Karlsruhe Institute of Technology

PRIVACY AND SECURITY

KASTEL

Understanding Disclosure Risk in DP

with Applications to Noise Calibration and Auditing

Patricia Guerra-Balboa[KIT] | Annika Sauer[KIT] | Héber H. Arcolezi[ÉTS, Inria] | Thorsten Strufe[KIT]

Paper ArXiv

Motivation

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

Interpretation Utility Auditing

- Literature mainly focus on **membership inference** attacks (MIAs).
- Recently, **ReRo** extends risk analysis to data **reconstruction attacks** (DRAs)! but...

η-ReRo Balle et al. [1]

$$\Pr_{\substack{Z_1\sim\pi \\ \theta\sim\mathcal{M}(D_{Z_1})}}[\ell(Z_1, A(\theta))) \leq \eta]$$

✗ Does not consider the impact of **target-specific auxiliary** knowledge,

✗ and **overestimates** participation **risk** due to prior knowledge and/or imputation.

η-RAD (Our solution)

$$\Pr_{\substack{Z_1\sim\pi \\ \theta\sim\mathcal{M}(D_{Z_1})}}[\ell(Z_1, A(\theta, a(Z_1))) \leq \eta] - \Pr_{\substack{Z_0,Z_1\sim\pi \\ \theta\sim\mathcal{M}(D_{Z_0})}}[\ell(Z_1, A(\theta, a(Z_1)) \leq \eta].$$

✔ Takes the auxiliary knowledge, $a(z)$, into account!

✔ Discounts imputation and prior knowledge!

Dataset	ReRo	RAD
Census	0.81	0
Texas	0.73	0

Our Novel Bounds

Th.4.3. (Universally Tight Bound)

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

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

Th. 4.2 (unknown aux)

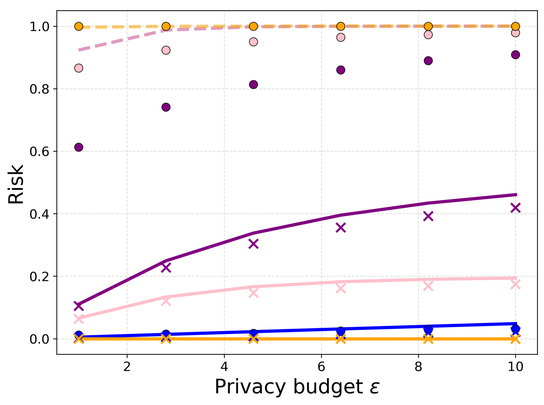
$$\eta\text{-RAD} \leq \text{TV}(\mathcal{M})(1-\kappa_\pi) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}(1-\kappa_\pi),$$

where TV is the total variation and κ_π the re-sampling probability.

✔ Independent of the auxiliary knowledge.

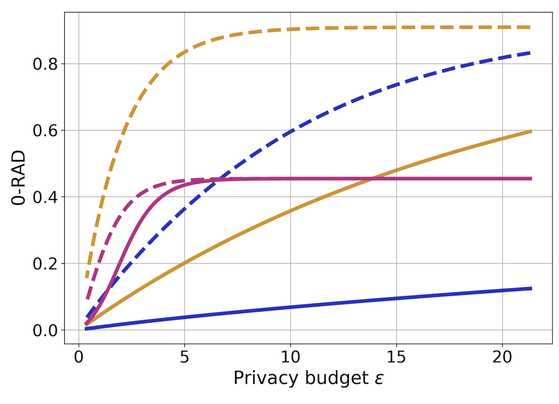
Notion	Assumptions	RAD	Composition	ReRo
TV	—	Th.4.2.	✓	∄
f-DP	$aux = \{\varnothing\}$	Th.5.1.	✓	[2]
$\mathcal{M}$	aux known	Th.4.3.	×	∄
(ε, δ)-DP	—	Th.4.2.	✓	∄
(ε, δ)-DP	$aux = \{\varnothing\}$	Prop.5.3.	✓	[1]
(ε, δ)-DP	$aux = \{\varnothing\}, \eta = 0$	Th.5.5.	✓	[1]

Th.4.3. captures the effect of the threshold η overcoming ReRo overestimation.



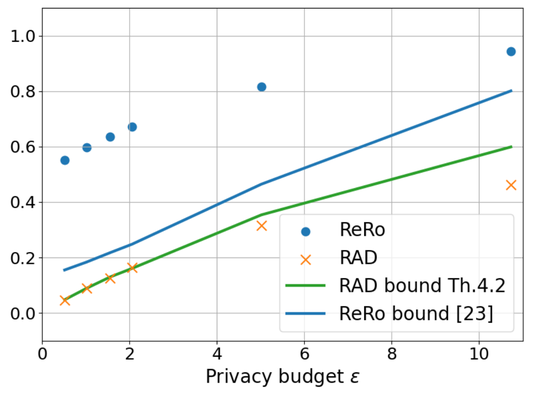
Our Optimal Attack Against Laplace Mech. ●=ReRo, ×=RAD, η =0, 40, 80, 100.

