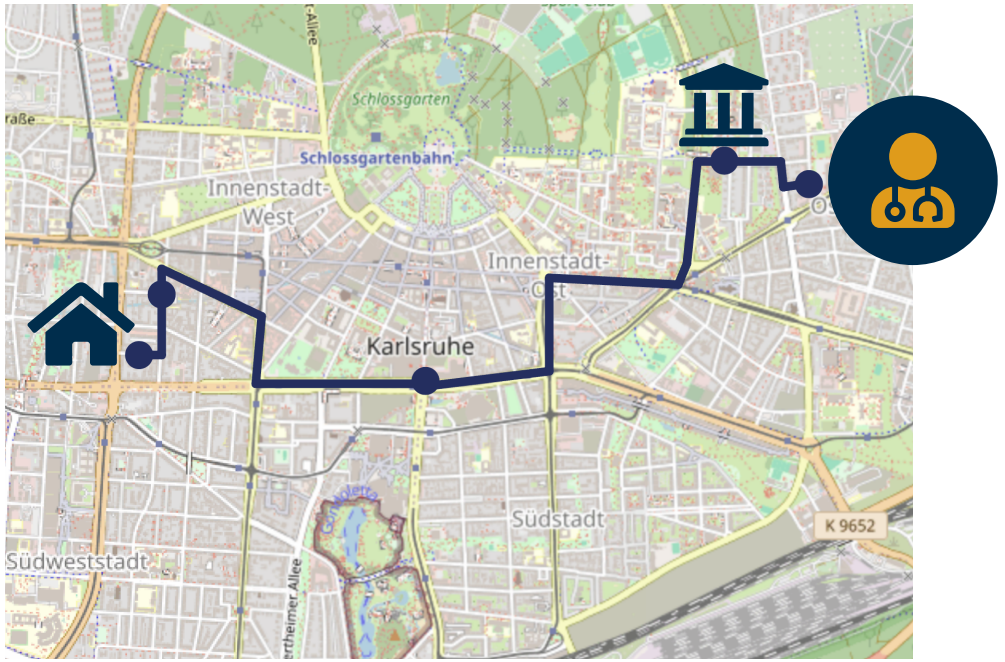




Threat Model: Reconstruction Attacks

$x \in \mathcal{X}$



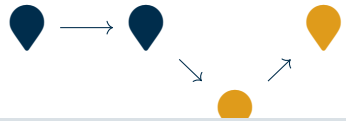


Auxiliary knowledge

- Alice's home
- Alice's job

$a(x) \in aux$

Attack's output



$A(\theta, a(x)) = \tilde{x}$

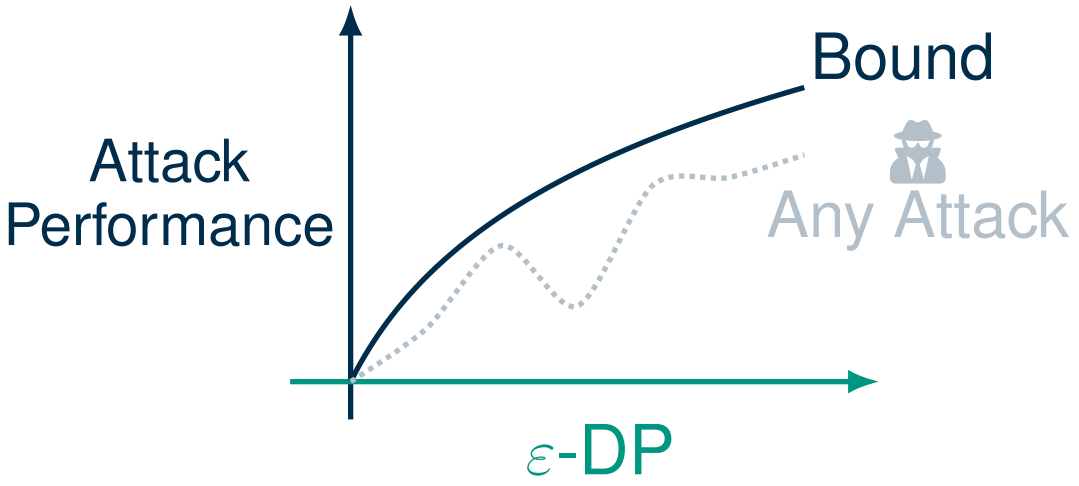
The attack is successful if the reconstruction error is less than η:

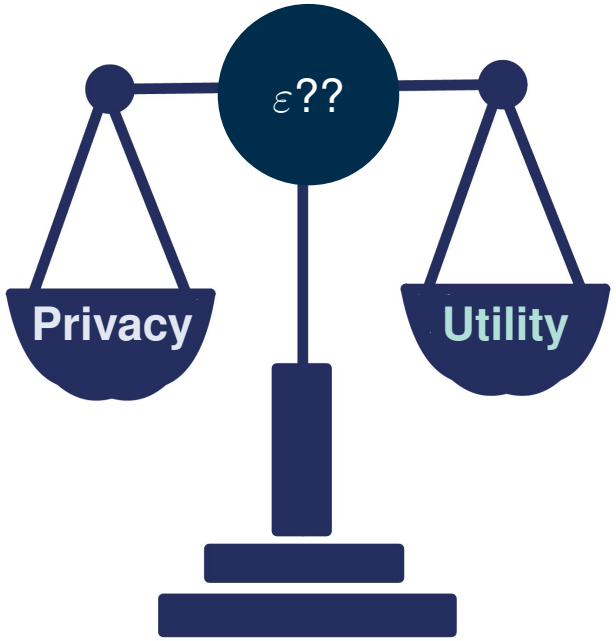
$\ell(x, \tilde{x}) \leq \eta$

Our Research Goal

Tight and accurate bounds connecting **noise injection** to **actual risk** are crucial for:

Interpretation Utility Auditing

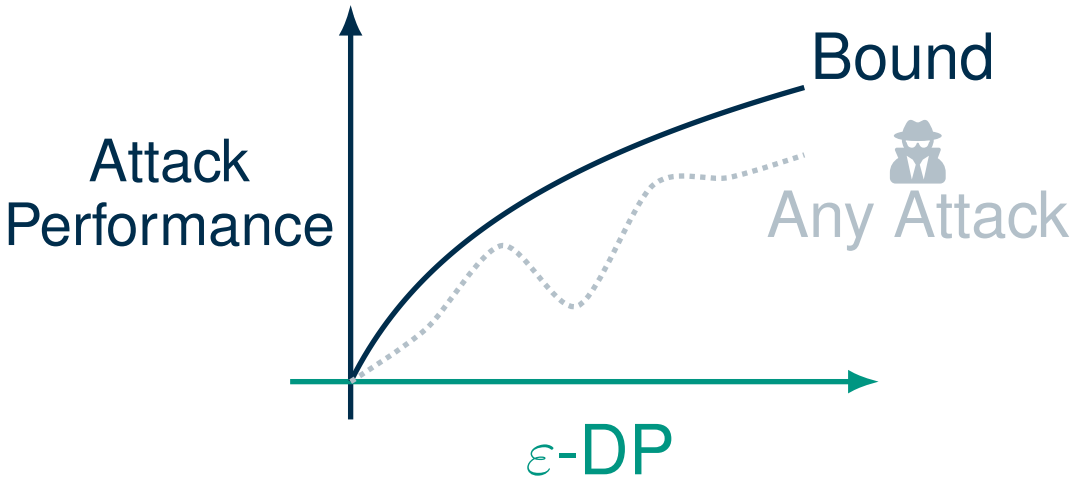


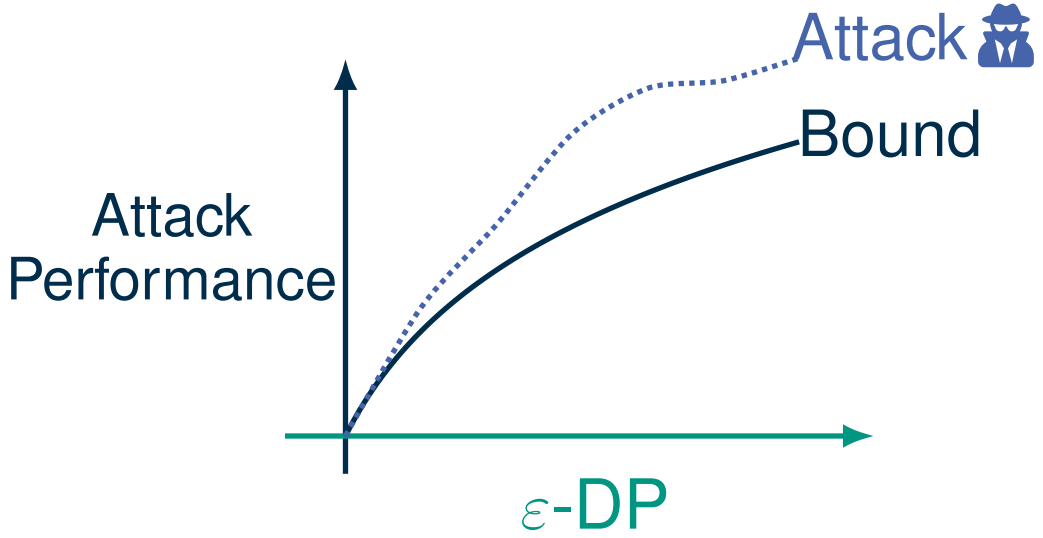


Our Research Goal

Tight and accurate bounds connecting **noise injection** to **actual risk** are crucial for:

Interpretation Utility Auditing

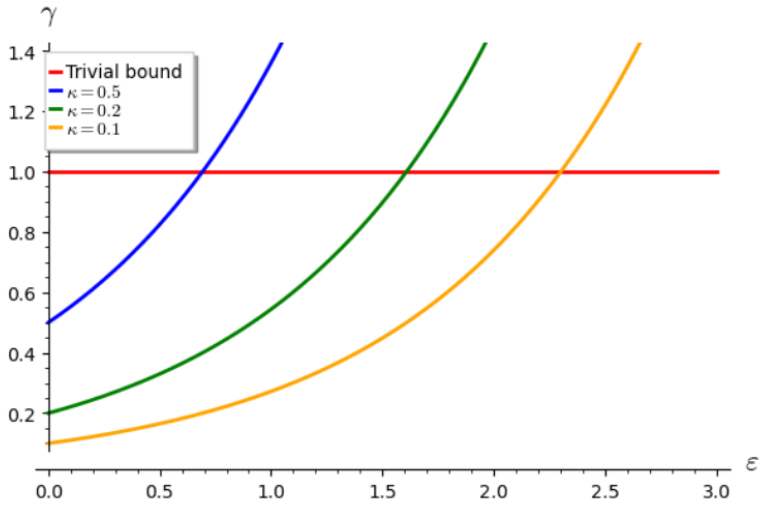




Limitations of Prior Work

η-ReRo
Balle et al.

