PRIVACY OR UTILITY FIRST? LEXICOGRAPHIC OPTIMIZATION OF THE PRIVACY–UTILITY BOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Privacy-preserving representation learning seeks to suppress sensitive information while retaining information required for a downstream task. Many existing neural information-theoretic approaches resolve this conflict through externally chosen constraints or weighted objectives, such that the resulting privacy-utility operating point depends on externally chosen hyperparameters rather than on a theoretically characterized privacy-utility bound. We derive a global information-theoretic lower bound on sensitive leakage as a function of retained utility and introduce qualified privacy mechanisms that characterize its equality cases. Furthermore, we show that interaction information constrains the compatibility of privacy and utility: qualified mechanisms can only exist for non-negative interaction information, whereas jointly ideal mechanisms can only exist for non-positive interaction information. In the positive-interaction regime, the extremal equality coordinates of this bound are distinct. We propose privacy-first and utility-first lexicographic optimization (priLO and utiLO) and prove that, whenever these coordinates are attainable by qualified mechanisms, the lexicographic objectives recover them exactly. For neural optimization, we realize these priorities using adaptive gradient surgery, eliminating the need for a fixed trade-off coefficient. Controlled synthetic experiments recover the predicted endpoint structure, while experiments on six real-world datasets show that the proposed lexicographic objectives provide competitive privacy-utility operating points without selecting a fixed trade-off coefficient.

1 INTRODUCTION

Privacy-preserving machine learning seeks to suppress sensitive information from neural representations while retaining information required for downstream tasks, motivated in part by regulatory requirements such as the *General Data Protection Regulation* (European Parliament and Council of the European Union, 2016) and the *California Consumer Privacy Act* (State of California, 2018). In this setting, a neural network computes a representation Z from input X that ideally eliminates marginal sensitive leakage about S while preserving all attainable utility information about U. These goals are non-trivial because S and U may depend on each other (Sankar et al., 2013), and their compatibility depends on how both types of information coexist in X.

Existing information-theoretic privacy approaches address this conflict by balancing sensitive leakage against a representation- or utility-preserving objective. The privacy funnel and its conditional extensions (Makhdoumi et al., 2014; Rodríguez-Gálvez et al., 2021) impose externally selected information thresholds, whereas the information bottleneck (Tishby et al., 1999) and CPFSI (de Freitas & Geiger, 2025) employ Lagrangian weights. These hyperparameters determine the resulting privacy-utility operating point but are not themselves determined by an information-theoretic criterion. Therefore, even when the optimization problem is solved exactly, the selected operating point remains dependent on the externally chosen hyperparameter.

Ideal-privacy and -utility extrema have been studied previously in more restrictive information-theoretic settings (Chakraborty et al., 2013; Johnson et al., 2018). We, in turn, derive a global lower bound on sensitive leakage as a function of retained utility and characterize its equality cases, yielding privacy- and utility-side extremal equality coordinates in the positive-interaction regime. We call privacy mechanisms *qualified* when they suppress sensitive information that is unrelated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to utility while preserving all attainable utility information that is unrelated to the sensitive variable. We show that these mechanisms exactly attain equality in this global lower bound. Interaction information constrains the compatibility of qualification and joint ideality (zero leakage while all attainable utility is retained): qualified mechanisms can only exist for non-negative interaction information, whereas jointly ideal mechanisms can only exist for non-positive interaction information. At zero interaction, both classes coincide; for positive interaction, joint ideality is impossible and the corresponding privacy and utility extremes are distinct. We refer to these as the privacy and utility bound endpoints, respectively. In the positive-interaction regime, these two endpoints naturally induce two strict optimization priorities: a privacy-first mechanism first minimizes sensitive leakage and then maximizes utility among privacy-optimal mechanisms, whereas a utility-first mechanism reverses this order. We formulate both using *lexicographic optimization* (LO) (Fishburn, 1974) and realize the neural optimization with the *gradient surgery* approach of Hsieh et al. (2024), augmented by an adaptive finite-step safeguard for incorporating the secondary objective. Whenever the corresponding qualified endpoints are attainable, we show that privacy-first LO (priLO) and utility-first LO (utiLO) recover them exactly.

The appropriate priority is application dependent. For example, privacy-preserving automatic speech recognition (ASR) may favor complete suppression of speaker identity even at some cost in recognition performance (Becker & Martin, 2025), whereas safety-critical medical classification may prioritize maximal diagnostic information while minimizing patient-information leakage subject to that requirement (Ziller et al., 2024; Tayebi Arasteh et al., 2024; Mohammadi et al., 2026).

Our main contributions are

- **a global privacy-utility lower bound and characterization of its equality cases** through qualified mechanisms,
- **interaction-information conditions for qualification and joint ideality**, separating the positive, zero, and negative interaction regimes and characterizing the qualified bound endpoints for positive interaction, and
- **dual privacy-first and utility-first lexicographic formulations**, which recover the corresponding qualified bound endpoints whenever they are attainable.

Preliminaries. For random variables A and B, the mutual information (MI) is defined in terms of Kullback-Leibler divergence as

$$I(A;B) = \mathrm{D}_{\mathrm{KL}}[\mathbb{P}_{A,B} \| \mathbb{P}_A \otimes \mathbb{P}_B], \tag{1}$$

where $\mathbb{P}_{A,B}$ is the joint probability law and $\mathbb{P}_A$ and $\mathbb{P}_B$ are the respective marginal probability laws. The MI quantifies the amount of information that is shared between A and B and it is zero iff A and B are statistically independent. The conditional mutual information is denoted by $I(A;B \mid C)$ and measures the information shared between A and B after conditioning on C. A generalization is given by the *interaction information* which is also commutative:

$$I(A;B;C) = I(A;B) - I(A;B \mid C), \tag{2}$$

The interaction information can become negative when conditioning on C increases the information shared between A and B. Unless stated otherwise, all reported information quantities are expressed in bits.

