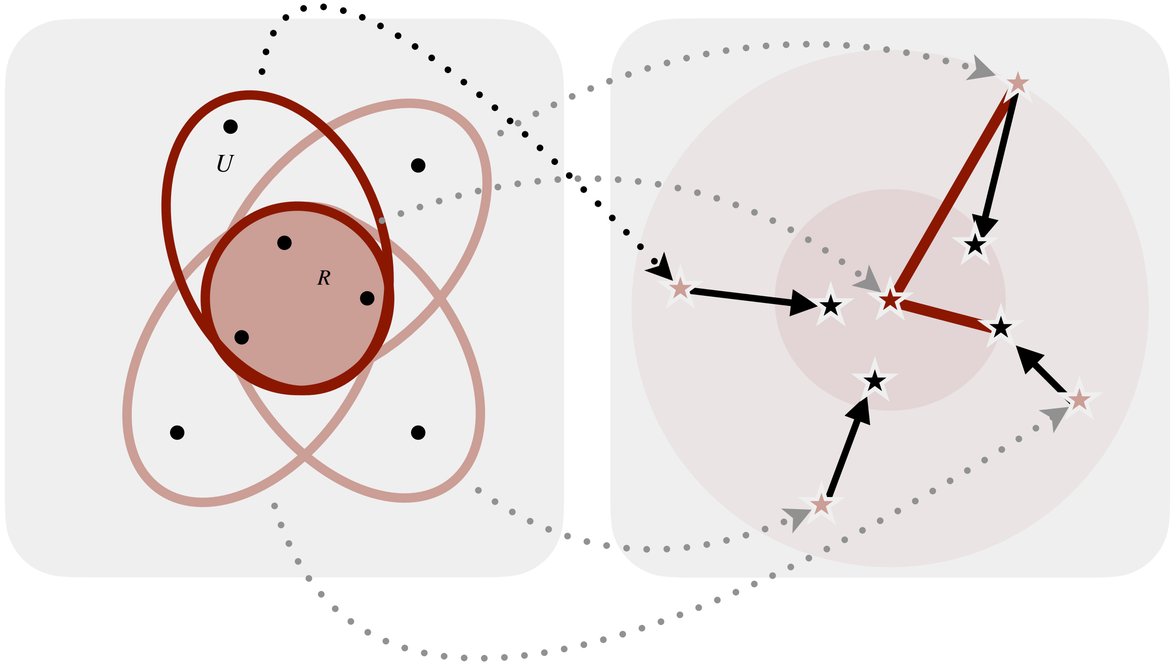
Less Noise, Same Certificate: Retain Sensitivity for Unlearning

Carolin Heinzler
Kasra Malihi
Amartya Sanyal



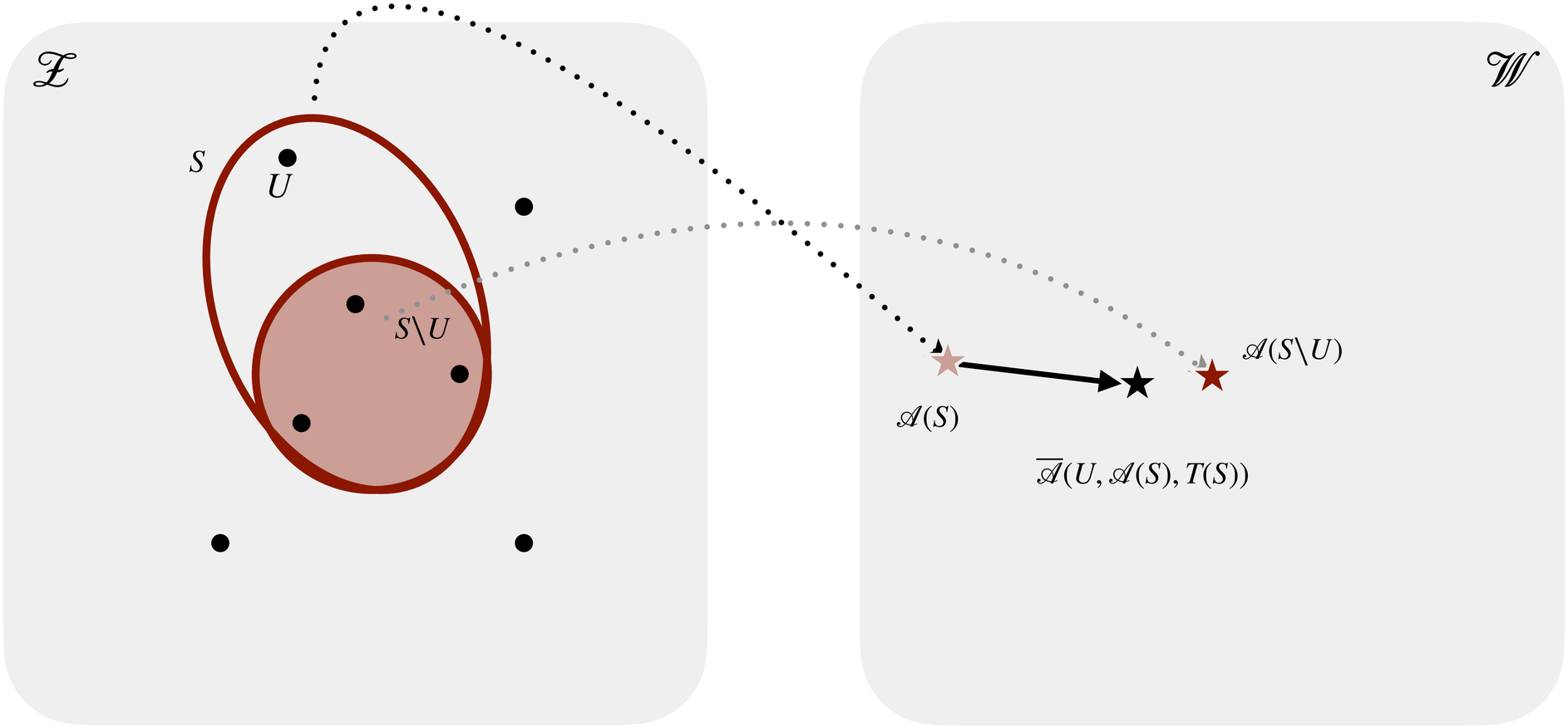
MACHINE LEARNING
UNIVERSITY OF COPENHAGEN