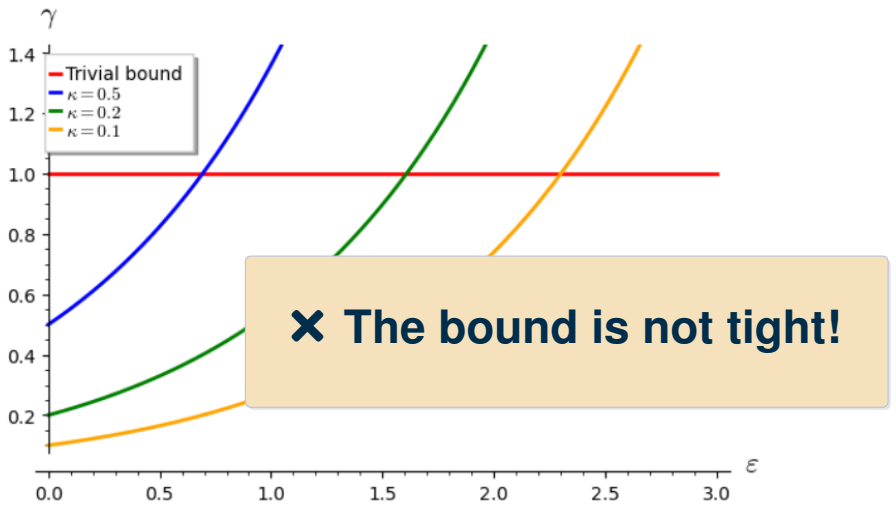
$$\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta)) \leq \eta]$$



Limitations of Prior Work

η-ReRo
Balle et al.

$$\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta)) \leq \eta]$$

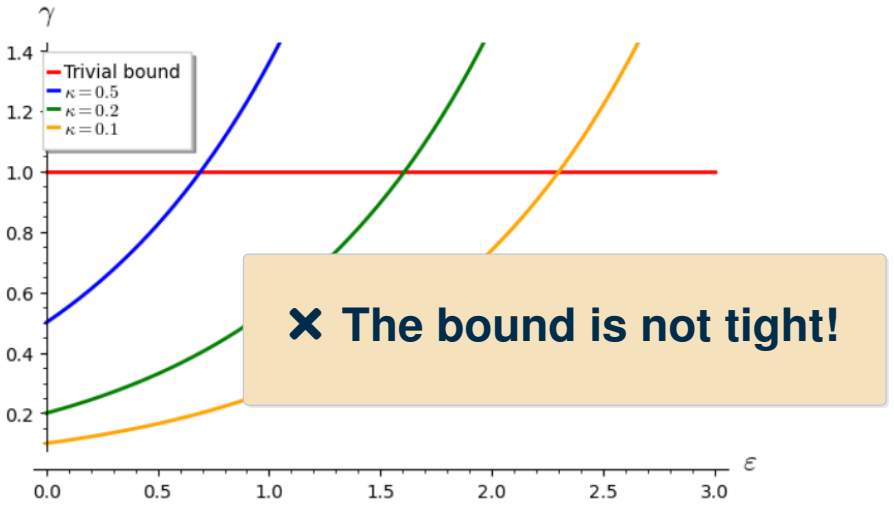


✖ **The bound is not tight!**

Limitations of Prior Work

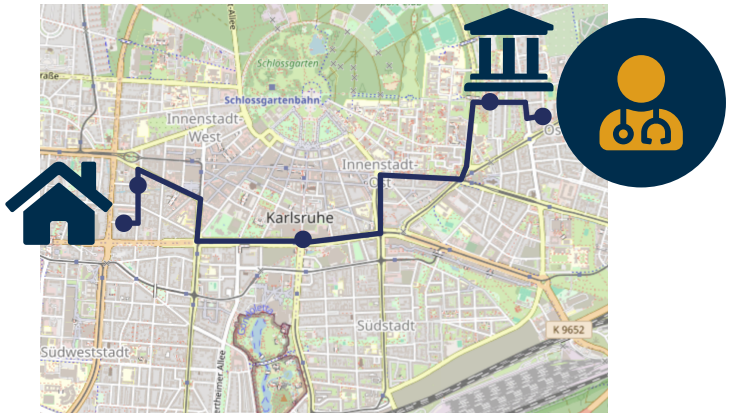
η-ReRo
Balle et al.

$$\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta)) \leq \eta]$$



✕ **The bound is not tight!**

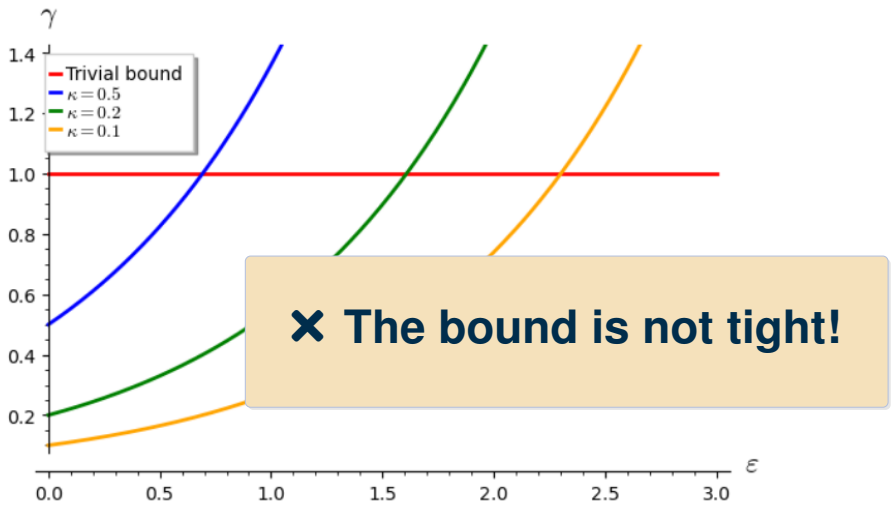
✕ **ReRo does not consider auxiliary knowledge!**



Limitations of Prior Work

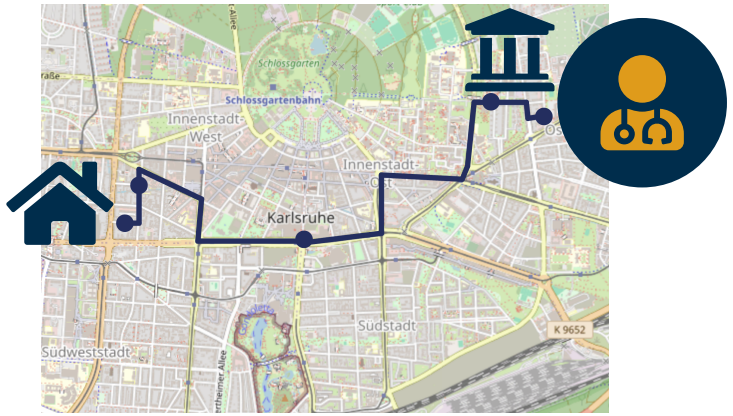
η-ReRo
Balle et al.

$$\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta)) \leq \eta]$$

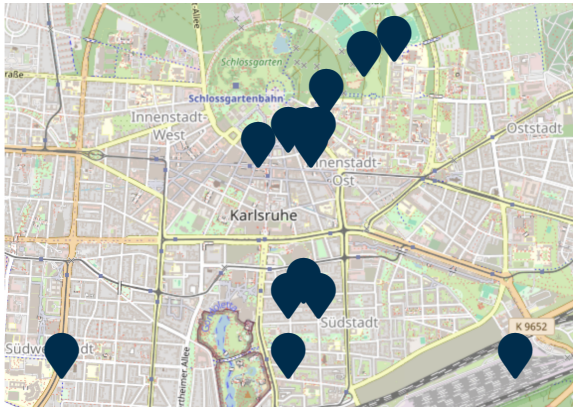


✕ **The bound is not tight!**

✕ **ReRo does not consider auxiliary knowledge!**



✕ **ReRo overestimates risk due to imputation & prior knowledge!**



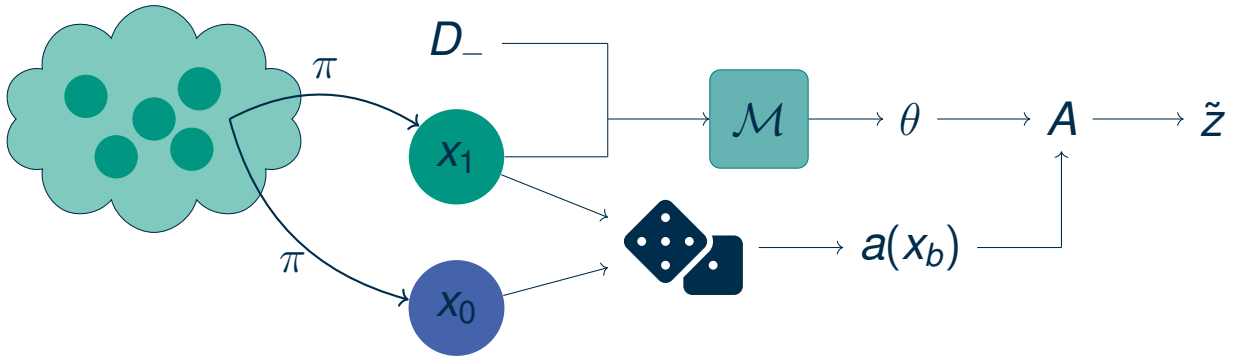
Our Contributions

Limitations of prior work:

- ✕ ReRo does not consider auxiliary knowledge.
- ✕ ReRo overestimates risk due to imputation & prior knowledge.
- ✕ The bound is not tight.

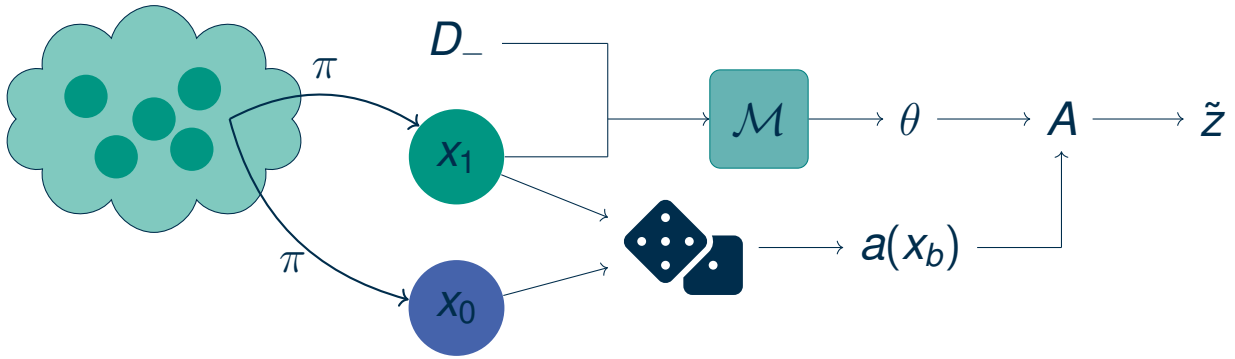
- C1. A novel attack performance metric.
- C2. Formal tight bounds connecting DP and the new metric.
- C3. A novel auditing framework using our bounds.

C1. Reconstruction Advantage (RAD)



$$\eta\text{-RAD} = \Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta] - \Pr_{\substack{X_0, X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_0})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta]$$

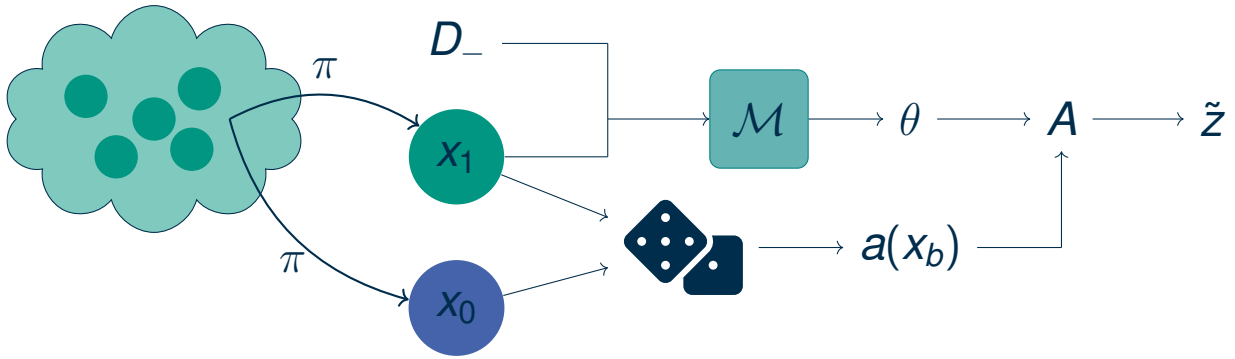
C1. Reconstruction Advantage (RAD)



$$\eta\text{-RAD} = \Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta] - \Pr_{\substack{X_0, X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_0})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta]$$

- ✔ Takes the **auxiliary knowledge** into account!