2 Related Work

Information Bottleneck. The information bottleneck (IB) was introduced for optimal compression (Tishby et al., 1999) and later formulated variationally for neural networks (Alemi et al., 2017). Related objectives have also been applied to privacy-preserving audio processing (Nelus & Martin, 2021). A Lagrangian form is

$$\min_{p(Z|X)} I(X;Z) - \beta I(U;Z), \tag{3}$$

where β trades compression against utility.

Privacy Funnel and Extensions. The privacy funnel (PF) (Makhdoumi et al., 2014) minimizes sensitive leakage subject to a representation-fidelity constraint:

$$\min_{p(Z|X)} I(S;Z) \text{ s.t. } I(X;Z) \geq \varepsilon_{\mathrm{PF}}. \tag{4}$$

This constraint preserves information in X without specifying utility U. The *conditional privacy funnel* (CPF) (Rodríguez-Gálvez et al., 2021) conditions fidelity on S, while the *conditional fairness bottleneck* (CFB) additionally promotes utility unexplained by S and suppresses information unrelated to both S and U.

Conditional Privacy Funnel with Side Information. CPFSI (de Freitas & Geiger, 2025) combines sensitive leakage, representation fidelity and utility through

$$\min_{p(Z|X)} \left[I(S;Z) - \gamma I(X;Z \mid S) - \beta I(U;Z \mid S) \right], \tag{5}$$

with $\gamma, \beta \geq 0$. Here, $I(X;Z \mid S)$ promotes representation fidelity beyond S, while $I(U;Z \mid S)$ promotes utility unexplained by S.

These approaches rely on externally chosen constraints or Lagrangian weights. Hence, even under exact optimization, their privacy-utility operating points remain dependent on externally chosen hyperparameters.

Privacy-utility extrema. For mutual information disclosure, Chakraborty et al. (2013) study maximum utility under perfect privacy and vice versa. Johnson et al. (2018) extend this to intermediate Pareto-optimal trade-offs using weighted mutual-information scalarization, assuming that useful and sensitive attributes are deterministic functions of the observed data. Sreekumar & Gündüz (2019) consider a more general latent-variable setting and characterize non-zero utility under perfect privacy subject to an additional rate constraint. In contrast, we assume only the Markov structure $(S,U) \to X \to Z$ and derive a global lower bound on sensitive leakage as a function of retained utility, including its equality cases and interaction-information constraints.

3 Privacy-first and Utility-first

We consider a privacy mechanism Π, possibly stochastic, which transforms the observable variable X into the representation Z. Our analysis is based on the following conditional-independence structure.

Assumption 1 (Markov chain for privacy mechanisms)**.** *For every privacy mechanism Π, the random variables satisfy*

$$(S,U) \to X \xrightarrow{\Pi} Z, \tag{6}$$

and hence

$$I(S,U;Z \mid X) = 0. \tag{7}$$

We denote the set of all privacy mechanisms satisfying this property by $\mathcal{P}$. Additionally, we assume that all I-terms appearing in the theoretical results are finite.

Accordingly, all information shared between Z and either S or U must also be contained in X. We refer to the information about S or U contained in X as *attainable information*, since a mechanism operating solely on X cannot access information outside of this region. By the data processing inequality, it holds that

$$I(S;Z) \leq I(S;X), \quad I(U;Z) \leq I(U;X), \tag{8}$$

$$I(S;Z \mid U) \leq I(S;X \mid U), \text{ and } I(U;Z \mid S) \leq I(U;X \mid S). \tag{9}$$

The identity mechanism $Z = X$ attains all four upper bounds simultaneously. This is, however, not desirable, since maximizing $I(S;Z)$ directly contradicts privacy. Instead, a privacy- and utility-preserving mechanism should exclude sensitive information that is unrelated to utility while preserving all attainable utility information that is unrelated to the sensitive variable. We refer to mechanisms satisfying both requirements as *qualified*.

Definition 1 (Qualified mechanisms)**.** *Let the set of all qualified privacy mechanisms be defined as*

$$\mathcal{P}_q := \{\Pi_q \in \mathcal{P} \mid I(S;Z \mid U) = 0 \ \wedge \ I(U;Z \mid S) = I(U;X \mid S)\}. \tag{10}$$

Independently of qualification, we call a mechanism *jointly ideal* when marginal sensitive leakage vanishes while all attainable marginal utility is preserved.

Definition 2 (Ideal mechanisms)**.** *Under Assumption 1, let the set of jointly ideal privacy mechanisms be defined as*

$$\mathcal{P}^* := \{\Pi^* \in \mathcal{P} \mid I(S;Z) = 0 \ \wedge \ I(U;Z) = I(U;X)\}. \tag{11}$$

The classes $\mathcal{P}^*$ and $\mathcal{P}_q$ are defined independently as subsets of $\mathcal{P}$; their relationship is characterized below by the interaction information.

The Markov structure additionally imposes a global lower bound on sensitive leakage for any attained utility level. Qualification precisely characterizes equality in this bound.

Lemma 1 (Global privacy-utility lower bound[1])**.** *Under Assumption 1, every privacy mechanism $\Pi \in \mathcal{P}$ satisfies*

$$I(S;Z) \geq I(U;Z) - I(U;X \mid S). \tag{12}$$

Moreover, equality holds if and only if Π is qualified:

$$\Pi \in \mathcal{P}_q \Leftrightarrow I(S;Z) = I(U;Z) - I(U;X \mid S). \tag{13}$$

Thus, in the $I(U;Z)$-$I(S;Z)$ plane, no mechanism can lie below the line in equation 13, and qualified mechanisms exactly attain equality.