Foundations of Responsible Machine Learning (FoRML)

University of Copenhagen,
Computer Science Department

FoRML
Foundations of Responsible Machine Learning

Unlearning


$\mathcal{Z}$
S
U
$S \backslash U$
$\mathcal{W}$
$\mathcal{A}(S \backslash U)$
$\mathcal{A}(S)$
$\overline{\mathcal{A}}(U, \mathcal{A}(S), T(S))$

Certified Unlearning

Let $\varepsilon, \delta > 0$. A learning-unlearning algorithm pair $(\mathscr{A}, \overline{\mathscr{A}})$ satisfies (ε, δ)-Unlearning, if:

- for all **Datasets** $S \subseteq \mathscr{Z}, |S| = n + m$
- and for all **Unlearning sets** $U \subseteq S, |U| \leq m$

$$\overline{\mathscr{A}}(U, \mathscr{A}(S), T(S)) \overset{\varepsilon,\delta}{\approx} \overline{\mathscr{A}}(\varnothing, \mathscr{A}(S \backslash U), T(S \backslash U))$$

Unlearning — Retraining

Both depend on the **same** $S \backslash U$

Certified Unlearning

Retain Perspective

Let $\varepsilon, \delta > 0$. A learning-unlearning algorithm pair $(\mathscr{A}, \overline{\mathscr{A}})$ satisfies (ε, δ)-Unlearning, if:

- for all **Retain Sets** $R \subseteq \mathscr{Z}$
- and for all **Unlearning sets** $U \subseteq \mathscr{Z}, |U| \leq m$, $|R \cup U| = n + m$

$$\overline{\mathscr{A}}(U, \mathscr{A}(R \cup U), T(R \cup U)) \overset{\varepsilon,\delta}{\approx} \overline{\mathscr{A}}(\emptyset, \mathscr{A}(R), T(R))$$

Unlearning | Retraining

Differential Privacy

Let $\varepsilon, \delta > 0$. A randomised mechanism $\mathscr{M}$ satisfies (ε, δ)-differential privacy if for all **neighbouring datasets** S, S' (add/remove adjacency)

$$\mathscr{M}(S) \overset{\varepsilon,\delta}{\approx} \mathscr{M}(S')$$

Assume

$$\overline{\mathscr{A}}(U, \mathscr{A}(S), T(S)) = \overline{\mathscr{A}}_0(U, \mathscr{A}(S), T(S)) + \nu$$

$\overline{\mathscr{A}}_0$ deterministic map

$\nu \sim \mathcal{N}(0, \sigma^2 I_d)$ i.i.d
σ may depend on available information $(U, T(S))$ or $(\varnothing, T(S \backslash U))$

What is minimum noise necessary to guarantee unlearning?

If $\overline{\mathscr{A}}_0(U, \mathscr{A}(S), T(S)) = \mathscr{A}(S)$ → Passive Unlearning

Else → Active Unlearning

Full Side Information $T(S) = S$

Noise ~ Sensitivity: DP

DP

Global Sensitivity: $\mathrm{GS}_f = \max_{S,S':d_{\pm}(S,S')=1} \|f(S) - f(S')\|$

No DP

Local Sensitivity: $\mathrm{LS}_f(S) = \max_{S':d_{\pm}(S,S')=1} \|f(S) - f(S')\|$

Noise ~ Sensitivity: Unlearning

For $f = \overline{\mathscr{A}}_0$,
$R = S \backslash U$,
$\sigma \sim$ Sensitivity

DP

Global Sensitivity: $\mathrm{GS}_f^{(m)} = \max_{S,S':d_{\pm}(S,S')=m} \|f(S) - f(S')\|$

Unlearning

No DP

Local Sensitivity: $\mathrm{LS}_f^{(m)}(S) = \max_{S':d_{\pm}(S,S')=m} \|f(S) - f(S')\|$

