Karlsruhe Institute of Technology

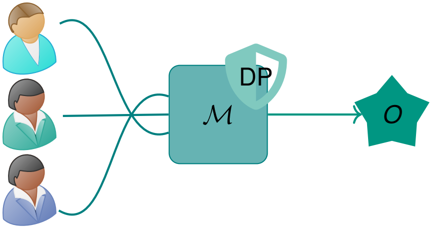
PRIVACY AND SECURITY

Analysis and Measurement of Attack Resilience of Differential Privacy

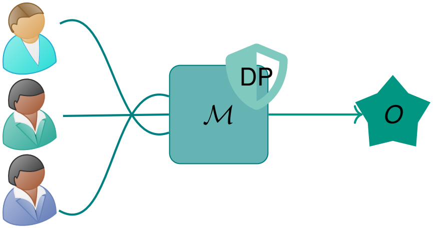
Workshop on Privacy in the Electronic Society 2024 (WPES2024)

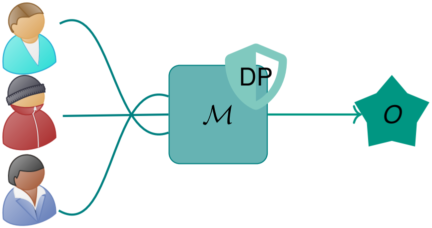
Patricia Guerra-Balboa, Annika Sauer, Thorsten Strufe | 14th October 2024

Differential Privacy (Bounded)



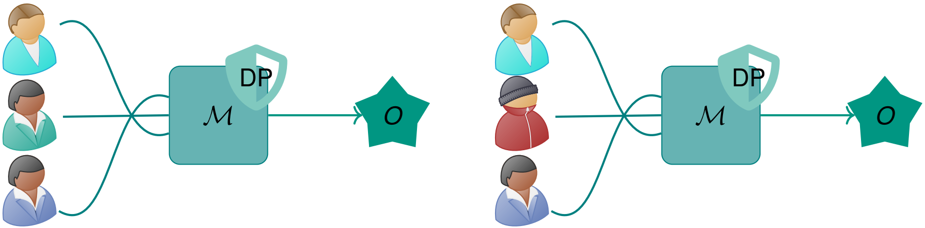
Differential Privacy (Bounded)





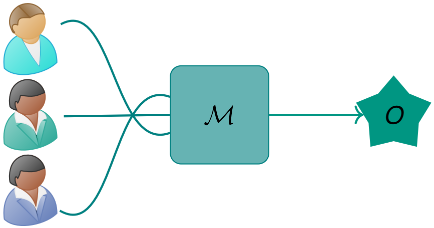
Differential Privacy (Bounded)

$$\Pr(O|\textit{Alice}) \leq e^{\varepsilon} \Pr(O|\textit{Bob})$$

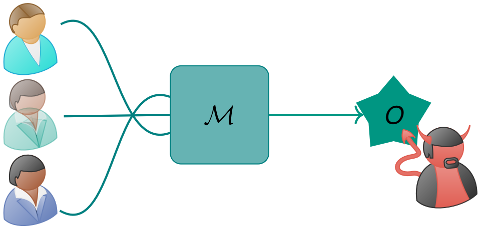


- More $\varepsilon \Rightarrow$ more indistinguishability & less utility
- How can we choose ε to mitigate the attacks?

Attacks on Private Data



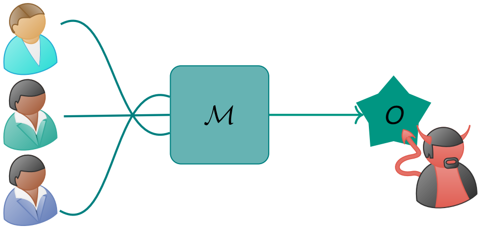
Attacks on Private Data



$z \in D$?