Qualification fixes both conditional objectives at their respective optima. On the marginal privacy-utility plane, the remaining degree of freedom is how much interaction information between S and U persists in Z. In fact, for every qualified mechanism, $I(S;U;Z) = I(S;Z)$ since $I(S;Z \mid U) = 0$, so by equation 13, $I(U;Z) = I(U;X|S) + I(S;U;Z)$. Thus, increasing the retained interaction information increases both marginal utility and leakage. Combining this bound with the data processing inequality yields necessary interaction-information conditions for both qualification and joint ideality.

Theorem 1 (Interaction constraints for qualified and ideal mechanisms[2])**.** *Under Assumption 1, qualification can only exist for non-negative interaction information:*

$$\mathcal{P}_q \neq \emptyset \quad \Rightarrow \quad I(S;U;X) \geq 0. \tag{14}$$

By contrast, joint ideality can only exist for non-positive interaction information:

$$\mathcal{P}^* \neq \emptyset \quad \Rightarrow \quad I(S;U;X) \leq 0. \tag{15}$$

More specifically, every qualified privacy mechanism $\Pi_q \in \mathcal{P}_q$ with $Z = \Pi_q(X)$ satisfies

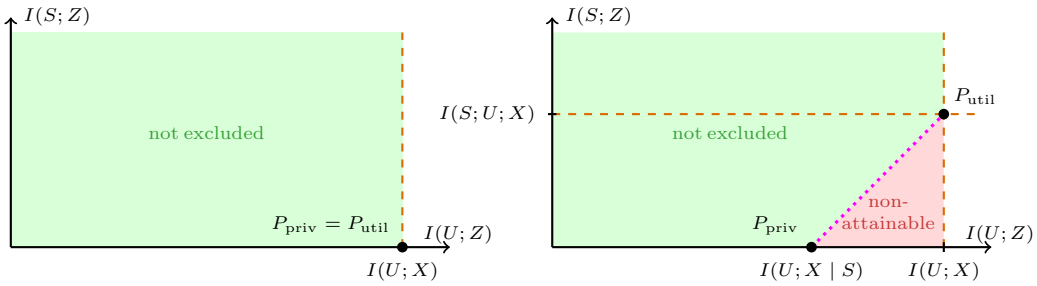
$$0 \leq I(S;Z) \leq I(S;U;X). \tag{16}$$

Consequently, a mechanism can be both qualified and jointly ideal only at zero interaction information:

$$\mathcal{P}^* \cap \mathcal{P}_q \neq \emptyset \quad \Rightarrow \quad I(S;U;X) = 0. \tag{17}$$

These sign conditions are necessary, not sufficient. The theorem separates three regimes. For $I(S;U;X) = 0$, $\mathcal{P}_q = \mathcal{P}^*$. If non-empty, both endpoints coincide at $(I(U;X), 0)$.[3] For $I(S;U;X) < 0$, no qualified mechanism exists, while joint ideality is not excluded.[4] We therefore focus on the non-trivial positive-interaction regime.

[1] Proof provided in Appendix A.1.
[2] Proof provided in Appendix A.2.
[3] Proof provided in Appendix A.3.
[4] Proof provided in Appendix A.4.



(a) Zero interaction information

(b) Positive interaction information

Figure 1: Privacy-utility planes for the zero- and positive-interaction regimes. At zero interaction, the privacy and utility endpoint coordinates coincide at $P_{\text{priv}} = P_{\text{util}} = (I(U;X), 0)$, whenever attained. For positive interaction, the qualified equality segment is bounded by $P_{\text{priv}} = (I(U;X \mid S), 0)$ and $P_{\text{util}} = (I(U;X), I(S;U;X))$. If both endpoints are attainable, Proposition 2 shows that the entire segment is attainable by qualified mechanisms.

Corollary 1 (Positive interaction[5]). *For $I(S;U;X) > 0$, no jointly ideal mechanism exists:*

$$\mathcal{P}^* = \emptyset. \tag{18}$$

Furthermore, for every $\Pi_q \in \mathcal{P}_q$ with $Z = \Pi_q(X)$ there exists

$$t = I(S;Z) \in [0, I(S;U;X)] \tag{19}$$

such that

$$\big(I(U;Z), I(S;Z)\big) = \big(I(U;X \mid S) + t, t\big). \tag{20}$$

Hence, every qualified mechanism lies on the line segment bounded by the privacy endpoint

$$P_{\text{priv}} = \big(I(U;X \mid S), 0\big) \tag{21}$$

and the utility endpoint

$$P_{\text{util}} = \big(I(U;X), I(S;U;X)\big). \tag{22}$$

If attained, P_{priv} has ideal privacy but not ideal utility, whereas P_{util} has ideal utility but not ideal privacy. These coordinates characterize the extremal equality points; this corollary does not by itself guarantee that either endpoint is attained by a qualified mechanism[6].

A visualization of the privacy-utility plane for zero interaction information and positive interaction information is given in Figure 1.

3.1 Lexicographic Optimization

The global lower bound in equation 12 directly exposes two natural priority extremes. At ideal privacy and ideal utility, respectively, it implies

$$I(S;Z) = 0 \Rightarrow I(U;Z) \leq I(U;X \mid S), \tag{23}$$

$$I(U;Z) = I(U;X) \Rightarrow I(S;Z) \geq I(S;U;X). \tag{24}$$

This motivates two strict priority orderings: maximizing utility among privacy-optimal mechanisms (priLO), and minimizing sensitive leakage among utility-optimal mechanisms (utiLO). We encode these priorities using *lexicographic optimization* (LO) (Fishburn, 1974; Isermann, 1982).

For two objectives $g(\zeta)$ and $f(\zeta)$, where f is primary, g is secondary and ζ denotes the optimization variable, lexicographic optimization first determines the optimum of f and subsequently optimizes g within the set of f-optimal solutions. It can be written as

$$\operatorname{lex} \min_{\zeta} \left[f(\zeta), g(\zeta)\right]. \tag{25}$$

[5]Proof provided in Appendix A.5.