C1. Reconstruction Advantage (RAD)



$$\eta\text{-RAD} = \Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta] - \underbrace{\Pr_{\substack{X_0, X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_0})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta]}$$

- ✔ Takes the **auxiliary knowledge** into account!
- ✔ Deducts **imputation** and prior knowledge success!

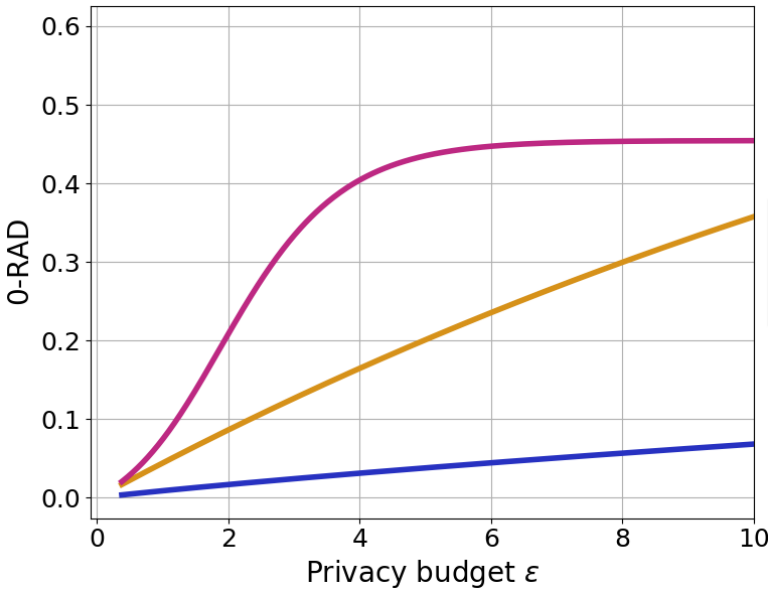
C2. Connecting DP and Reconstruction Advantage

Theorem 4.3. Universally Tight Bound (Informal)
Given a mechanism $\mathcal{M}$, a prior π, and a function $a\colon \mathcal{X} \to aux$, every attack satisfies:

$$\eta\text{-RAD} \leq \gamma(p_{\mathcal{M}}, \pi, aux)$$

— Th.4.3., $aux = \{\varnothing\}$, $\eta = 0$.

γ



Gaussian, OUE, and Laplace.

C2. Connecting DP and Reconstruction Advantage

Theorem 4.3. Universally Tight Bound (Informal) Given a mechanism $\mathcal{M}$, a prior π, and a function $a\colon \mathcal{X} \to aux$, every attack satisfies:

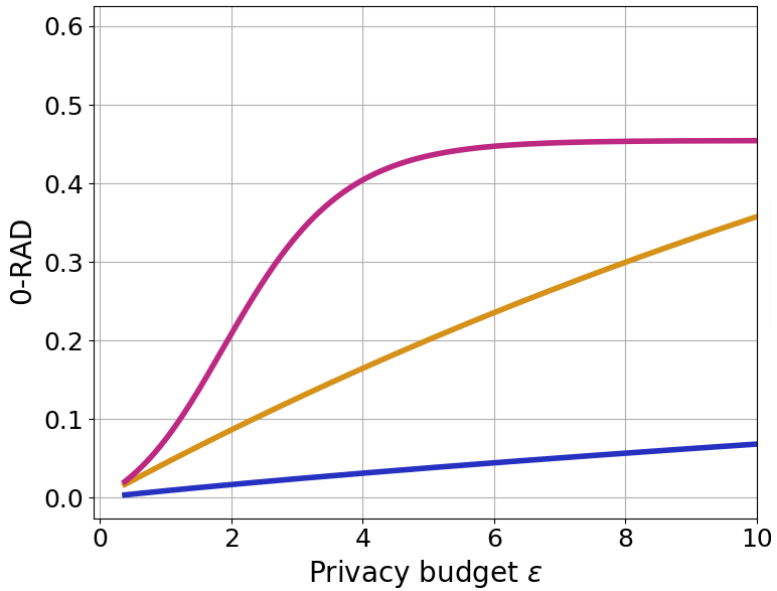
$$\eta\text{-RAD} \leq \gamma(p_{\mathcal{M}}, \pi, aux)$$

The attack is optimal

The bound is tight

We design a **novel attack** and prove that it **achieves the bound**.

— Th.4.3., $aux = \{\varnothing\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

C2. Connecting DP and Reconstruction Advantage

Theorem 4.3. Universally Tight Bound (Informal) Given a mechanism $\mathcal{M}$, a prior π, and a function $a\colon \mathcal{X} \to aux$, every attack satisfies:

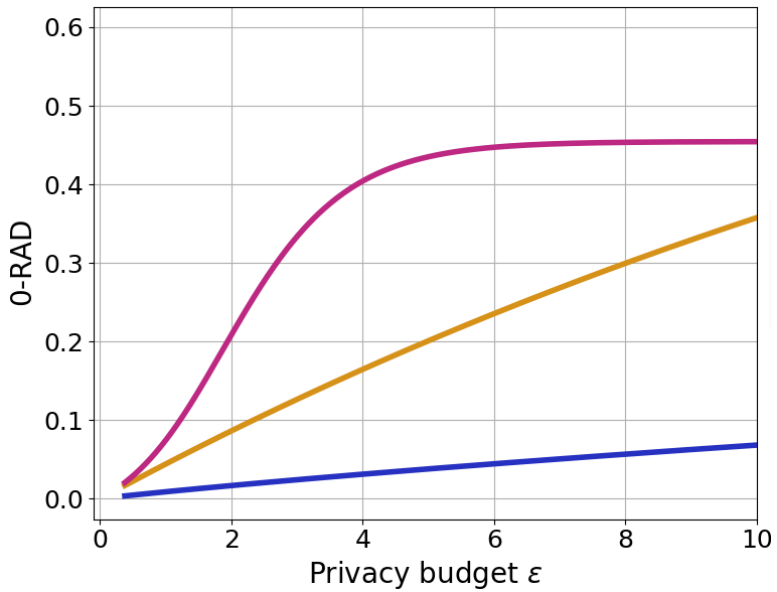
$$\eta\text{-RAD} \leq \gamma(p_{\mathcal{M}}, \pi, aux)$$

The attack is optimal

The bound is tight

✔ We design a **novel attack** and prove that it **achieves the bound**.

— Th.4.3., $aux = \{\varnothing\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

Limitations:

$$\gamma(p_{\mathcal{M}}, \pi, aux) = \sum_{\theta \in \Theta} \sum_{z \in aux} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x)\pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

✘ The bound (γ) does not always have a **closed form.**

C2. Connecting DP and Reconstruction Advantage

Theorem 4.3. Universally Tight Bound (Informal) Given a mechanism $\mathcal{M}$, a prior π, and a function $a\colon \mathcal{X} \to aux$, every attack satisfies:

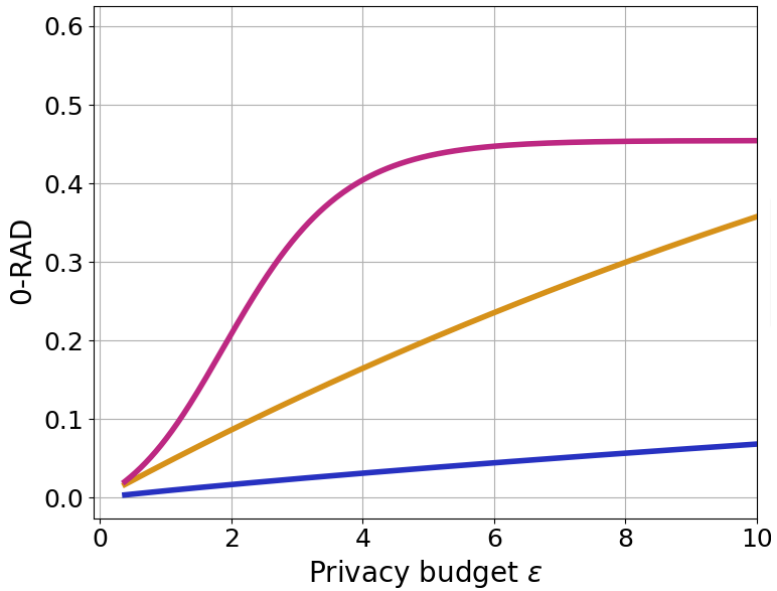
$$\eta\text{-RAD} \leq \gamma(p_{\mathcal{M}}, \pi, aux)$$

The attack is optimal

The bound is tight

We design a **novel attack** and prove that it **achieves the bound**.

— Th.4.3., $aux = \{\varnothing\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

Limitations:

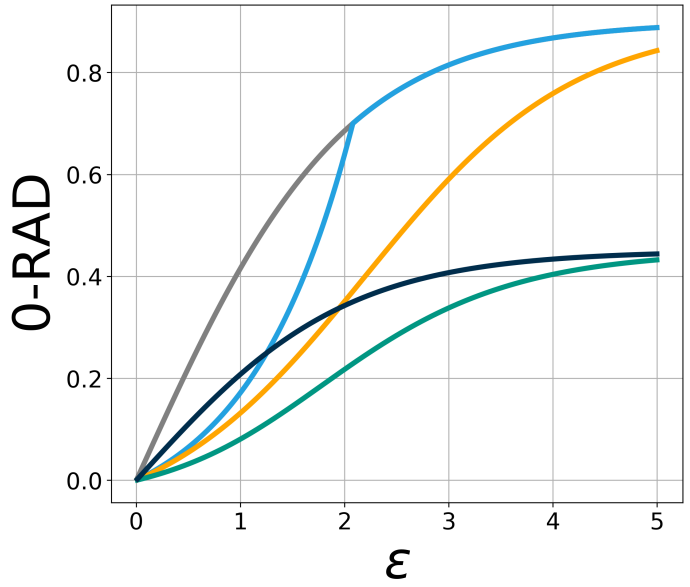
$$\gamma(p_{\mathcal{M}}, \pi, aux) = \sum_{\theta\in\Theta} \sum_{z\in aux} \max_{x_\theta\in\mathcal{X}} \sum_{\substack{\ell(x,x_\theta)\leq\eta \\ a(x)=z}} w(\theta, x)\pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

- ✗ The bound (γ) does not always have a **closed form.**
- ✗ We need access to the **auxiliary knowledge**.

C2. Upper Bounds on Reconstruction Advantage

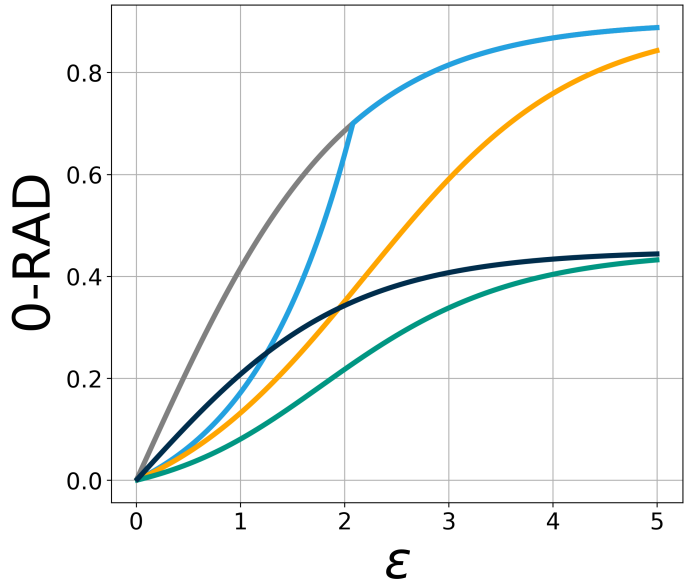
- ✔ We provide **closed-form upper-bounds** to apply when our optimal bound is not computable.