Different protection under same ε! Knowing *aux* significantly improves the risk estimation.



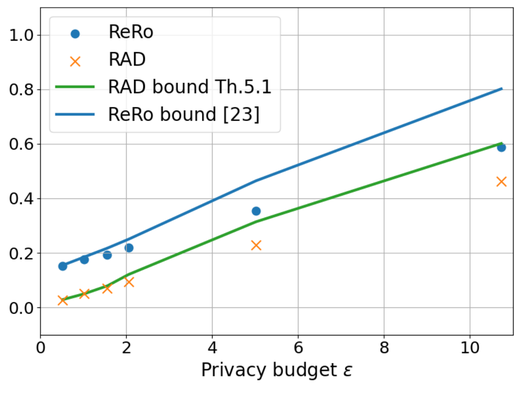
RAD bounds for Gaussian, OUE, and Laplace. –– Th.4.2., — Th.4.3.

Th. 4.2. nearly tight bound! ReRo bounds are broken when $aux \neq \{\varnothing\}$.



Optimal Attack with $a(z) = z$ in DP-SGD

Th. 5.1. serves as a reasonable upper bound when Th.4.3 can not be computed.

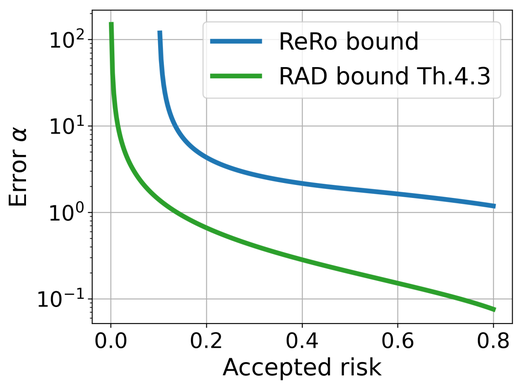


Optimal Attack with $a(z) = \varnothing$ in DP-SGD

Impact on Utility and Auditing

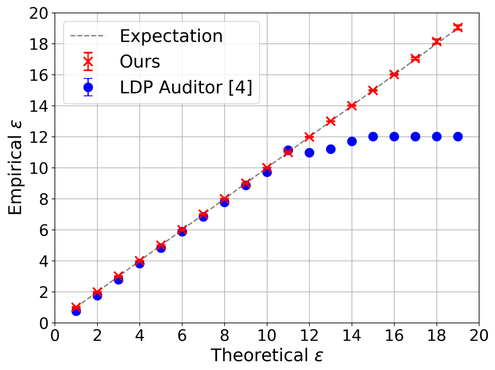
- From Th.4.3: **Optimal Noise Injection** for any given admitted **risk**.
- When *aux*, $\mathcal{M}$ or π are not known, **our upper-bounds** offer reasonable risk estimations and enjoy composition.
- Combining **our optimal attack** algorithm and our bounds we obtain a **DP auditing method**.
- RAD-based auditing **scales better** than ML-based methods while offering **better accuracy** than LDP Auditor.

Improved Utility by Recalibrating Laplace Noise to RAD

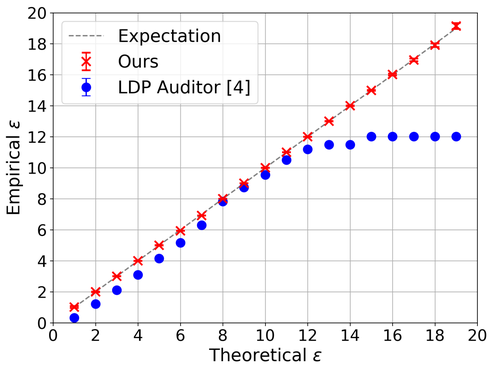


Accuracy Improvement of RAD-based Auditing

(GRR)



(SS)



Conclusion

RAD together with our theoretical results enable a more interpretable reasoning about the offered privacy guarantees, calibrating the noise exactly to the desired risk level, detecting the differences between mechanisms and enabling more accurate and scalable DP auditing.

[1] Borja Balle, Giovanni Cherubin, and Jamie Hayes. "Reconstructing Training Data with Informed Adversaries". In: *IEEE Symposium on Security and Privacy*. 2022, pp. 1138–1156. DOI: 10.1109/SP46214.2022.9833677.

[2] Jamie Hayes, Borja Balle, and Saeed Mahloujifar. "Bounding training data reconstruction in DP-SGD". in: *Advances in Neural Information Processing Systems (NeurIPS'23)*. Vol. 36. 2023, pp. 78696–78722.