[6]A counterexample with positive interaction information but an unattainable privacy endpoint is provided in Appendix A.6

Unlike a weighted sum, this formulation imposes a strict priority: an improvement of the secondary objective cannot compensate for a deterioration of the primary objective.

Applying LO to privacy-first and utility-first gives

$$\text{privacy-first LO (priLO):} \qquad \operatorname{lex}\min_{\Pi\in\mathcal{P}} \left[I(S;Z), -I(U;Z)\right] \tag{26}$$

$$\text{utility-first LO (utiLO):} \qquad \operatorname{lex}\min_{\Pi\in\mathcal{P}} \left[-I(U;Z), I(S;Z)\right]. \tag{27}$$

These lexicographic priority orderings are defined independently of whether the corresponding equality coordinates of the bound are attainable. When the qualified bound endpoints are attainable, the following proposition shows that the corresponding lexicographic optima are attained and coincide with these endpoints.

Proposition 1 (Lexicographic recovery of attainable bound endpoints[7]). *Let $I(S;U;X) > 0$ and assume qualified mechanisms attain both bound endpoints P_{priv} and P_{util} from Corollary 1. Then priLO and utiLO admit solutions. Let Π_{priv} and Π_{util} be arbitrary such solutions, with $Z_{\text{priv}} = \Pi_{\text{priv}}(X)$ and $Z_{\text{util}} = \Pi_{\text{util}}(X)$. Then,*

$$\big(I(U;Z_{\text{priv}}), I(S;Z_{\text{priv}})\big) = P_{\text{priv}}, \tag{28}$$

$$\big(I(U;Z_{\text{util}}), I(S;Z_{\text{util}})\big) = P_{\text{util}}. \tag{29}$$

In particular, every such lexicographic solution is qualified.

Weighted mutual-information scalarizations have previously been used to recover Pareto-optimal privacy-utility trade-offs (Johnson et al., 2018). Without assuming attainability of the analytical bound endpoints, recovering a prescribed lexicographic extreme by scalarization requires choosing the relative weight according to the geometry of the attainable privacy-utility region. Since this region is generally unknown a priori, the required scalarization weight cannot be determined from the desired priority alone.[8]

Gradient surgery. The definitions above specify the desired lexicographic solutions but not how they are obtained numerically. In neural network training, the optimum value f^* and the corresponding optimal set are generally unavailable explicitly (Gong et al., 2021; Hsieh et al., 2024). We therefore adopt the gradient-projection geometry of Hsieh et al. (2024) and augment it with an adaptive finite-step safeguard. Let f denote the primary objective and g the secondary objective, and define

$$\boldsymbol{a}_t = \nabla f(\boldsymbol{\theta}_t),\ \boldsymbol{b}_t = \nabla g(\boldsymbol{\theta}_t) - \frac{\nabla g(\boldsymbol{\theta}_t)^\top \boldsymbol{m}_t}{\|\boldsymbol{m}_t\|_2^2 + \varepsilon}\boldsymbol{m}_t, \tag{30}$$

where $\boldsymbol{m}_t$ is an exponential moving average of preceding primary gradients. The candidate update direction is

$$\boldsymbol{d}_t(\lambda_t) = \boldsymbol{a}_t + \lambda_t \boldsymbol{b}_t, \qquad 0 \le \lambda_t \le 1. \tag{31}$$

Rather than treating λ_t as a fixed trade-off coefficient, we determine it independently at every optimization step. Starting from $\lambda_t = 1$, we evaluate geometrically decreasing candidates using the actual finite SGD step. A positive candidate is admissible only if

$$f(\boldsymbol{\theta}_{t+1}(\lambda_t)) \le f(\boldsymbol{\theta}_t),\ g(\boldsymbol{\theta}_{t+1}(\lambda_t)) \le g(\boldsymbol{\theta}_t). \tag{32}$$

Among admissible candidates, we select the one yielding the smallest secondary objective. If no positive candidate is admissible, we set $\lambda_t = 0$ and apply the pure-primary update. Thus, λ_t is a step-local optimization quantity rather than a privacy-utility trade-off hyperparameter. For priLO and utiLO, f and g are assigned according to their respective priority ordering above.

Mutual information estimation. Our proposed methods priLO and utiLO require either the minimization or maximization of marginal or conditional MI. We use the Barber-Agakov (BA) lower bound (Barber & Agakov, 2003) for MI-maximization and the variational Contrastive Log-ratio Upper Bound (vCLUB) (Cheng et al., 2020) for MI-minimization.

[7] Proof provided in Appendix A.7.

[8] A formal derivation of this geometry dependence is provided in Appendix A.8.

Table 1: Overview of datasets (left) and benchmark hyperparameters (right). U and S denote the utility and sensitive variables, $|\mathcal{U}|$ and $|\mathcal{S}|$ their respective numbers of classes, N_{obs} the number of observations after preprocessing, and the final column the shape of the input X. On the right, the power of 2 indicates a Cartesian product with itself.