No Unlearning for $\mathrm{LS}_f^{(m)}(S)$

Unlearning for $\mathrm{LS}_f^{(m)}(R)$

Ours

No DP

Retain Sensitivity: $\mathrm{RS}_f^{(m)}(S) = \max_{Z \subseteq \mathscr{Z}:|Z| \le m} \|f(S) - f(S \cup Z)\|$

Unlearning for $\mathrm{RS}_f^{(m)}(R)$

No DP

Down Sensitivity: $\mathrm{DS}_f^{(m)}(S) = \max_{Z \subseteq \mathscr{Z}:|Z| \le m} \|f(S) - f(S \backslash Z)\|$

No Unlearning for $\mathrm{DS}_f^{(m)}(S)$

Result

Let $\varepsilon, \delta > 0$. For all $R \subseteq \mathcal{Z}$, for all $U \subseteq \mathcal{Z}$, $|U| \leq m$, $|R| + |U| = n + m$:

$$\overline{\mathcal{A}}(U, \mathcal{A}(R \cup U), R \cup U) = \overline{\mathcal{A}}_0(U, \mathcal{A}(R \cup U), R \cup U) + \mathcal{N}(0, \sigma(R)^2 I_d) \overset{\varepsilon,\delta}{\approx} \overline{\mathcal{A}}(\emptyset, \mathcal{A}(R), R) = \overline{\mathcal{A}}_0(\emptyset, \mathcal{A}(R), R) + \mathcal{N}(0, \sigma(R)^2 I_d)$$

with $\sigma(R) \sim \mathrm{RS}^{(m)}_{\overline{\mathcal{A}}_0}(R)$.

Why? 1. Unlearning **and** Retraining can reconstruct R

Result

Let $\varepsilon, \delta > 0$. For all $R \subseteq \mathcal{Z}$, for all $U \subseteq \mathcal{Z}$, $|U| \leq m$, $|R| + |U| = n + m$:

$R \cup U \setminus U = R$

$$\overline{\mathcal{A}}(U, \mathcal{A}(R \cup U), R \cup U) = \overline{\mathcal{A}}_0(U, \mathcal{A}(R \cup U), R \cup U) + \mathcal{N}(0, \sigma(R)^2 I_d)$$

$\overset{\varepsilon,\delta}{\approx}$

$R \setminus \varnothing = R$

$$\overline{\mathcal{A}}(\varnothing, \mathcal{A}(R), R) = \overline{\mathcal{A}}_0(\varnothing, \mathcal{A}(R), R) + \mathcal{N}(0, \sigma(R)^2 I_d)$$

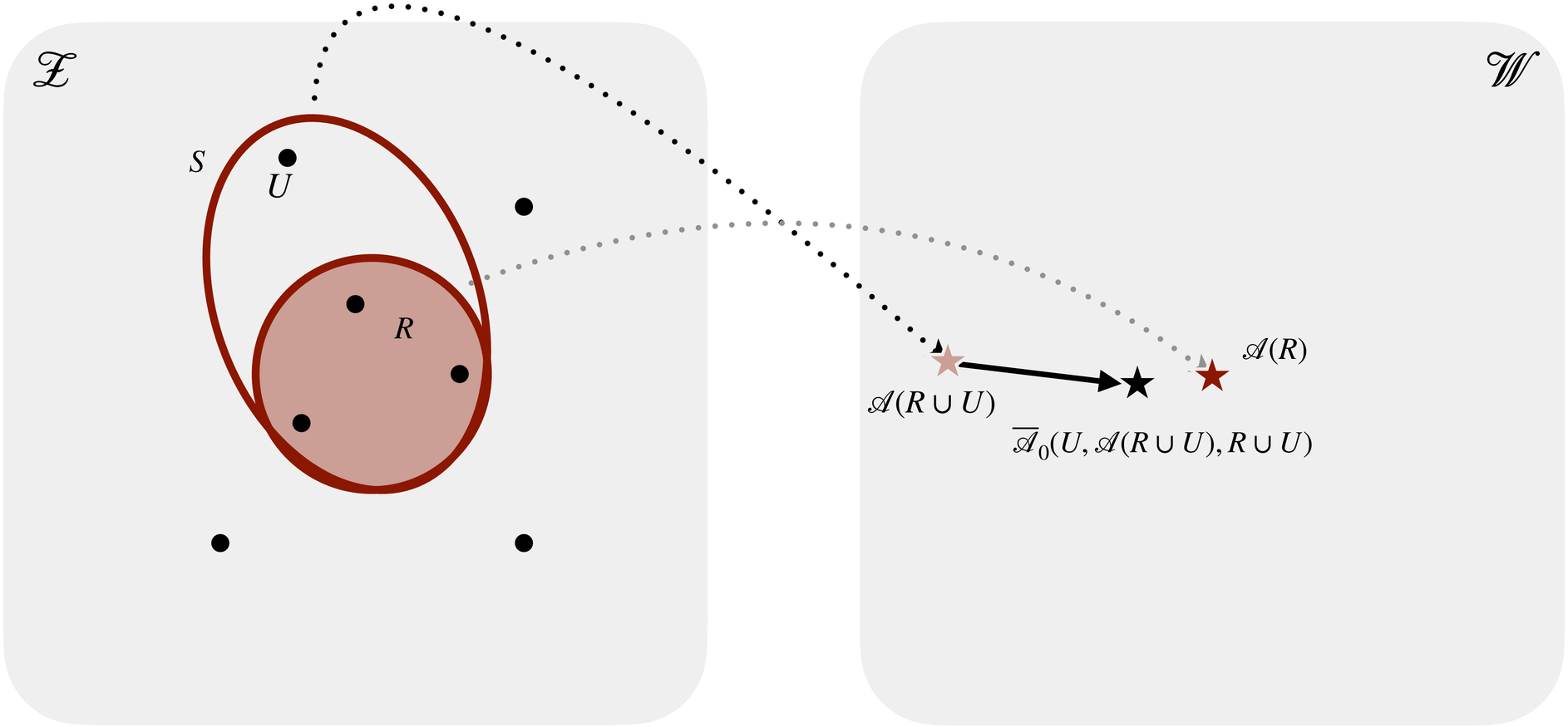
with $\sigma(R) \sim \mathrm{RS}^{(m)}_{\overline{\mathcal{A}}_0}(R)$.