Membership Inference Attack

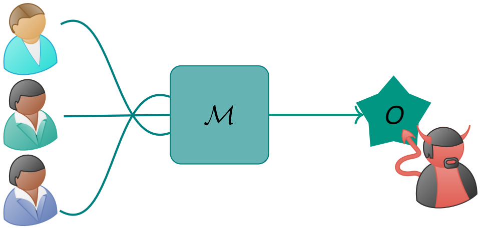
Attacks on Private Data



$z \in D?$ $\qquad$ $\varphi(z)?$

Membership Inference Attack $\subset$ Attribute Inference Attack

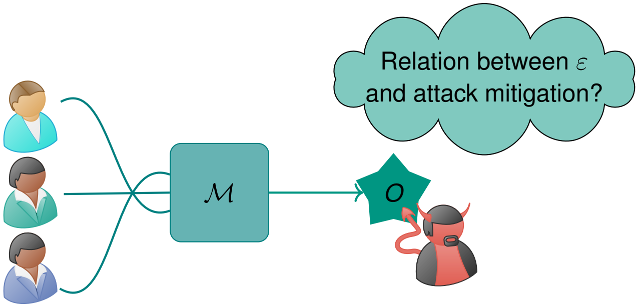
Attacks on Private Data



$z \in D?$ | $\varphi(z)?$ | $z'? \,:\, l(z, z') \leq \eta$

Membership Inference Attack $\subset$ Attribute Inference Attack $\subset$ Data Reconstruction Attack

Attacks on Private Data

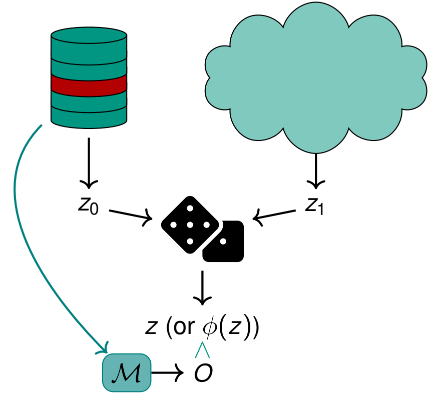


$z \in D?$ | $\varphi(z)?$ | $z'? \; : \; l(z, z') \leq \eta$

Membership Inference Attack $\subset$ Attribute Inference Attack $\subset$ Data Reconstruction Attack

Adversarial bounds until now

Membership & Attribute Advantage



Given that $\theta \sim \mathcal{M}(D)$ and $D \sim \pi^n$, then:

- **Membership Advantage** (Adv_{MIA})

$$\Pr(A(\theta) = 0 | b = 0) - \Pr(A(\theta) = 0 | b = 1)$$

- **Attribute Advantage** (Adv_{AIA})

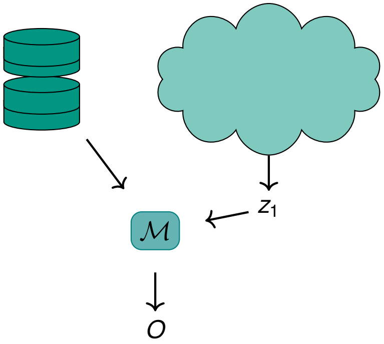
$$\Pr(A(\theta) = \varphi(z) | b = 0) - \Pr(A(\theta) = \varphi(z) | b = 1)$$

- Existing bounds (Humphries et al.):

MIA	AIA
$\mathrm{Adv}^{s}_{\mathrm{MIA}} \leq \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$	✘

Adversarial bounds until now

Reconstruction robustness



- **Reconstruction Robustness** ((η, γ)-ReRo)

$$\Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_Z)}} [I(Z, A(\theta)) \leq \eta] \leq \gamma.$$

- Existing bound (Balle et al.):

$$\gamma = \kappa_{\pi,I}(\eta)e^{\varepsilon}$$

Where $\kappa_{\pi,I}(\eta) = \sup_{z'\in\mathcal{Z}} \Pr_{Z\sim\pi}[I(Z, z') \leq \eta]$

Q1: Can we find tighter bounds ?