Dataset	U	$\lvert\mathcal{U}\rvert$	S	$\lvert\mathcal{S}\rvert$	N_{obs}	Input
Adult	Inc. $>$ 50k	2	Sex	2	45,222	102
COMPAS	Recidivism	2	Ethn.	2	6,172	19
Dutch	Occupation	2	Sex	2	60,420	48
Credit	Default	2	Sex	2	30,000	31
HAR	Activity	6	Subject	30	9,112	9×128
AudioMNIST	Digit	10	Speaker	60	30,000	64×98

Method	Param.	Values
IB	β	$\{0.1, 1, 10\}$
PF	β	$\{0.1, 1, 10\}$
CPF	γ	$\{1, 64, 1024\}$
CFB	β	$\{1, 64, 1024\}$
CPFSI	β, γ	$\{1, 64, 1024\}^2$

For a discrete variable $Y \in \{S, U\}$ and an optional conditioning variable C, the conditional MI can be expressed as

$$\begin{aligned} I(Y; Z \mid C) &= H(Y \mid C) - H(Y \mid Z, C) \\ &= H(Y \mid C) + \mathbb{E}_{p(Y,Z,C)} \left[\log p(Y \mid Z, C)\right]. \end{aligned} \tag{33}$$

Replacing the generally unknown posterior $p(Y \mid Z, C)$ by a variational classifier $q_\psi(Y \mid Z, C)$ yields the BA lower bound

$$I(Y; Z \mid C) \geq H(Y \mid C) + \mathbb{E}_{p(Y,Z,C)} \left[\log q_\psi(Y \mid Z, C)\right] =: I_{\text{BA}}(Y; Z \mid C). \tag{34}$$

Hence, maximizing I_{BA} provides a surrogate for preserving information.

For MI minimization, CLUB instead contrasts the conditional log-likelihood of samples from the joint distribution with samples drawn independently conditional on C. Its variational form is

$$I_{\text{vCLUB}}(Y; Z \mid C) = \mathbb{E}_{p(Y,Z,C)} \left[\log q_\phi(Y \mid Z, C)\right] - \mathbb{E}_{p(C)p(Z|C)p(Y|C)} \left[\log q_\phi(Y \mid Z, C)\right]. \tag{35}$$

For $C = \varnothing$, both expressions reduce to their corresponding marginal-MI estimators. CLUB is an upper bound when the true conditional distribution is available. With the variational approximation q_ϕ, (Cheng et al., 2020) show that vCLUB remains an upper bound under an appropriate approximation condition and otherwise approaches the true MI as the variational conditional becomes sufficiently accurate. The variational distributions q_ψ and q_ϕ are implemented as classifiers and fitted by minimizing the cross-entropy on samples from $p(Y, Z, C)$, equivalently by maximizing their conditional log-likelihood. We use vCLUB exclusively for MI-minimization objectives and BA exclusively for MI-maximization objectives.

4 Experiments

Endpoint recovery. We first test whether the neural realization of priLO and utiLO approaches analytically known privacy-utility endpoints (cf. Figure 1). We use finite-support binary problems in two interaction regimes. At zero interaction, the theoretical privacy and utility endpoints coincide at $P_{\text{priv}} = P_{\text{util}} = (1, 0)$. With one bit of positive interaction, they separate to $P_{\text{priv}} = (1, 0)$ and $P_{\text{util}} = (2, 1)$. For each regime and priority ordering, we train the neural mechanism for 15 random seeds using the marginal BA/vCLUB objectives and adaptive gradient surgery, and evaluate the resulting representation with an independent estimator that does not reuse the training MI critics. Figure 2 shows the individual runs and their means. In the zero-interaction regime, both methods cluster close to the common endpoint. With one bit of interaction, priLO reaches $(1.001, 0.002)$ on average, while utiLO reaches $(2.000, 1.083)$: full attainable utility is retained, with a small residual leakage gap to the theoretical utility endpoint. A separate exact finite-support construction and exhaustive deterministic-mechanism enumeration, which remove neural optimization and MI-estimation error entirely, are provided in Appendix A.10 and Figure 5.

Dimensional scaling. We next isolate the effect of observation dimensionality using a synthetic Gaussian problem whose informative privacy-utility structure and analytical endpoints remain fixed while independent nuisance dimensions are added.[9] We vary the input dimensionality from $d = 2$ to $d = 1024$ and apply a fixed orthogonal rotation such that the informative coordinates are not

[9]The complete construction and experimental protocol are provided in Appendix A.11.

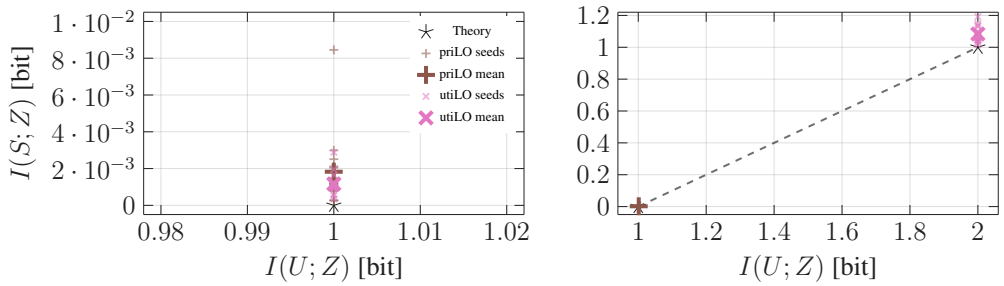


Figure 2: Neural finite-support endpoint recovery over 15 random seeds for zero interaction information (left) and one-bit interaction information (right). The dashed line denotes the qualified-bound segment between the analytical privacy and utility endpoints.

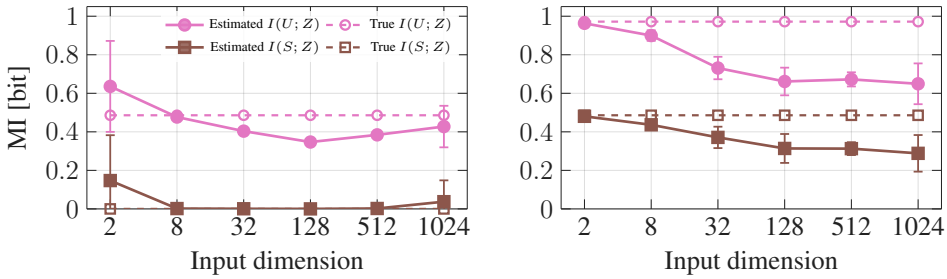


Figure 3: Dimensional scaling for priLO (left) and utiLO (right) over 10 random seeds. Solid curves show mean held-out MI estimates, error bars denote one standard deviation, and dashed curves show the analytical endpoints.

axis-aligned. Figure 3 shows that priLO maintains approximately zero leakage for most dimensions, but generally under-recovers the attainable utility. For utiLO, the deviation from the utility endpoint becomes substantial at larger dimensions, with increased variance at $d = 1024$. Hence, increasing dimensionality can impair the practical recovery of the endpoints even though their analytical coordinates remain unchanged.