Why? 1. Unlearning **and** Retraining can reconstruct R

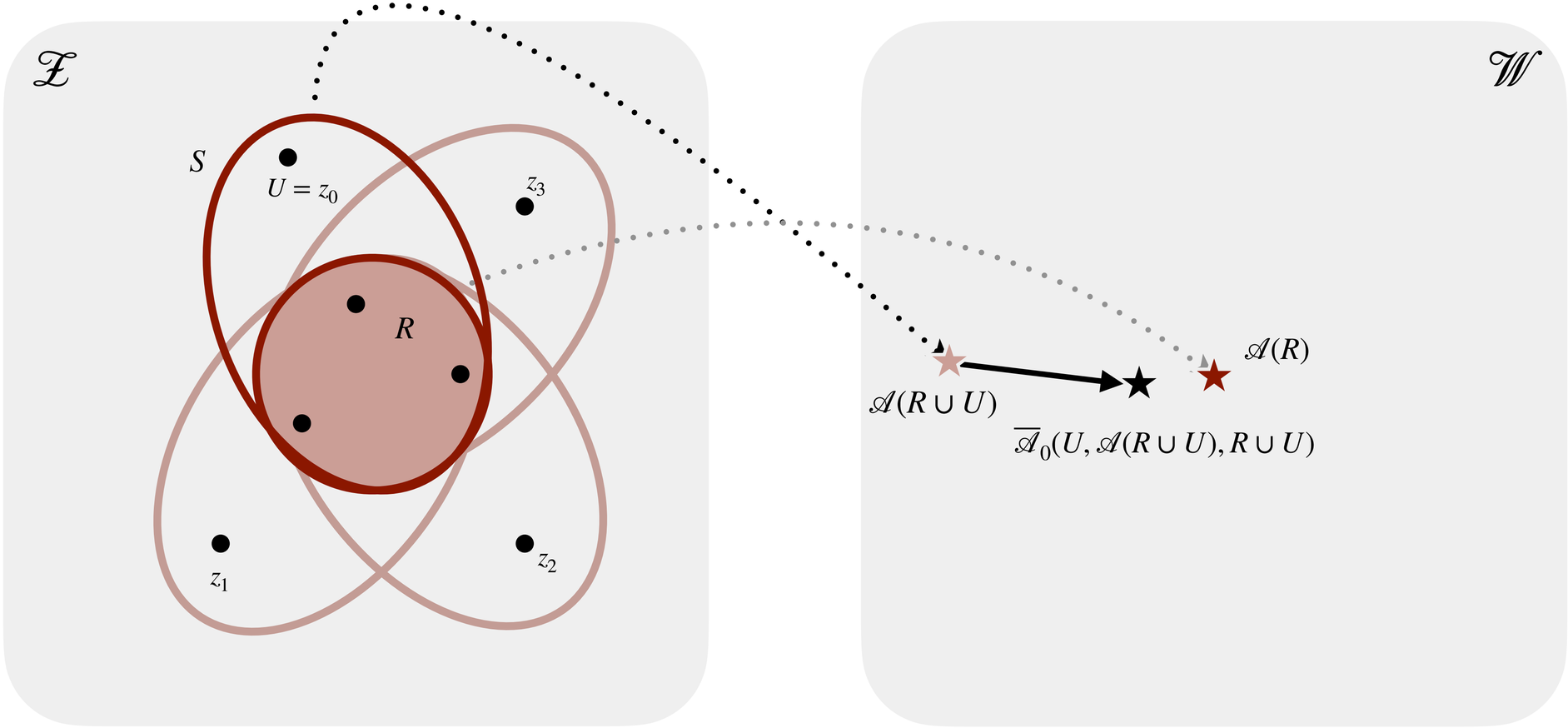
2. **Sufficient* noise** to hide unlearning of any Z from $R \cup Z$ (in particular U from $S = R \cup U$)

*and **necessary** for additive mechanisms with Gaussian noise by mean shift Lemma