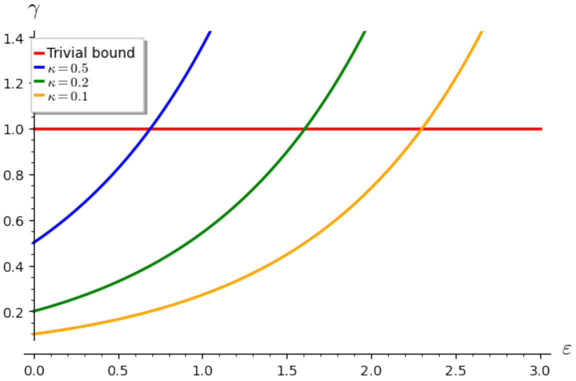


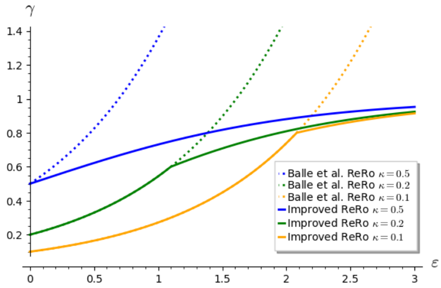
Figure: Balle et al. bound for ReRo

A1: Our Improved bound for Perfect Reconstruction Robustness

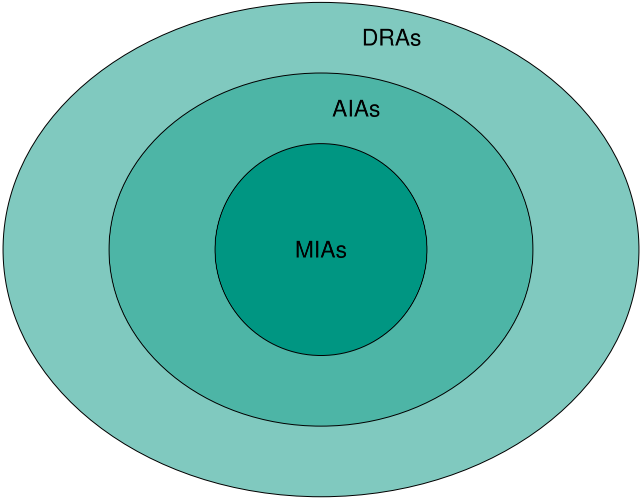
Improved Bound for ReRo against perfect reconstruction

If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ satisfies ε-DP, then it also satisfies $(0, \gamma)$-ReRo with

$$\gamma \leq \min\{\kappa_0 e^{\varepsilon}, \kappa_0 \left(1 + (m-1)\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\right)\}$$

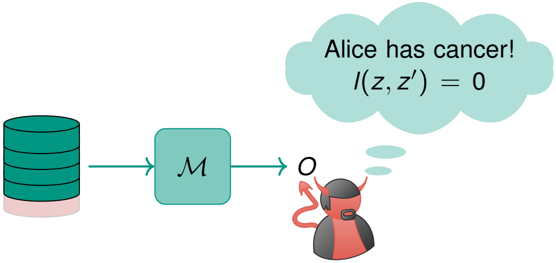


Q2: Can we ReRo as general attack performance metric that allows comparison?