Notion	Assumptions	RAD
TV	—	Th.4.2.
f-DP	$aux = \{\varnothing\}$	Th.5.1.
$\mathcal{M}$	aux known	Th.4.3.
(ε, δ)-DP	—	Th.4.2.
(ε, δ)-DP	$aux = \{\varnothing\}$	Prop.5.2.
(ε, δ)-DP	$aux = \{\varnothing\}, \eta = 0$	Th.5.3.

C2. Upper Bounds on Reconstruction Advantage

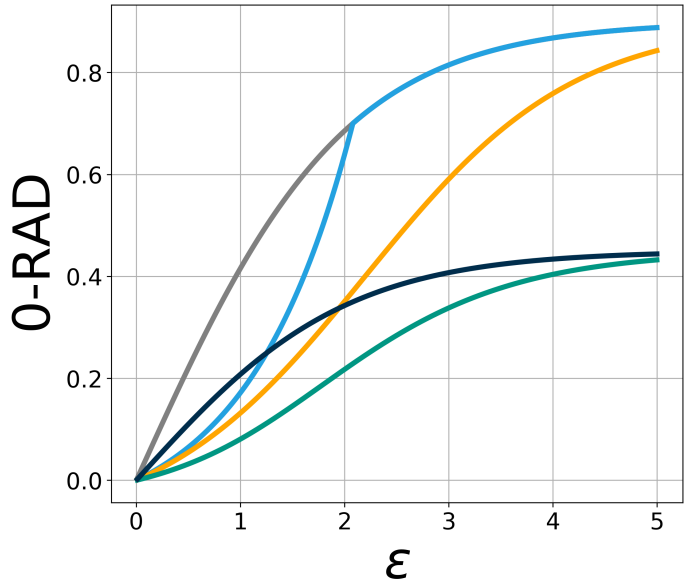
- ✔ We provide **closed-form upper-bounds** to apply when our optimal bound is not computable.



Notion	Assumptions	RAD
TV	—	Th.4.2.
f-DP	$aux = \{\varnothing\}$	Th.5.1.
$\mathcal{M}$	aux known	Th.4.3.
(ε, δ)-DP	—	Th.4.2.
(ε, δ)-DP	$aux = \{\varnothing\}$	Prop.5.2.
(ε, δ)-DP	$aux = \{\varnothing\}, \eta = 0$	Th.5.3.

C2. Upper Bounds on Reconstruction Advantage

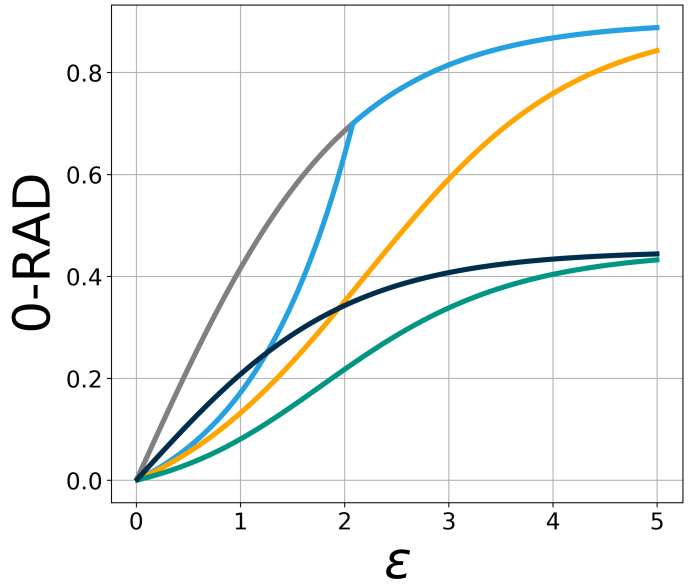
- ✔ We provide **closed-form upper-bounds** to apply when our optimal bound is not computable.



Notion	Assumptions	RAD
TV	—	Th.4.2.
f-DP	$aux = \{\varnothing\}$	Th.5.1.
$\mathcal{M}$	aux known	Th.4.3.
(ε, δ)-DP	—	Th.4.2.
(ε, δ)-DP	$aux = \{\varnothing\}$	Prop.5.2.
(ε, δ)-DP	$aux = \{\varnothing\}, \eta = 0$	Th.5.3.

C2. Upper Bounds on Reconstruction Advantage

- ✔ We provide **closed-form upper-bounds** to apply when our optimal bound is not computable.



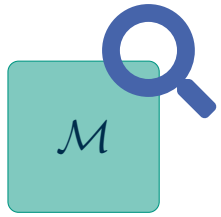
Notion	Assumptions	RAD
TV	**—**	**Th.4.2.**
***f*-DP**	$aux = \{\varnothing\}$	**Th.5.1.**
$\mathcal{M}$	**aux known**	**Th.4.3.**
(ε, δ)**-DP**	**—**	**Th.4.2.**
(ε, δ)**-DP**	$aux = \{\varnothing\}$	**Prop.5.2.**
(ε, δ)**-DP**	$aux = \{\varnothing\}, \eta = 0$	**Th.5.3.**

Our experiments in

Machine Learning Aggregated Statistics Location Data

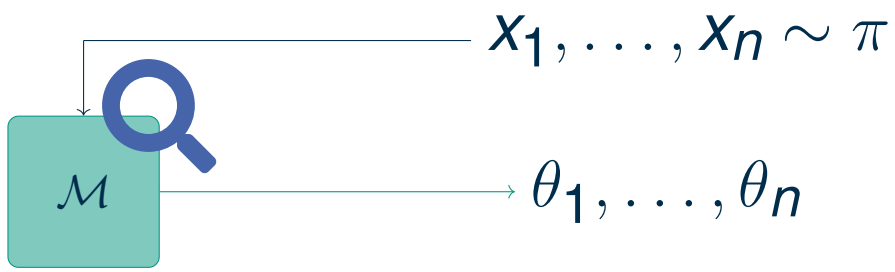
show the applicability and tightness of our results.

C3. RAD-based DP Auditing



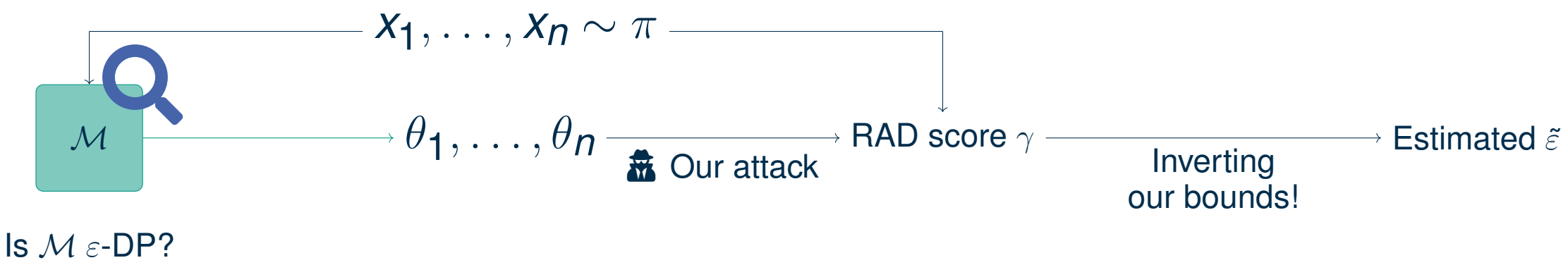
Is $\mathcal{M}$ ε-DP?

C3. RAD-based DP Auditing

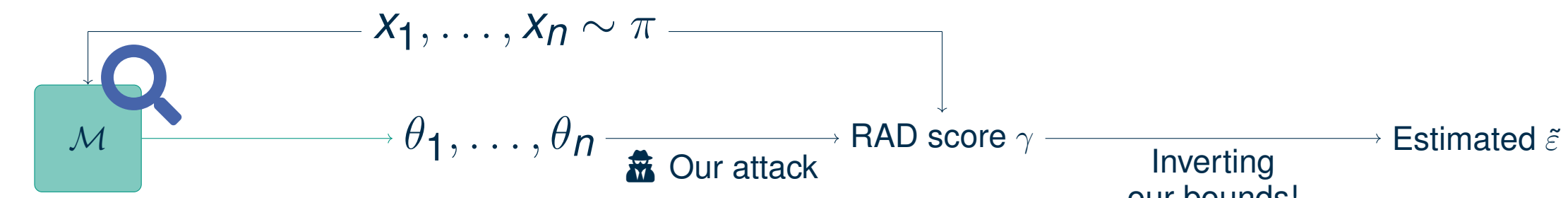


Is $\mathcal{M}$ ε-DP?

C3. RAD-based DP Auditing



C3. RAD-based DP Auditing



Is $\mathcal{M}$ ε-DP?

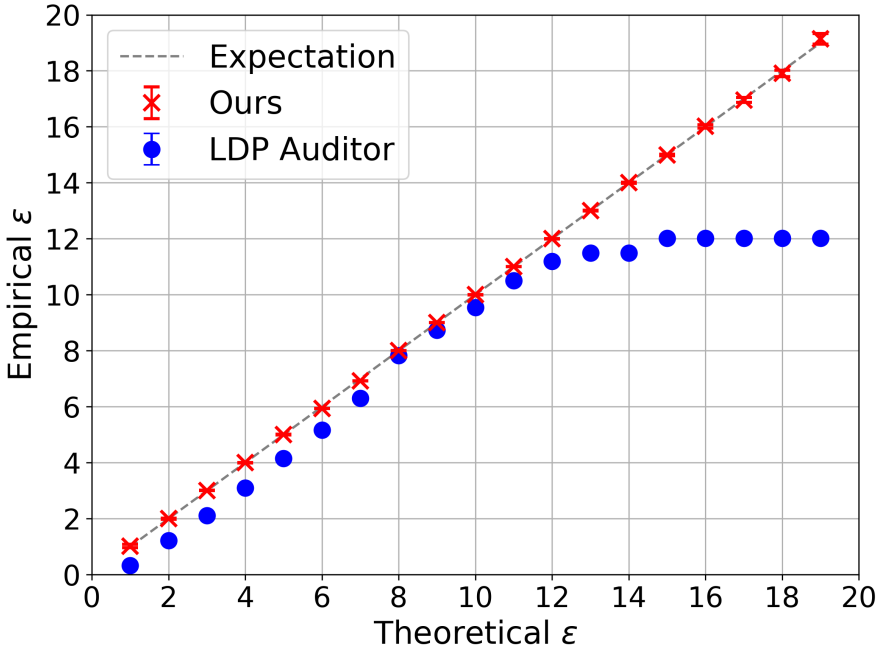
We improve scalability over previous auditing methods!

C3. RAD-based DP Auditing

Is $\mathcal{M}$ ε-DP?

We improve scalability over previous auditing methods!

We improve accuracy over LDP Auditor!



Conclusions

ε is not enough!

RAD and our bounds reveal the actual participation risk

- ✔ We **avoid overestimating risk** and **improve utility** over prior approaches.
- ✔ We can c**alibrate noise** to **exactly match the risk** we are willing to accept.

Conclusions

ε is not enough!

RAD and our bounds reveal the actual participation risk

- ✔ We **avoid overestimating risk** and **improve utility** over prior approaches.
- ✔ We can c**alibrate noise** to **exactly match the risk** we are willing to accept.

Our RAD bounds and attack enable auditing

- ✔ We **detect errors and misconduct more accurately**.
- ✔ Our method **scales to large databases** and high-dimensional categorical data.

Conclusions

ε is not enough!

RAD and our bounds reveal the actual participation risk

- ✔ We **avoid overestimating risk** and **improve utility** over prior approaches.
- ✔ We can c**alibrate noise** to **exactly match the risk** we are willing to accept.

Our RAD bounds and attack enable auditing

- ✔ We **detect errors and misconduct more accurately**.
- ✔ Our method **scales to large databases** and high-dimensional categorical data.

Paper

Code

Conclusions

ε is not enough!