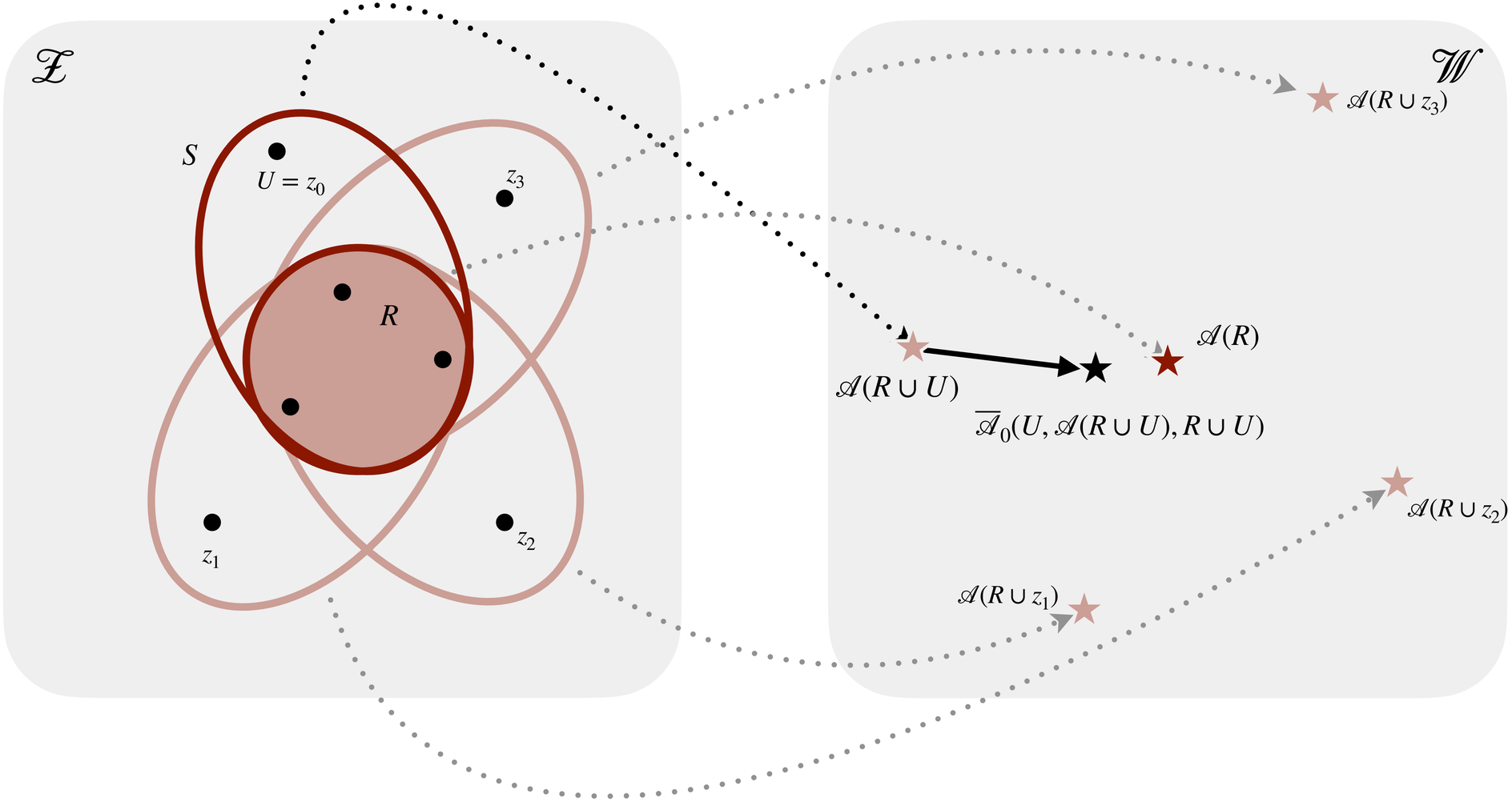
Unlearning


$\mathcal{Z}$
S
U
R
$\mathcal{W}$
$\mathcal{A}(R)$
$\mathcal{A}(R \cup U)$
$\overline{\mathcal{A}}_0(U, \mathcal{A}(R \cup U), R \cup U)$