A2: ReRo and the privacy fallacy

A successful reconstruction $\not\Rightarrow$ Privacy Leakage



Our Solution: Unbiased Reconstruction Robustness

Karlsruhe Institute of Technology

Unbiased Reconstruction Robustness (U-ReRo)

A randomized learning mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ is (η, γ)-U-ReRo, with respect to π and I if for any dataset $D \in \mathcal{Z}^{n-1}$ and any reconstruction attack $A\colon \Theta \to \mathcal{Z}$ we have

$$\Pr_{\substack{Z \sim \pi \\ \theta \sim \mathcal{M}(D_Z)}} [I(Z, A(\theta)) \leq \eta] - \mathbb{E}_{Z_0 \sim \pi} \left(\Pr_{\substack{Z \sim \pi \\ \theta \sim \mathcal{M}(D_{Z_0})}} [I(Z, A(\theta)) \leq \eta] \right) \leq \gamma.$$

U-ReRo is consistent with previous advantages

Karlsruhe Institute of Technology

$\mathrm{Adv}_{\mathrm{AIA}} \Leftrightarrow$ U-ReRo

$$\mathcal{M} \text{ is } (0,\gamma)\text{-U-ReRo} \iff \mathrm{Adv}_{\mathrm{AIA}}(A, \mathcal{M}, \pi^n) \leq \gamma \text{ for all } A.$$

$\mathrm{Adv}_{\mathrm{MIA}} \Leftrightarrow$ U-ReRo

$$\mathcal{M} \text{ is } (0,\gamma)\text{-U-ReRo} \iff \mathrm{Adv}_{\mathrm{MIA}}(A, \mathcal{M}, \pi^n) \leq \gamma,$$

Additionally, if A^s is a strong MIA under uniform priors, then

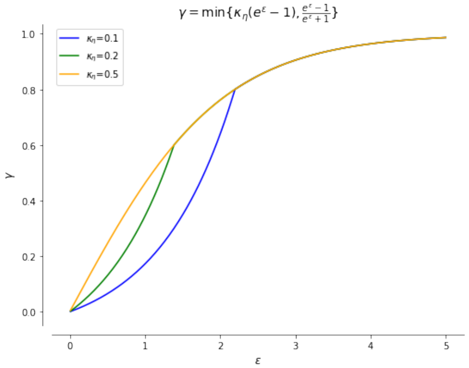
$$\mathcal{M} \text{ is } (0,\frac{\gamma}{2})\text{-U-ReRo} \iff \mathrm{Adv}^s_{\mathrm{MIA}}(A, \mathcal{M}, \pi^n) \leq \gamma.$$

New Adversarial bounds for U-ReRo

ε-DP $\Rightarrow$ (η, γ)-U-ReRo

If $\mathcal{M}$ satisfies ε-DP, then it also satisfies (η, γ)-U-ReRo with

$$\gamma = \min\{\kappa_\eta(e^\varepsilon - 1), \frac{e^\varepsilon - 1}{e^\varepsilon + 1}\}$$

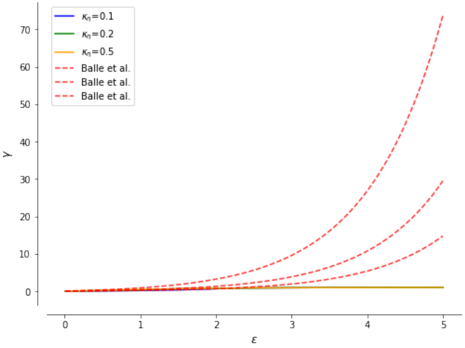


New Adversarial bounds for U-ReRo

ε-DP $\Rightarrow (\eta, \gamma)$-U-ReRo

If $\mathcal{M}$ satisfies ε-DP, then it also satisfies (η, γ)-U-ReRo with

$$\gamma = \min\{\kappa_\eta(\mathrm{e}^\varepsilon - 1), \frac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + 1}\}$$



New Adversarial bounds for U-ReRo

ε-DP $\Rightarrow$ (η, γ)-U-ReRo

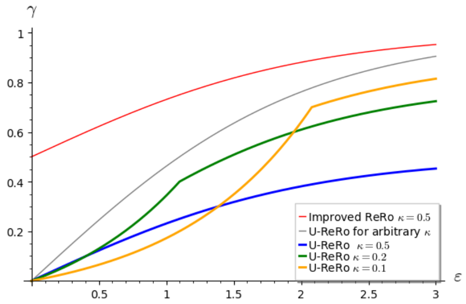
If $\mathcal{M}$ satisfies ε-DP, then it also satisfies (η, γ)-U-ReRo with

$$\gamma = \min\{\kappa_\eta(e^\varepsilon - 1), \frac{e^\varepsilon - 1}{e^\varepsilon + 1}\}$$