RAD and our bounds reveal the actual participation risk

- ✔ We **avoid overestimating risk** and **improve utility** over prior approaches.
- ✔ We can c**alibrate noise** to **exactly match the risk** we are willing to accept.

Our RAD bounds and attack enable auditing

- ✔ We **detect errors and misconduct more accurately**.
- ✔ Our method **scales to large databases** and high-dimensional categorical data.

More results and details in our poster!

Paper

Code

Additional Slides

Theorem 4.2

Theorem ((ε, δ)-DP implies η-RAD)

Let $\pi, \ell, \eta \geq 0$ as in Def. 5.1., and $\kappa_\pi = \Pr_{X,X' \sim \pi}[X = X']$. If a mechanism $\mathcal{M} \colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ satisfies (ε, δ)-DP, then for any attack $A \colon \Theta \times aux \to \mathcal{D}(\mathcal{X})$, and database D_- we have

$$\eta\text{-RAD} \leq \mathrm{TV}(\mathcal{M})(1 - \kappa_\pi) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}(1 - \kappa_\pi).$$

Theorem 4.3

Theorem

Given $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ and $a\colon \mathcal{X} \to$ aux measurable, then for any attack $A\colon \Theta \times \text{aux} \to \mathcal{D}(\mathcal{X})$, we have

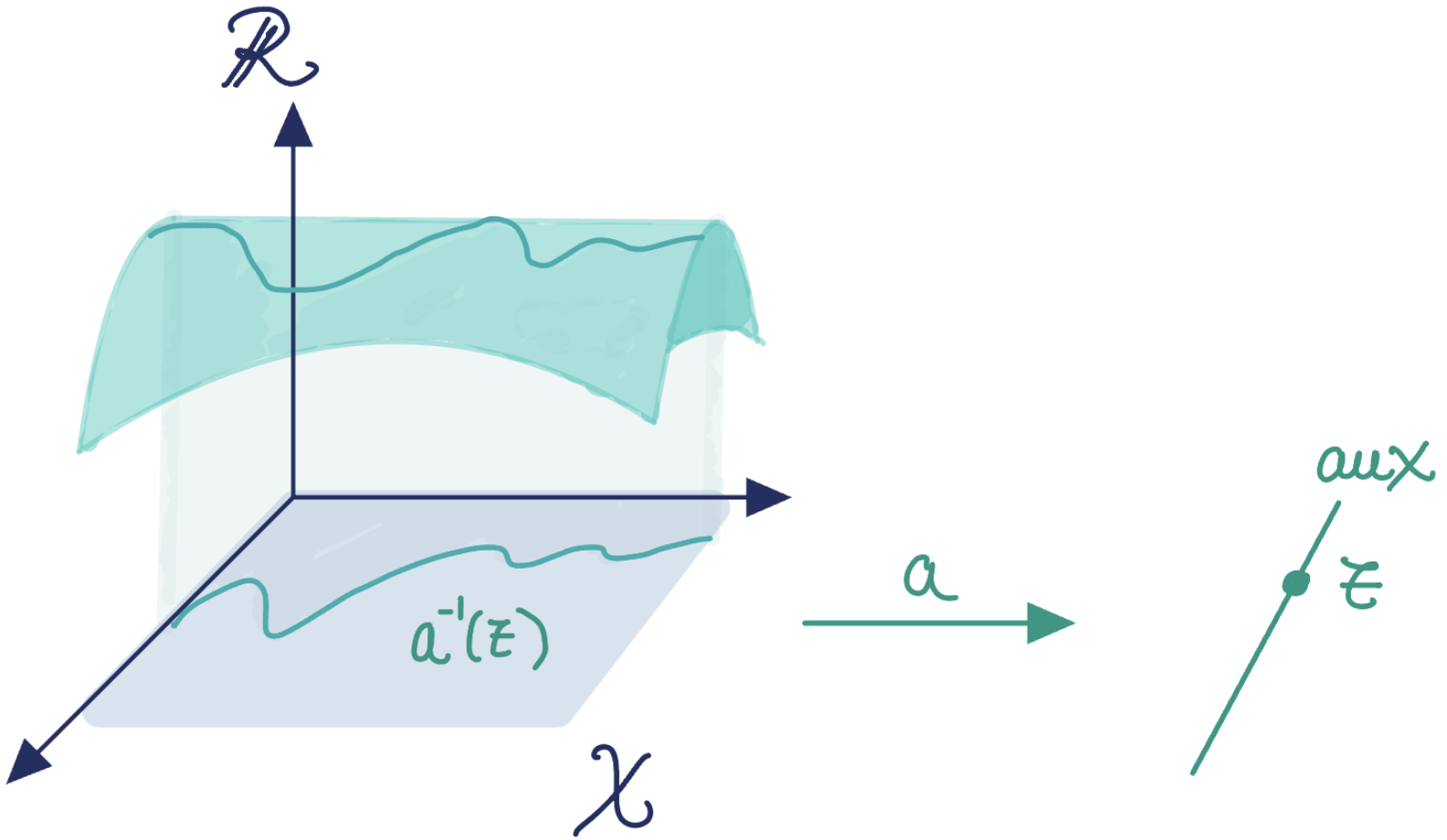
$$\eta\text{-RAD} \leq \int_{\Theta} \int_{\text{aux}} \max_{x_\theta \in \mathcal{X}} \left(\int_{S_\eta^z(x_\theta)} w(\theta, x) \pi_x \, \mathrm{d}\mu_z(x) \right) \mathrm{d}\nu(z) \, \mathrm{d}\mu(\theta),$$

where μ the counting (or Lebesgue measure) and π_x mass (or density function) in the discrete (or continuous case). Additionally, $w(x, \theta) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$, $S_\eta^z(x_\theta) = \{x\colon a(x) = z \wedge \ell(x_\theta, x) \leq \eta\}$, $\nu(z) = \mu \circ a^{-1}(z)$ and μ_z the disintegration theorem measure.
The discrete case simplifies to

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x$$

by direct application of Remark 2.29.

Disintegration Theorem



Optimal Bound for Particular Cases

Membership Inference Attack (MIA):

- $a(x) = x$
- $\ell(x, x') = 1$ if $x = x'$ and 0 otherwise.

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{\substack{x \in \mathcal{X} \\ w(\theta,x)>0}} w(\theta, x)\pi_x$$

Perfect reconstruction:

- $a(x) = \varnothing$ for every $x \in \mathcal{X}$
- $\eta = 0$

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in aux} \max_{\substack{x \in \mathcal{X} \\ w(\theta,x)>0}} w(\theta, x)\pi_x$$

Generalized Randomized Response

$$\eta\text{-RAD} \leq \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + m - 1}(1 - \kappa_\pi)$$

Laplace mechanism, uniform prior

$$\eta\text{-RAD} \leq \left(1 - e^{-\frac{\varepsilon}{2(m-1)}}\right)\frac{m-1}{m}$$

Optimal Unary Encoding

$$\eta\text{-RAD} \leq \frac{1}{2}\frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1}(1 - \kappa_\pi)$$

Gaussian mechanism, uniform prior

$$\eta\text{-RAD} \leq \left(2\Phi\left(\frac{1}{2\sigma(m-1)}\right)\right)\frac{m-1}{m}$$

Upper-bounds for $aux = \{\varnothing\}$

Theorem

If a mechanism $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ satisfies f-DP, then for any attack with $aux = \{\varnothing\}$, $A\colon \Theta \to \mathcal{D}(\mathcal{X})$, it satisfies

$$\eta\text{-RAD} \leq \max_{\alpha \in [\kappa^-_{\pi,\ell}(\eta), \kappa^+_{\pi,\ell}(\eta)]} 1 - f(\alpha) - \alpha.$$

If $\mathcal{X}$ is discrete, then it also holds

$$\eta\text{-RAD} \leq (1 - \kappa_\pi) \max_{\alpha \in [0, \frac{\kappa^+_{\pi,\ell}(\eta)}{1-\kappa_\pi}]} 1 - f(\alpha) - \alpha.$$

Theorem (0-RAD under (ε, δ)-DP)

Given $|\mathcal{X}| = m$ with prior $\pi_1(1-\pi_1) \geq \cdots \geq \cdots \geq \pi_m(1-\pi_m)$ and $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ an (ε, δ)-DP mechanism, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{X})$

$$0\text{-RAD} \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1} K_\pi + R \max_{i>K} \pi_i$$

where $K \in [m]$ is the largest index satisfying $R = (m-1)\frac{e^\varepsilon - 1 + m\delta}{e^\varepsilon + m - 1} - (K - \sum_{i=1}^K \pi_i)\frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1} \geq 0$ and $K_\pi = \sum_i^K (1-\pi_i)\pi_i$.

Optimal Attack against DP-SGD

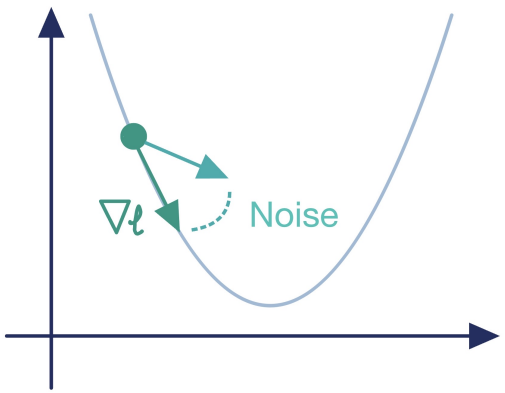
- $\mathcal{X} = \{x_1, \ldots, x_m\}$, π uniform.
- Intermediate weights $\theta = (\theta_1, \ldots, \theta_T)$ and
- privatized gradients $\{\bar{g}_1, \ldots, \bar{g}_T\}$ released during training:

$$\bar{g}_t = \sum_x \text{clip}_C(\nabla_\theta \ell(\theta_t, x)) + \mathcal{N}(0, c^2\sigma^2 I),$$

where $\text{clip}_C(\vec{v}) = \vec{v}\min(1, \frac{C}{\|\vec{v}\|_2})$.

- Noisy contribution of the target's gradient t:

$$g_t = \bar{g}_t - \sum_{x \in D_-} \text{clip}_C(\nabla_\theta \ell(\theta_t, x)).$$



Algorithm 1: Optimal Attack for DP-SGD

Input : $\theta = (\theta_1, \ldots, \theta_T)$, $a(x) = z$ and $g = (g_1, \ldots, g_T)$
Output : $\tilde{x}$
for $x : a(x) = z$ **do**
  compute $\sum_t W(\bar{g}_t, \text{clip}_C(\nabla_\theta \ell(\theta_t, x)))$;
Select $x^* = \arg\max_{x:\, a(x)=z} \sum_t W(\bar{g}_t, \text{clip}_C(\nabla_\theta \ell(\theta_t, x)))\pi_x$;
if $W(\bar{g}_t, x^*) > 0$ **then**
  $\tilde{x} = x^*$;
else
  $\tilde{x} \longleftarrow U[\mathcal{X} \setminus \{x : a(x) = z\}]$;

We can simplify w maximization to

$$\arg\max_{x:a(x)=z} w(\theta, x) = \arg\max_{x:a(x)=z} \sum_t W(g_t, \text{clip}_C(\nabla_\theta \ell(\theta_t, x)))$$

where $W(u, v) = \langle u, v \rangle - \frac{1}{m} \sum_{x_i \in \mathcal{X}} \langle u, \text{clip}_C(\nabla_\theta \ell(\theta_t, x_i)) \rangle$.

Optimal Attack Against DP-SGD

Why can we simplify w to W?