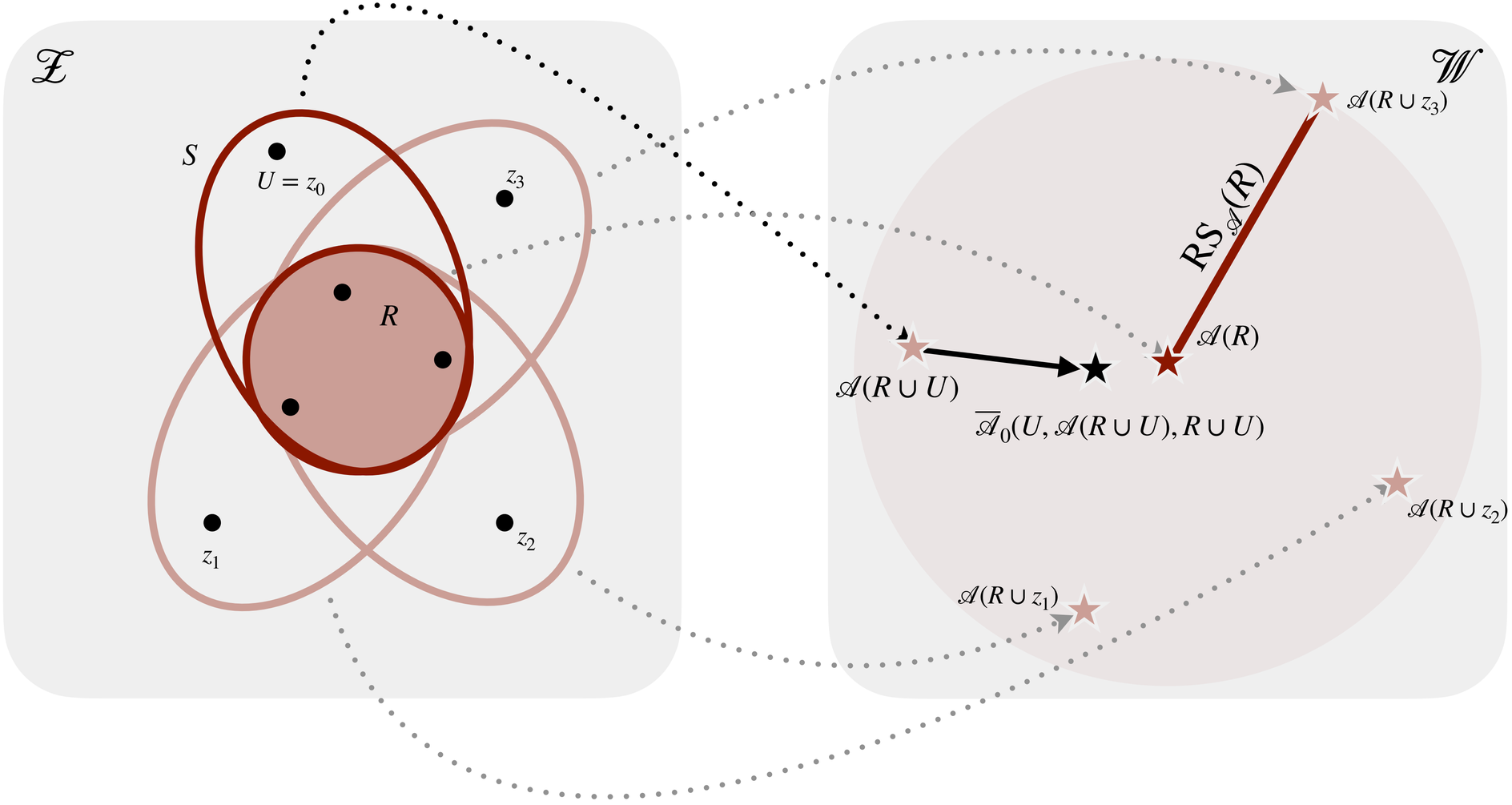
Unlearning


$\mathcal{Z}$
S
$U = z_0$
z_3
R
z_1
z_2
$\mathcal{W}$
$\mathcal{A}(R)$
$\mathcal{A}(R \cup U)$
$\overline{\mathcal{A}}_0(U, \mathcal{A}(R \cup U), R \cup U)$