Real-data benchmark. We evaluate on Adult (Becker & Kohavi, 1996), COMPAS (Larson et al., 2016), Dutch Census (van der Laan, 2000), Credit (Yeh, 2016; Yeh & Lien, 2009), HAR (Reyes-Ortiz et al., 2013), and AudioMNIST (Becker et al., 2024). Dataset details, hyperparameters, and protocols are summarized in Table 1 and Appendices A.12–A.13.

Figure 4 shows the resulting utility-privacy operating points over 15 random seeds, evaluated post hoc using held-out BA lower bounds with train-fitted, validation-selected posteriors independent of the critics used during mechanism optimization. PF and CPF generally sacrifice substantial utility for limited privacy gains. IB performs particularly well on HAR and AudioMNIST but collapses or becomes less competitive on several tabular datasets, whereas CPFSI exhibits an almost complementary pattern. CFB is the strongest conventional baseline overall, but depends strongly on its trade-off coefficient and several settings collapse to near-zero utility. In contrast, at least one of priLO or utiLO yields a competitive endpoint on every dataset without selecting such a coefficient, illustrating the dataset dependence of scalarized objectives.

Operational utility and privacy attacks. To complement the MI evaluation, we freeze each representation and train matched post-hoc classifiers for utility $Z \to U$ and sensitive-attribute recovery $Z \to S$. We report macro-F_1 for $Z = X$, priLO, utiLO, CPFSI-P/U selected from the MI operating points independently of the post-hoc results, and CFB with $\beta = 1024$. Complete results and protocols are provided in Appendix A.14 and Appendix A.13. The post-hoc results mirror the information-theoretic geometry in Figure 4: HAR and AudioMNIST exhibit little apparent privacy-utility conflict, with several methods attaining high utility and weak sensitive-attribute predictability, whereas datasets with a more pronounced conflict show a clearer separation between privacy- and utility-oriented solutions. In particular, utiLO retains near-maximal utility on HAR and AudioMNIST while reducing sensitive-attribute prediction to approximately chance level.

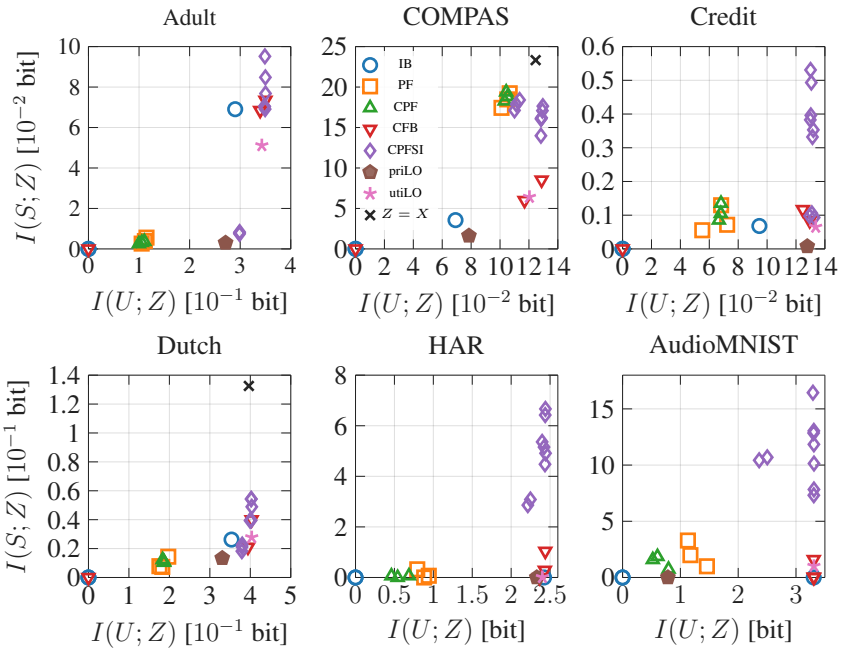


Figure 4: Benchmark results over 15 seeds. $Z = X$ is omitted where vertically distant from the learned methods.

Table 2: Post-hoc macro-F_1 (in %) for utility (F_1^U) and leakage (F_1^S) on frozen representations.

	Utility $F_1^U \uparrow$						Attacker $F_1^S \downarrow$					
Dataset	$Z = X$	priLO	utiLO	CPFSI-P	CPFSI-U	CFB	$Z = X$	priLO	utiLO	CPFSI-P	CPFSI-U	CFB
Adult	78.7	60.2	78.6	76.1	**78.9**	78.8	83.0	**40.3**	52.9	41.2	53.4	52.8
COMPAS	68.1	49.0	67.0	**67.8**	67.6	67.6	69.1	**48.8**	61.1	62.7	65.6	62.9
Dutch	81.8	77.0	**81.6**	80.7	81.5	81.5	67.7	**51.1**	60.0	56.9	60.3	59.4
Credit	67.8	65.5	**68.1**	67.8	68.0	**68.1**	56.1	**38.5**	39.0	39.2	42.6	38.7
HAR	93.3	94.4	**98.6**	89.7	96.8	96.8	36.0	2.8	**2.0**	4.8	10.9	4.8
AudioMNIST	99.6	31.6	**99.9**	99.8	**99.9**	99.8	98.0	1.9	**1.7**	8.6	11.1	2.1

Limitations. Mutual information measures average dependence and provides neither differential-privacy nor worst-case disclosure guarantees. Moreover, priLO and utiLO optimize BA/vCLUB surrogates, so finite-sample estimation, variational approximation, and optimization error can prevent exact endpoint recovery. On AudioMNIST, priLO attains near-zero leakage but under-recovers utility, consistent with secondary-gradient starvation. We interpret this as a limitation of the neural realization rather than evidence against the lexicographic formulation.