Given θ, x, under DP-SGD the privatized gradient at step t is

$$g_t \sim \mathcal{N}\left(\mu_x,\ C^2\sigma^2 I\right), \qquad \mu_x = \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x)),$$

where C is the clipping parameter and I the identity function of dimension d. Hence

$$p_{\mathcal{M}}(g_t \mid x) = \underbrace{\frac{1}{(2\pi C^2\sigma^2)^{d/2}}}_{A} \exp\Big(\underbrace{-\frac{1}{2C^2\sigma^2}}_{B} \|g_t - \mu_x\|^2\Big),$$

where both A, B are independent from x. Consequently,

$$w(g,x) = \prod_t p_{\mathcal{M}}(g_t \mid x) - \prod_t p_{\mathcal{M}}(g_t) > 0 \Leftrightarrow A^T \left(\prod_t e^{B\langle g_t, \mu_x\rangle} - \prod_t \frac{1}{m}\sum_i e^{B\langle g_t, \mu_{x_i}\rangle}\right) > 0 \Leftrightarrow \tag{1}$$

$$e^{B\sum_t \langle g_t, \mu_x\rangle} > \prod_t \frac{1}{m}\sum_i e^{B\langle g_t, \mu_{x_i}\rangle} \Leftrightarrow B\sum_t \langle g_t, \mu_x\rangle > \sum_t \ln\left(\frac{1}{m}\sum_i e^{B\langle g_t, \mu_{x_i}\rangle}\right) \Leftrightarrow \tag{2}$$

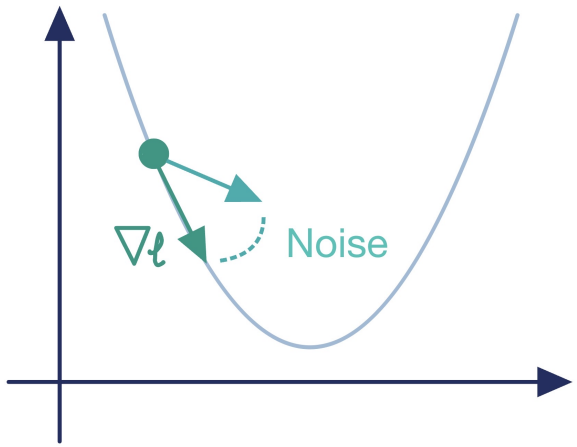
$$B\sum_t \langle g_t, \mu_x\rangle > \sum_t \frac{1}{m}\sum_z \ln(e^{B\langle g_t, \mu_{x_i}\rangle}) \Leftrightarrow B\sum_t \langle g_t, \mu_x\rangle > \sum_t \frac{B}{m}\sum_i \langle g_t, \mu_{x_i}\rangle \Leftrightarrow \tag{3}$$

$$\sum_t \langle g_t, \mu_x\rangle - \sum_t \frac{1}{m}\sum_{x_i} \langle g_t, \mu_{x_i}\rangle > 0 \Leftrightarrow \sum_t W(g_t, x) > 0. \tag{4}$$

Experimental Details: Mechanisms

ML Image Classification

DP-(S)GD



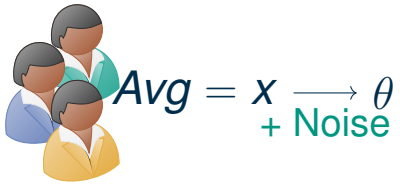
Architecture: Standard multilayer perceptron (MLP)

Hyperparameters:

- Batch sampling rate $q = 1$, Learning rate 9.
- Epochs $T = 100$, Clipping rate $C = 0.1$.

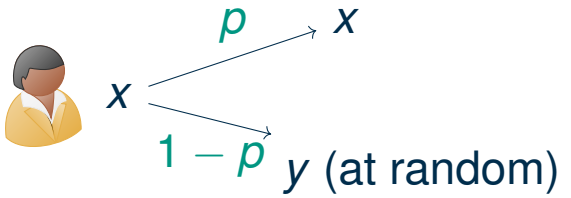
Aggregate Statistics

Laplace Mechanism

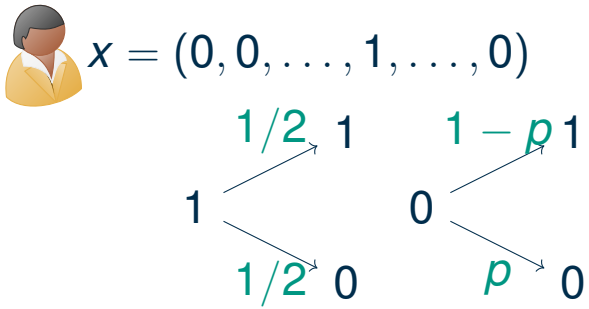


LDP for location data

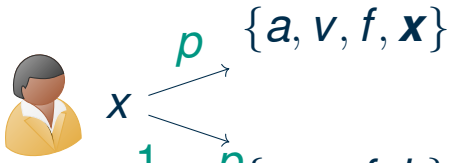
Generalized Randomized Response



Optimal Unary Encoding



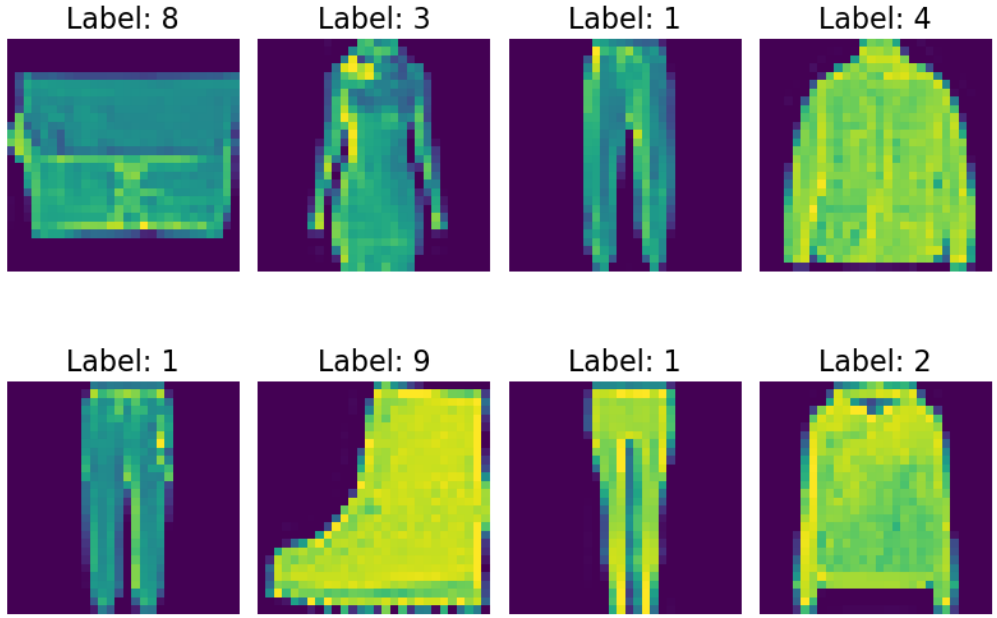
Subset Selection



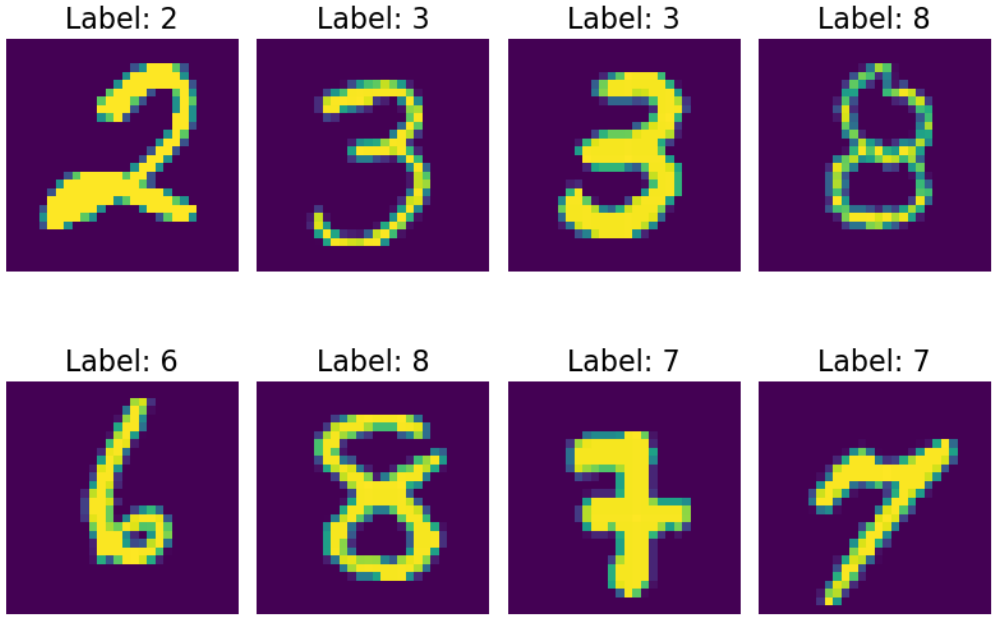
Experimental Details: Datasets

DP-SGD

Fashion-MNIST



MNIST

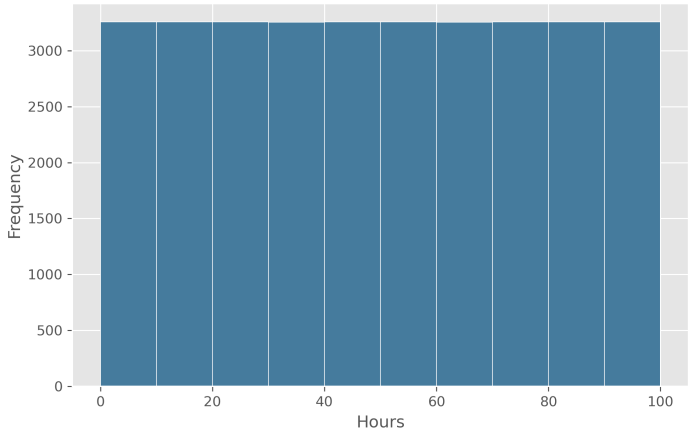


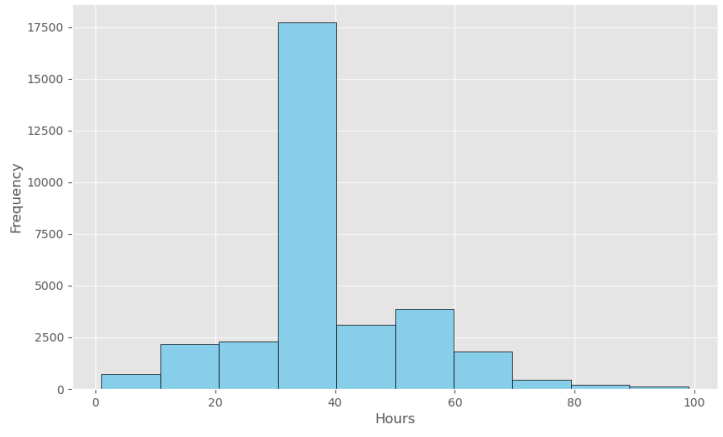
- We set the uniform prior with size $|\mathcal{X}| = 8$ (disjoint from D_-), meaning that $\kappa^+_{\pi,0} = \kappa_\pi = 0.125$.
- We train on $D = 1\,000$ images.

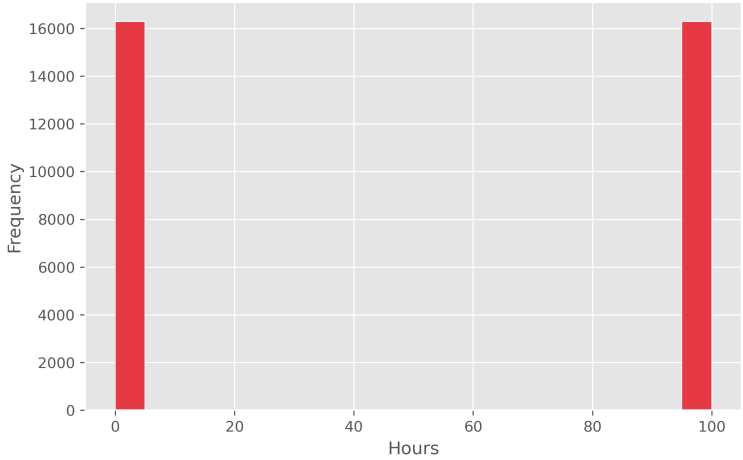
Experimental Details: Datasets

Laplace Mechanism

- Adult dataset (Census), $|D| = 32\,561$
- $\mathcal{X} = [0, \ldots, 100]$ working hours per week.