ε-DP $\Rightarrow$ $(0, \gamma)$-U-ReRo (AIA)

If $\mathcal{M}$ satisfies ε-DP, then it also satisfies $(0, \gamma)$-ReRo with

$$\gamma = \min\{\kappa_0(e^\varepsilon - 1), \kappa_0(m - 1)\frac{e^\varepsilon - 1}{e^\varepsilon + 1} + \kappa_0 - \kappa_0^-\},$$



Conclusions

Conclusions

- ✗ ReRo overestimates the privacy leakage

Attack	Our Improved Bound
MIA Strongest	$\frac{\gamma}{2} \leq \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$
MIA Informed	$\gamma \leq \min\{\frac{1}{m}(e^{\varepsilon}-1), \frac{m-1}{m}\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
AIA Informed	$\gamma \leq \min\{\frac{e^{\varepsilon}}{m}, \frac{m-1}{m}\left(\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}+1\right)\}$
AIA Inf. Uniform	$\gamma \leq \min\{\frac{1}{m}(e^{\varepsilon}-1), \frac{m-1}{m}\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_{\eta}(e^{\varepsilon}-1), \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_0(e^{\varepsilon}-1), \kappa_0(m-1)\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}+\kappa_0-\kappa_0^{-}\}$

Conclusions

Karlsruhe Institute of Technology

- ✗ ReRo overestimates the privacy leakage
- ✓ our (η, γ)-U-ReRo generalizes the membership and attribute advantages to arbitrary reconstruction attacks

Attack	Our Improved Bound
MIA Strongest	$\frac{\gamma}{2} \leq \frac{e^\varepsilon - 1}{e^\varepsilon + 1}$
MIA Informed	$\gamma \leq \min\{\frac{1}{m}(e^\varepsilon - 1), \frac{m-1}{m}\frac{e^\varepsilon-1}{e^\varepsilon+1}\}$
AIA Informed	$\gamma \leq \min\{\frac{e^\varepsilon}{m}, \frac{m-1}{m}\left(\frac{e^\varepsilon-1}{e^\varepsilon+1} + 1\right)\}$
AIA Inf. Uniform	$\gamma \leq \min\{\frac{1}{m}(e^\varepsilon - 1), \frac{m-1}{m}\frac{e^\varepsilon-1}{e^\varepsilon+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_\eta(e^\varepsilon - 1), \frac{e^\varepsilon-1}{e^\varepsilon+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_0(e^\varepsilon - 1), \kappa_0(m-1)\frac{e^\varepsilon-1}{e^\varepsilon+1} + \kappa_0 - \kappa_0^-\}$

Conclusions

Karlsruhe Institute of Technology

- ✗ ReRo overestimates the privacy leakage
- ✓ our (η, γ)-U-ReRo generalizes the membership and attribute advantages to arbitrary reconstruction attacks
- ✓ Our results allow to choose lower privacy parameters (ε), achieving better utility without increasing privacy risks

Attack	Our Improved Bound
MIA Strongest	$\frac{\gamma}{2} \leq \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$
MIA Informed	$\gamma \leq \min\{\frac{1}{m}(e^{\varepsilon}-1), \frac{m-1}{m}\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
AIA Informed	$\gamma \leq \min\{\frac{e^{\varepsilon}}{m}, \frac{m-1}{m}\left(\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}+1\right)\}$
AIA Inf. Uniform	$\gamma \leq \min\{\frac{1}{m}(e^{\varepsilon}-1), \frac{m-1}{m}\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_{\eta}(e^{\varepsilon}-1), \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_0(e^{\varepsilon}-1), \kappa_0(m-1)\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}+\kappa_0-\kappa_0^{-}\}$

Conclusions

- ✗ ReRo overestimates the privacy leakage
- ✓ our (η, γ)-U-ReRo generalizes the membership and attribute advantages to arbitrary reconstruction attacks
- ✓ Our results allow to choose lower privacy parameters (ε), achieving better utility without increasing privacy risks
- ✓ We use U-ReRo to prove a novel bound for the advantage of an arbitrary AIA under DP.