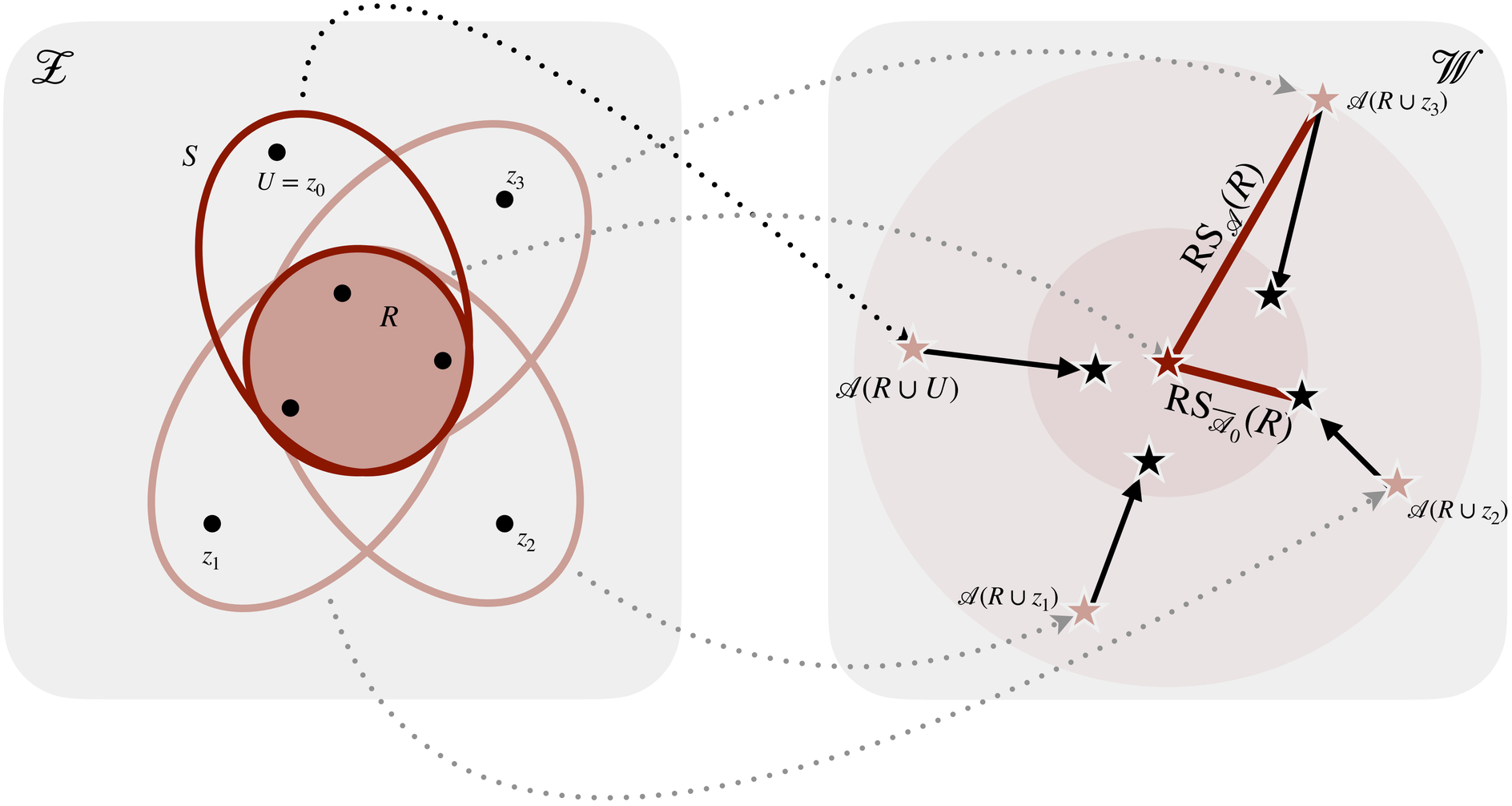
Unlearning


$\mathcal{Z}$
S
$U = z_0$
z_3
R
z_1
z_2
$\mathcal{W}$
$\mathcal{A}(R \cup z_3)$
$\mathcal{A}(R)$
$\mathcal{A}(R \cup U)$
$\overline{\mathcal{A}}_0(U, \mathcal{A}(R \cup U), R \cup U)$
$\mathcal{A}(R \cup z_2)$
$\mathcal{A}(R \cup z_1)$

Unlearning


$\mathcal{Z}$
S
$U = z_0$
z_3
R
z_1
z_2
$\mathcal{W}$
$\mathcal{A}(R \cup z_3)$
$RS_{\mathcal{A}}(R)$
$\mathcal{A}(R)$
$\mathcal{A}(R \cup U)$
$\overline{\mathcal{A}}_0(U, \mathcal{A}(R \cup U), R \cup U)$
$\mathcal{A}(R \cup z_2)$
$\mathcal{A}(R \cup z_1)$

Unlearning


$\mathcal{Z}$
S
$U = z_0$
z_3
R
z_1
z_2
$\mathcal{W}$
$\mathcal{A}(R \cup z_3)$
$RS_{\mathcal{A}}(R)$
$\mathcal{A}(R \cup U)$
$RS_{\overline{\mathcal{A}}_0}(R)$
$\mathcal{A}(R \cup z_2)$
$\mathcal{A}(R \cup z_1)$

Noise ~ Properties of R

$|R| = n,$
$\|x\| \leq B$

Better retain-set 'conditioning' → less noise for same certificate→ higher utility after unlearning

Problem	Global Sensitivity	Retain Sensitivity
Median	$\mathrm{GS}^{(m)}_{\mathrm{median}} = B$	*local spacings* around median(R)
weight of MST	$\mathrm{GS}^{(m)}_{\mathrm{MST}} = mB$	maximum total weight of m *replaceable MST edges* in R
PCA (rank-k projector)	$\mathrm{GS}^{(m)}_{\mathrm{PCA}} \leq \sqrt{2k}$	*eigengap* of covariance of R
SVM	$\mathrm{GS}^{(m)}_{\mathrm{SVM}} = \frac{1}{\gamma}$	*empirical margin* γ_R
ERM	$\mathrm{GS}^{(m)}_{\mathrm{ERM}} \leq \frac{mL}{n\lambda}$	*data-dependent strong convexity* λ_R of $\hat{F}_R$: $\mathrm{RS}^{(m)}_{\mathrm{ERM}}(R) \leq \frac{mL}{n\lambda_R}$

Noise ~ Properties of R: Active Methods

$|R| = n,$
$\|x\| \leq B$

ERM $\quad \mathrm{GS}^{(m)}_{\mathrm{ERM}} \leq \frac{mL}{n\lambda}$ $\quad$ *data-dependent strong convexity* λ_R of $\hat{F}_R$ $\quad \mathrm{RS}^{(m)}_{\mathrm{ERM}}(R) \leq \frac{mL}{n\lambda_R}$