Experimental Details: Datasets

LDP for Locations

- **Porto**: 83, 409, 386 location reports that we map to the OSM roadgraph at Porto's city center (41.1475° N, 8.5870° W) with a 2.7 km radius, capturing the urban core. $|\mathcal{X}| = 3\,052$.
- **Geolife**: 24 876 978 locations that we mapped to an OSM graph centered near Tiananmen Square (39.9130° N, 116.3703° E) with a 5 km radius covering major central districts, $|\mathcal{X}| = 5\,356$.

Experimental Details: Datasets

Imputation Attack in Census Data

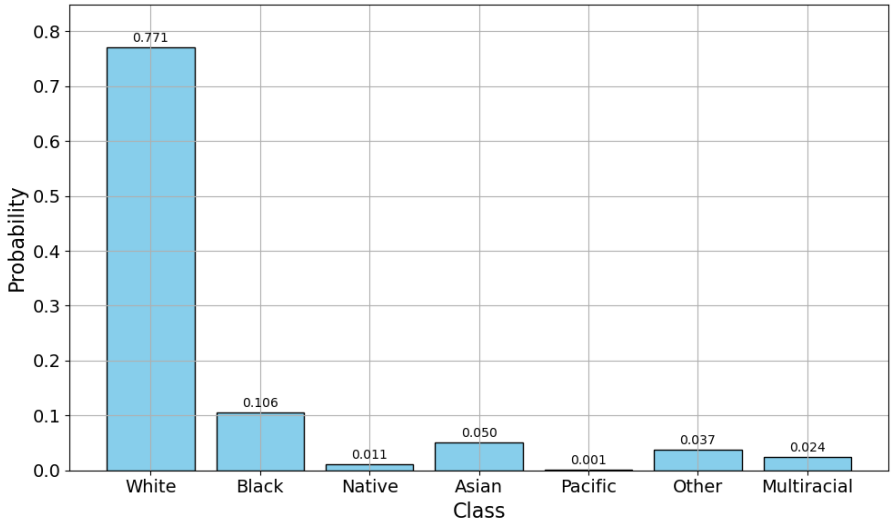


Figure: Ethnicity (Census).

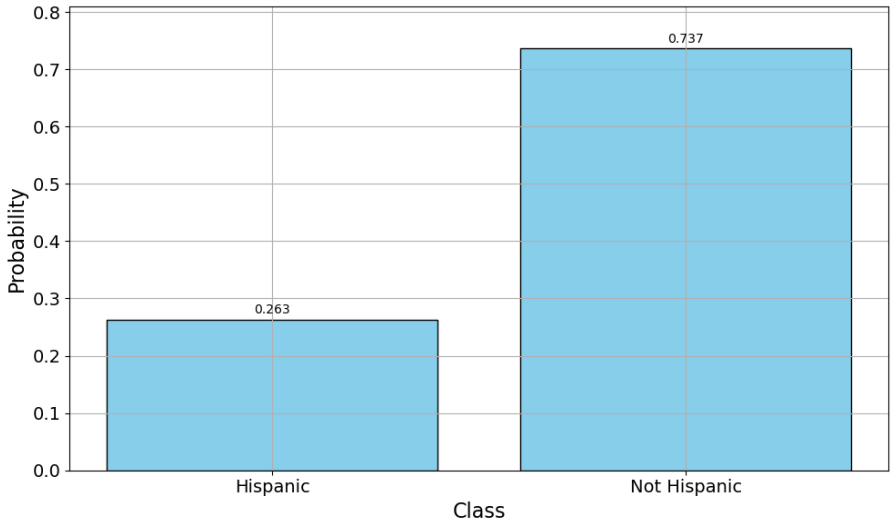
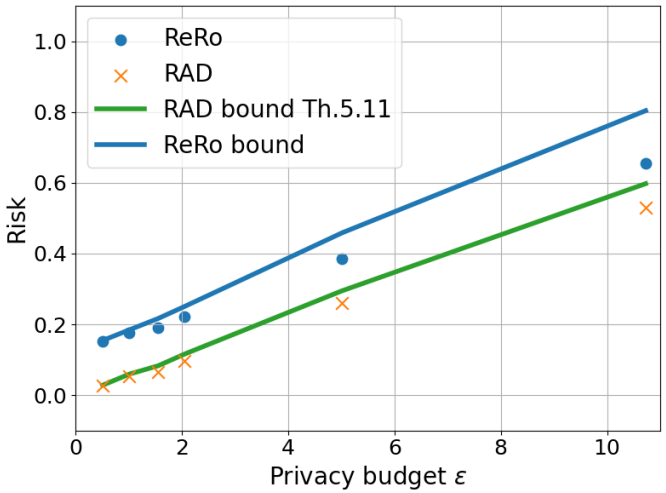
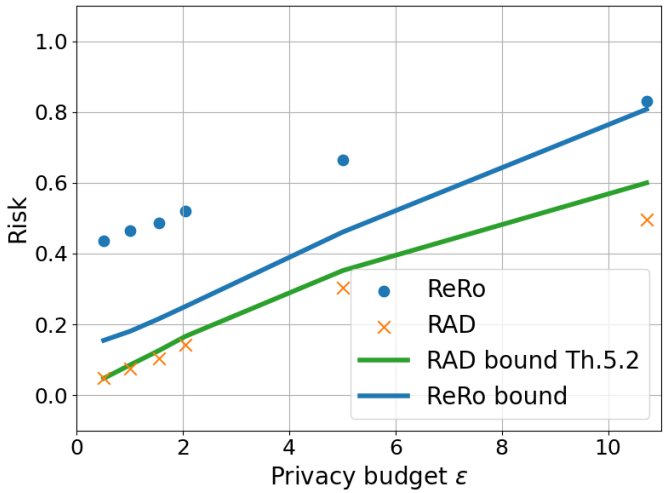


Figure: Ethnicity (Texas-100X).

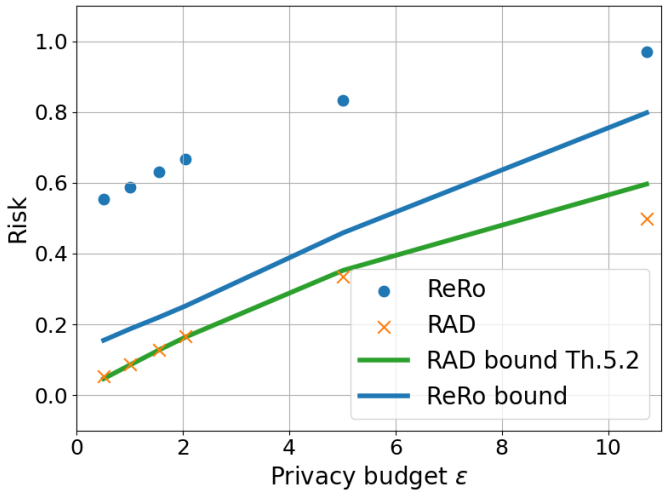
Results Optimal Attack against DP-SGD



DRA, $aux = \{\varnothing\}$.



DRA, $a(x) =$ image label.



MIA, $a(x) = x$.

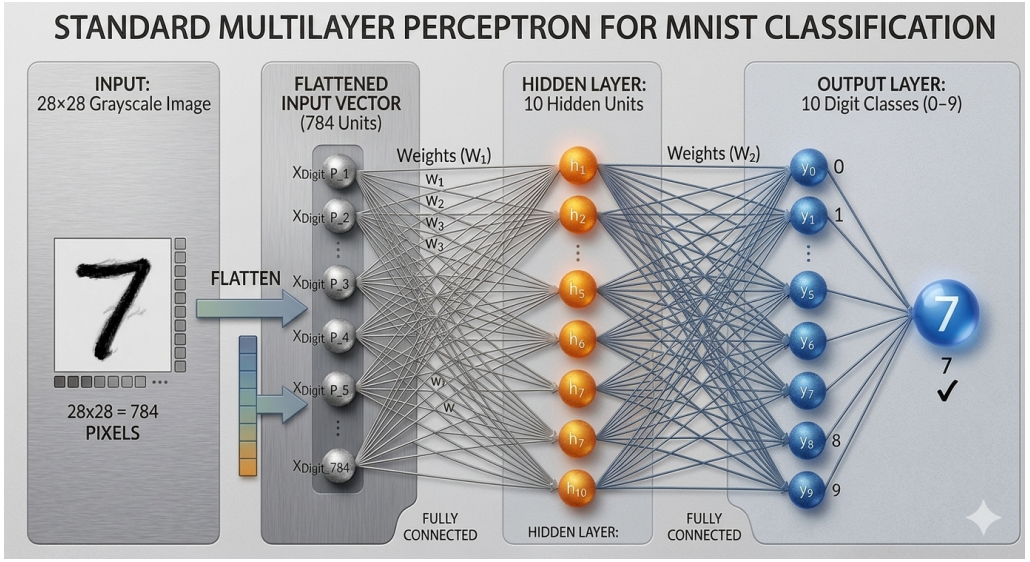
Experimental details:

- $\eta = 0$
- We set $|D_-| = 999$ and $|D_- \cup \{x\}| = 1\,000$
- Full-batch DP-SGD for $T = 100$ steps.
- We set the clipping rate, $C = 0.1$ and $\delta = 10^{-5}$

Details about DP-SGD Experiments

Detailed Architecture:

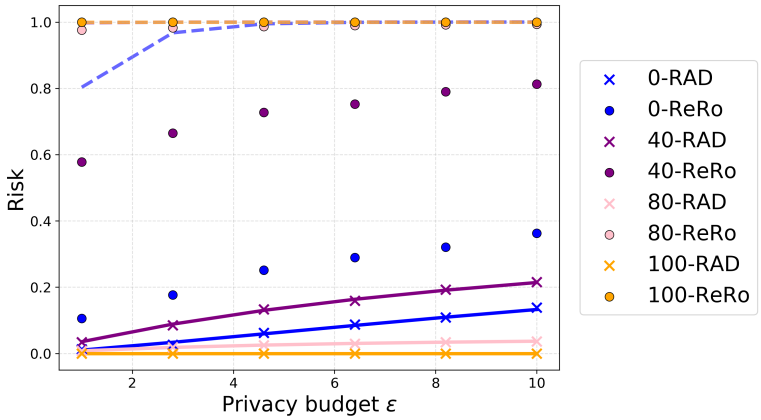
- Standard **multilayer perceptron** (MLP) with 2 layers.
- **Input:** 28×28 grayscale image (784-dimensional vector).
- The input passes through a **fully connected** layer with 10 hidden units, followed by a **ReLU** activation.
- The output layer consists of **10 units**, one for each digit class (0–9).



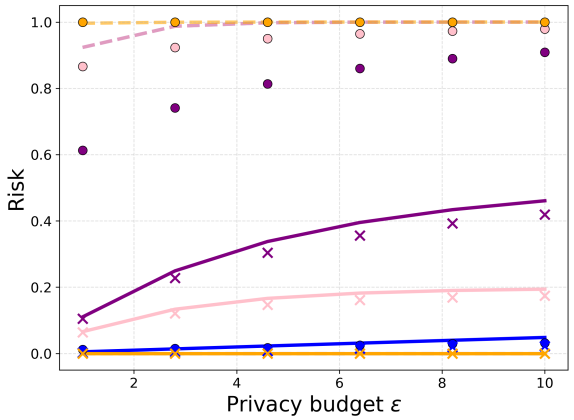
ε	MNIST	Fashion
1	0.46	0.46
10	0.75	0.75
20	0.75	0.77
50	0.75	0.80

Table: Average test accuracy for different ε values under DP-SGD on MNIST and Fashion.