5 Conclusions

We derived a global privacy-utility lower bound, characterized its equality cases, and showed how interaction information governs the compatibility of privacy and utility. The controlled endpoint-recovery experiments recover the predicted lexicographic endpoints, while the dimensional-scaling experiment exposes practical limitations of their neural recovery. Positive interaction separates the privacy and utility bound endpoints, whereas near-zero interaction can yield almost joint privacy and utility. Non-positive interaction is a necessary condition for exact joint ideality. The real-data benchmarks show that different conventional methods are competitive in different settings, while at least one of priLO or utiLO remains competitive across all investigated datasets. This is particularly relevant when the underlying privacy-utility structure is unknown, since scalarized methods require both a suitable objective formulation and a trade-off coefficient. Lexicographic optimization instead targets the desired endpoint directly through priority ordering, providing a principled alternative when little prior domain knowledge is available. In the future, we will generalize our method using ensemble learning such that the operating points between the lexicographic extrema can be controlled.

AI USE STATEMENT

We used generative AI tools for language editing, literature discovery, adversarial checking of scientific claims, methodological criticism, and assistance with mathematical notation and derivations. We also used AI coding agents to assist with the implementation of experimental code. The central research ideas, narrative and theorem structure were developed exclusively by the authors. All AI-assisted mathematical and methodological content was reviewed by the authors and, where applicable, verified against the corresponding references. Both authors also reviewed and verified the AI-assisted experimental code.

ETHICS STATEMENT

This work studies methods for reducing sensitive-information leakage from learned representations. Our experiments use existing public benchmark datasets and do not involve the collection of new human-subject data. The proposed information-theoretic privacy criterion measures average statistical dependence and should not be interpreted as a worst-case or differential-privacy guarantee. Consequently, deployment in privacy-sensitive applications would require additional analysis appropriate to the corresponding threat model. The methods may also be used to quantify sensitive information contained in representations, and should therefore be applied with appropriate safeguards.

REPRODUCIBILITY STATEMENT

We anonymously provide the source code and experiment configurations used to generate the results in this work. The code contains the implementations of the considered baseline methods, priLO and utiLO, as well as the synthetic experiments and real-data benchmarks. It includes dataset preparation, training and evaluation scripts, the random seeds and configurations used in the reported experiments, and scripts for reproducing the figures. The accompanying README describes environment setup and the commands required to reproduce the experiments. In the camera-ready paper, a link to a git-repository will be provided. Detailed dataset and experimental protocols, as well as the mathematical proofs, are provided in the Appendix.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

David Barber and Felix V. Agakov. Information maximization in noisy channels: A variational approach. In *Advances in Neural Information Processing Systems*, volume 16, pp. 201–208, 2003.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL `https://doi.org/10.24432/C5XW20`.

Luca Becker and Rainer Martin. Contrastive representation learning for privacy-preserving finetuning of audio-visual speech recognition. In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2025. doi: 10.1109/WASPAA66052.2025.11230949.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, January 2024. ISSN 0016-0032. doi: 10.1016/j.jfranklin.2023.11.038.

Supriyo Chakraborty, Nicolas Bitouzé, Mani Srivastava, and Lara Dolecek. Protecting data against unwanted inferences. In *2013 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2013. doi: 10.1109/ITW.2013.6691224.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *Proceedings of the 37th International*

Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pp. 1779–1788. PMLR, 2020.

João Machado de Freitas and Bernhard C. Geiger. Trustworthy representation learning via information funnels and bottlenecks. *Machine Learning*, 114(12):267, 2025. doi: 10.1007/s10994-025-06924-9.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, pp. 1–88, 2016.

Peter C. Fishburn. Lexicographic orders, utilities and decision rules: A survey. *Management Science*, 20(11):1442–1471, 1974. doi: 10.1287/mnsc.20.11.1442.

Chengyue Gong, Xingchao Liu, and Qiang Liu. Automatic and harmless regularization with constrained and lexicographic optimization: A dynamic barrier approach. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29630–29642, 2021.

Yu-Guan Hsieh, James Thornton, Eugene Ndiaye, Michal Klein, Marco Cuturi, and Pierre Ablin. Careful with that scalpel: Improving gradient surgery with an EMA. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19085–19100. PMLR, 2024. URL https://proceedings.mlr.press/v235/hsieh24a.html.

Heinz Isermann. Linear lexicographic optimization. *Operations-Research-Spektrum*, 4(4):223–228, 1982. doi: 10.1007/BF01782758.

Matthew P. Johnson, Liang Zhao, and Supriyo Chakraborty. Achieving Pareto-optimal MI-based privacy-utility tradeoffs under full data. *IEEE Journal of Selected Topics in Signal Processing*, 12 (5):1093–1105, 2018. doi: 10.1109/JSTSP.2018.2850021.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the COMPAS recidivism algorithm. ProPublica, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In *2014 IEEE Information Theory Workshop (ITW)*, pp. 501–505, 2014. doi: 10.1109/ITW.2014.6970882.

Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, IoTDI '19, pp. 49–58, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6283-2. doi: 10.1145/3302505.3310068.

Marziyeh Mohammadi, Mohsen Vejdanihemmat, Mahshad Lotfinia, Mirabela Rusu, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. Differential privacy for medical deep learning: Methods, tradeoffs, and deployment implications. *npj Digital Medicine*, 9(1):93, 2026. doi: 10.1038/s41746-025-02280-z.