Attack	**Our Improved Bound**
MIA Strongest	$\frac{\gamma}{2} \leq \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$
MIA Informed	$\gamma \leq \min\{\frac{1}{m}(e^{\varepsilon}-1), \frac{m-1}{m}\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
AIA Informed	$\gamma \leq \min\{\frac{e^{\varepsilon}}{m}, \frac{m-1}{m}\left(\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}+1\right)\}$
AIA Inf. Uniform	$\gamma \leq \min\{\frac{1}{m}(e^{\varepsilon}-1), \frac{m-1}{m}\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_{\eta}(e^{\varepsilon}-1), \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_0(e^{\varepsilon}-1), \kappa_0(m-1)\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}+\kappa_0-\kappa_0^{-}\}$

Conclusions

- ✗ ReRo overestimates the privacy leakage
- ✓ our (η, γ)-U-ReRo generalizes the membership and attribute advantages to arbitrary reconstruction attacks
- ✓ Our results allow to choose lower privacy parameters (ε), achieving better utility without increasing privacy risks
- ✓ We use U-ReRo to prove a novel bound for the advantage of an arbitrary AIA under DP.

Attack	Our Improved Bound
MIA Strongest	$\frac{\gamma}{2} \leq \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$
MIA Informed	$\gamma \leq \min\{\frac{1}{m}(e^{\varepsilon}-1), \frac{m-1}{m}\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
AIA Informed	$\gamma \leq \min\{\frac{e^{\varepsilon}}{m}, \frac{m-1}{m}\left(\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}+1\right)\}$
AIA Inf. Uniform	$\gamma \leq \min\{\frac{1}{m}(e^{\varepsilon}-1), \frac{m-1}{m}\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_{\eta}(e^{\varepsilon}-1), \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\}$
DRA Informed	$\gamma \leq \min\{\kappa_0(e^{\varepsilon}-1), \kappa_0(m-1)\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}+\kappa_0-\kappa_0^{-}\}$

Thanks for your attention!

Backup 1: The different advantages

Formal definition of advantage:

$$\mathrm{Adv}_{\mathrm{MIA}}^{(*)} = 2\Pr[\mathrm{Exp}_{(*)}^{\mathrm{MIA}}] - 1 \tag{1}$$

We have the following relationship between the advantage of a strong membership experiment with resampling and without resampling:

$$\mathrm{Adv}_{\mathrm{MIA}} = 2\Pr[\mathrm{Exp}^{MIA}] - 1 = \Pr[\mathrm{Exp}_{s}^{MIA}] - \frac{1}{2} = \frac{1}{2}\mathrm{Adv}_{MIA}^{s}$$

This is coherent with the fact that $\mathrm{Adv}_{\mathrm{MIA}}$ is upper-bounded by $\frac{1}{2}$ in a strong membership experiment.

Backup 2: Worse case or average case?

- In general attack performance metrics are **average-case**

$$\Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_Z)}}(I(A(\theta),Z)=0)=\sum_{z\in\mathcal{Z}}\Pr_{\theta\sim\mathcal{M}(D_z)}(I(A(\theta),z)=0)\pi(z)$$

- **We can make them worse-case** by modifying the universe distribution to $z\in\{z_0,z_1\}$
  - we can choose z_0 to be the worse case
  - the bound will still hold but the baseline error will adjust it.