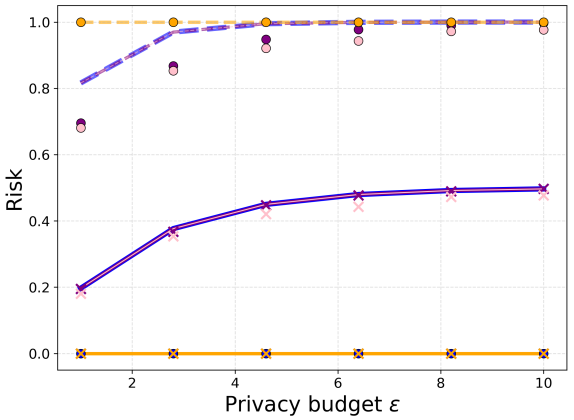
Results Against Laplace Mechanism



Adult distribution.

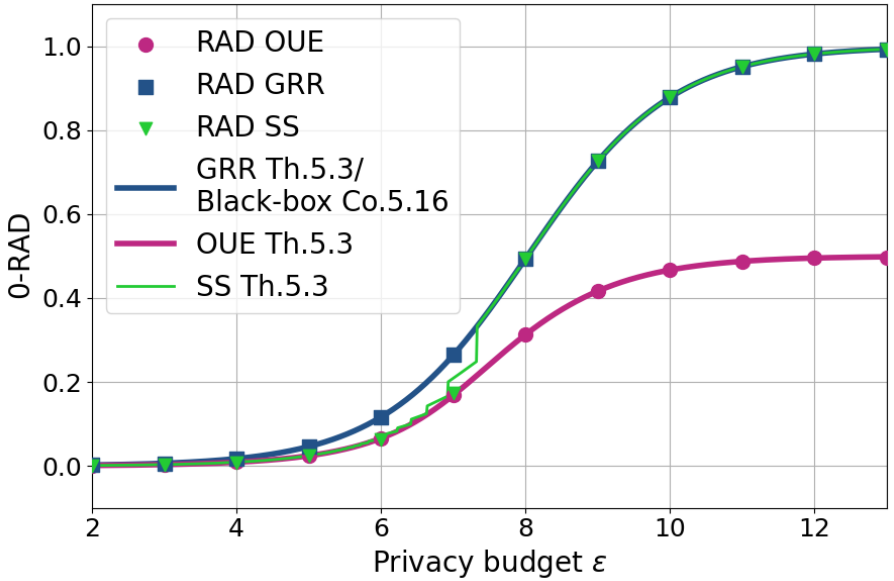


Uniform distribution.

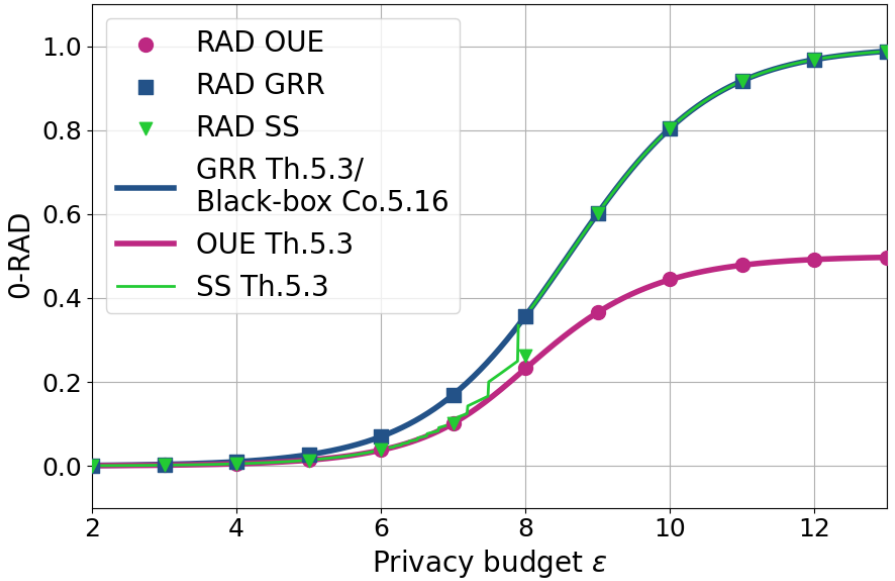


$\pi(0) = \pi(100) = \frac{1}{2}$.

Results Against LDP mechanisms

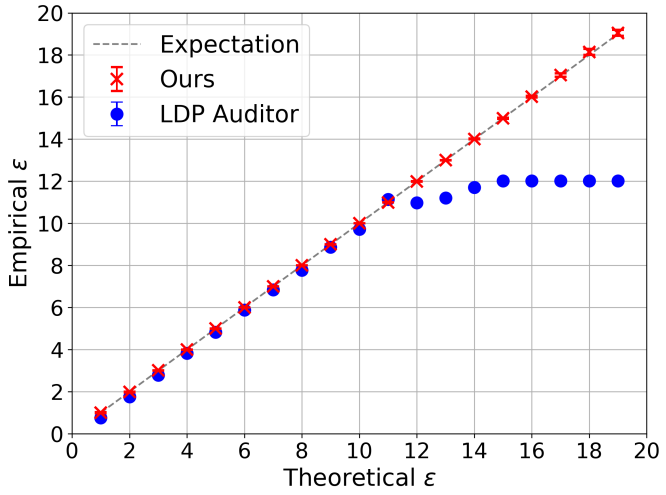


Porto

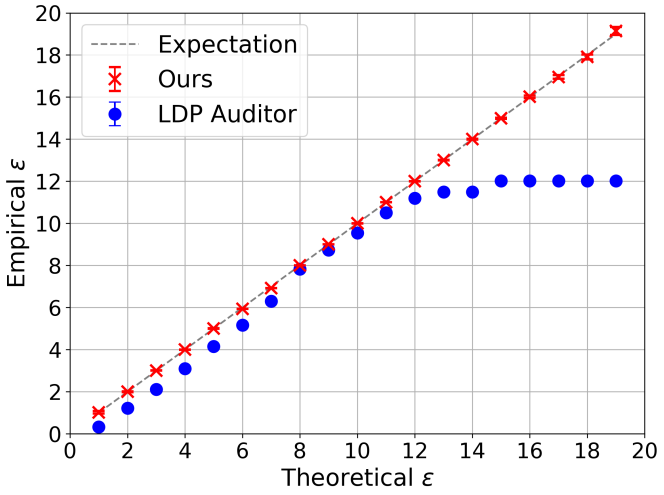


Geolife

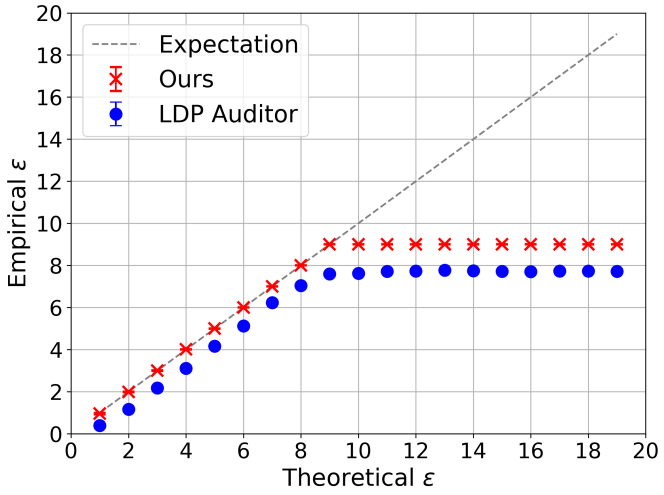
Results RAD-based Auditor



(GRR)



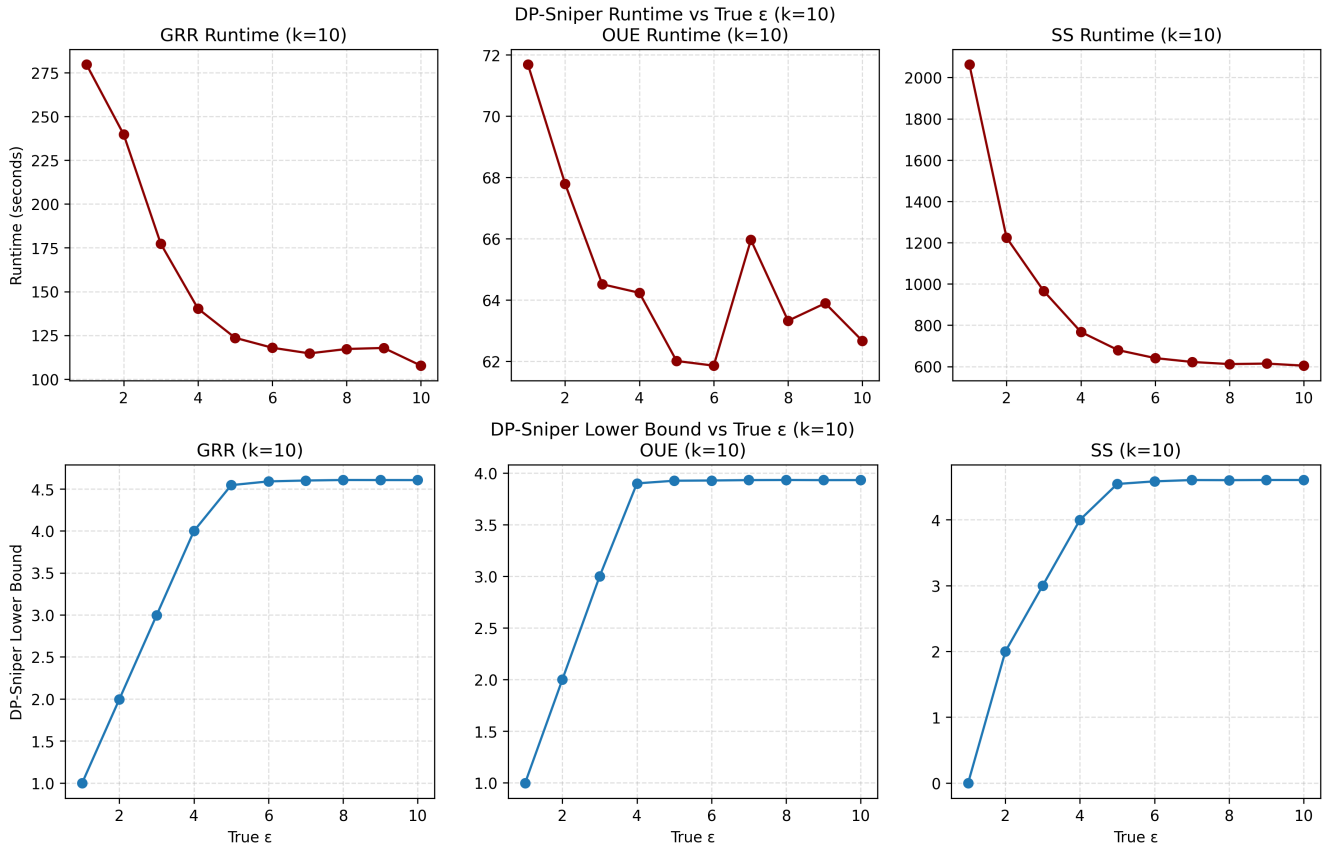
(SS)



(OUE)

Scalability Problem of ML-based Auditing

- $k = |\mathcal{X}|$, Porto, $|\mathcal{X}| = 3\,052$, Geolife, $|\mathcal{X}| = 5\,356$.



Top: inferred DP lower bounds.
Bottom: wall-clock runtime per ε value.

- DP-Sniper saturates near $\varepsilon \approx 4.6$ due to its internal clipping mechanism,
- and requires several minutes per audit.
- Conclusion: computationally infeasible for large categorical LDP mechanisms.

DP-Sniper evaluation on a small categorical domain ($k = 10$) for GRR, OUE, and SS.