Descent-to-Delete [Neel'20]: (gradient descent on retain data)

$$\overline{\mathscr{A}}(U, w_{R\cup U}, R \cup U) = \mathrm{Proj}_{\mathscr{W}}\left(w_{R\cup U} - \eta_R \nabla \widehat{F}_R(w_{R\cup U})\right)^I + \nu$$

Newton Step Update [Sekhari'21]: (Hessian update)

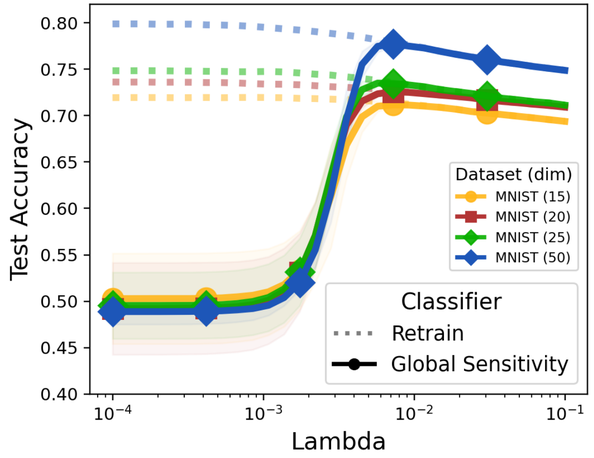
$$\overline{\mathscr{A}}(U, w_{R\cup U}, R \cup U) = w_{R\cup U} + \frac{1}{n}(\nabla^2 \hat{F}_R(w_{R\cup U}))^{-1} \sum_{z\in U} \nabla f(w_{R\cup U}, z)$$

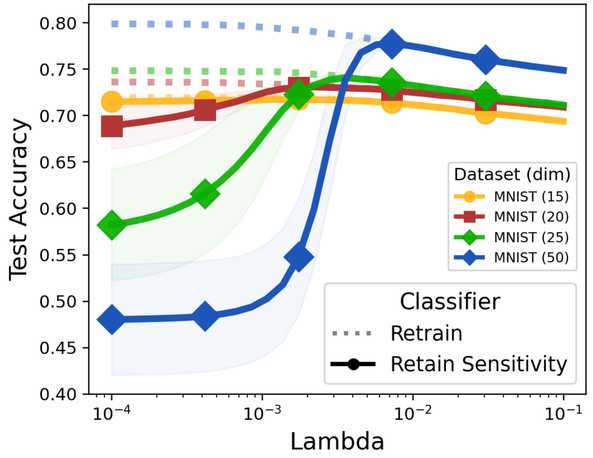
Noise ~ Properties of R: Active Methods

Noise ~ GS

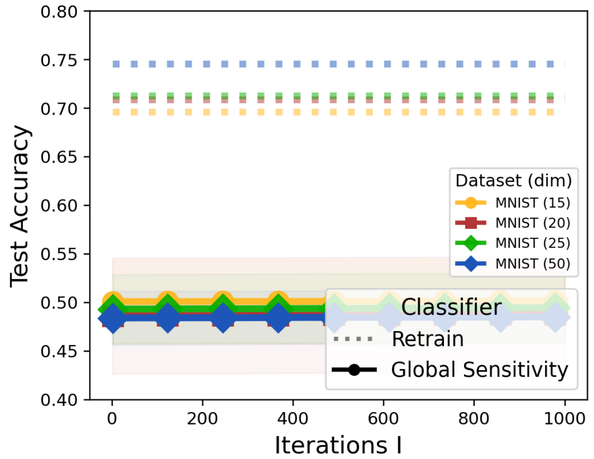
Noise ~ RS

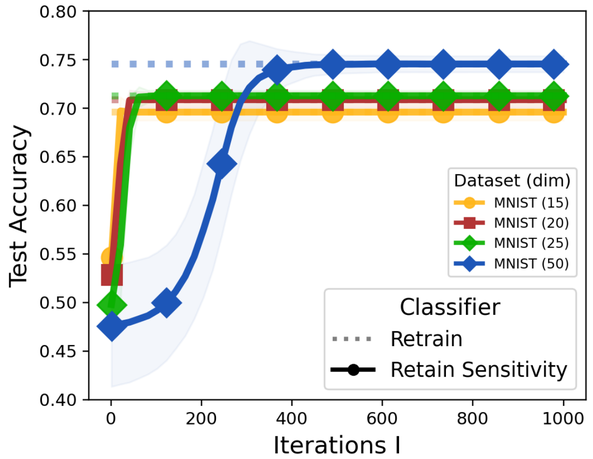
Newton method





D2D





Discussion

Unlearning mechanisms can rely on noise scaling with Retain Sensitivity at R

Consequence: ε-DP-algorithm is ε'-Unlearning with **lower** ε' than by just group privacy

No **privacy requirement for** R in plain unlearning - desirable? See [Cohen'26]

Limitations: **Efficient computation** of necessary statistics at Unlearning/Training time

Assumption of **Full Side Information** to construct RS: sufficient — but necessary?