Alexandru Nelus and Rainer Martin. Privacy-preserving audio classification using variational information feature extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2864–2877, 2021. doi: 10.1109/TASLP.2021.3108063.

Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2013. URL https://doi.org/10.24432/C54S4K.

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. A variational approach to privacy and fairness. In *2021 IEEE Information Theory Workshop (ITW)*, pp. 1–6, 2021. doi: 10.1109/ITW48936.2021.9611429.

Lalitha Sankar, S. Raj Rajagopalan, and H. Vincent Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6): 838–852, 2013. doi: 10.1109/TIFS.2013.2253320.

Sreejith Sreekumar and Deniz Gündüz. Optimal privacy-utility trade-off under a rate constraint. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2159–2163, 2019. doi: 10.1109/ISIT.2019.8849330.

State of California. California Consumer Privacy Act of 2018. Assembly Bill No. 375, Chapter 55, Statutes of 2018, 2018. Codified in California Civil Code, Title 1.81.5.

Soroosh Tayebi Arasteh, Alexander Ziller, Christiane Kuhl, Marcus Makowski, Sven Nebelung, Rickmer Braren, Daniel Rueckert, Daniel Truhn, and Georgios Kaissis. Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging. *Communications Medicine*, 4(1):46, 2024. doi: 10.1038/s43856-024-00462-6.

Chao Tian, Jun Chen, and Suhas N. Diggavi. Multiuser successive refinement and multiple description coding. *IEEE Transactions on Information Theory*, 54(2):921–931, February 2008. doi: 10.1109/TIT.2007.913435.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999.

Paul van der Laan. The 2001 census in the Netherlands: Integration of registers and surveys, 2000. Paper prepared for the INSEE-EUROSTAT Seminar on Censuses After 2001, Paris, France, 20–21 November 2000.

I-Cheng Yeh. Default of credit card clients. UCI Machine Learning Repository, 2016. URL `https://doi.org/10.24432/C55S3H`.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2): 2473–2480, 2009. doi: 10.1016/j.eswa.2007.12.020.

Alexander Ziller, Tamara T. Mueller, Simon Stieger, Leonhard F. Feiner, Johannes Brandt, Rickmer Braren, Daniel Rueckert, and Georgios Kaissis. Reconciling privacy and accuracy in AI for medical imaging. *Nature Machine Intelligence*, 6(7):764–774, 2024. doi: 10.1038/s42256-024-00858-y.

A Appendix

A.1 Global privacy-utility lower bound

Under Assumption 1, the chain rule of mutual information gives

$$I(U;Z) - I(S;Z) = I(U;Z \mid S) - I(S;Z \mid U) \tag{36}$$
$$\leq I(U;Z \mid S) \tag{37}$$
$$\leq I(U;X \mid S), \tag{38}$$

where the first inequality follows from the non-negativity of conditional mutual information and the second from the conditional data processing inequality. Rearranging gives

$$I(S;Z) \geq I(U;Z) - I(U;X \mid S). \tag{39}$$

Equality in the complete chain holds iff

$$I(S;Z \mid U) = 0, \tag{40}$$
$$I(U;Z \mid S) = I(U;X \mid S), \tag{41}$$

which are exactly the two conditions in Definition 1. Hence equality holds iff $\Pi \in \mathcal{P}_q$. For an ideal mechanism $\Pi^* \in \mathcal{P}^*$, Definition 2 gives

$$I(S;Z) = 0 \quad \text{and} \quad I(U;Z) = I(U;X). \tag{42}$$

For such an ideal mechanism, equality in the bound holds iff

$$0 = I(U;X) - I(U;X \mid S) = I(S;U;X). \tag{43}$$

A.2 Interaction constraints for qualified and ideal mechanisms

For any qualified privacy mechanism, equality in Lemma 1 gives

$$I(S;Z) = I(U;Z) - I(U;X \mid S). \tag{44}$$

By Assumption 1, the data processing inequality yields

$$I(U;Z) \le I(U;X), \tag{45}$$

and therefore

$$I(S;Z) \le I(U;X) - I(U;X \mid S) \tag{46}$$
$$= I(S;U;X). \tag{47}$$

Since mutual information is non-negative,

$$0 \le I(S;Z) \le I(S;U;X), \tag{48}$$

which can only hold if

$$I(S;U;X) \ge 0. \tag{49}$$

If $\mathcal{P}^* \neq \emptyset$, let $\Pi^* \in \mathcal{P}^*$ and $Z = \Pi^*(X)$. By Definition 2,

$$I(S;Z) = 0 \quad \text{and} \quad I(U;Z) = I(U;X). \tag{50}$$

Applying Lemma 1 gives

$$0 = I(S;Z) \tag{51}$$
$$\ge I(U;Z) - I(U;X \mid S) \tag{52}$$
$$= I(U;X) - I(U;X \mid S) \tag{53}$$
$$= I(S;U;X), \tag{54}$$

so $I(S;U;X) \le 0$. Finally, if $\mathcal{P}^* \cap \mathcal{P}_q \neq \emptyset$, the qualification result gives $I(S;U;X) \ge 0$ while the ideality result gives $I(S;U;X) \le 0$. Hence,

$$I(S;U;X) = 0. \tag{55}$$

A.3 Zero interaction

First let $\Pi_q \in \mathcal{P}_q$ and $Z = \Pi_q(X)$. For $I(S;U;X) = 0$, Theorem 1 yields

$$0 \le I(S;Z) \le 0, \tag{56}$$

and hence $I(S;Z) = 0$. Equality in Lemma 1 then gives

$$I(U;Z) = I(U;X \mid S). \tag{57}$$

Since

$$I(S;U;X) = I(U;X) - I(U;X \mid S) = 0, \tag{58}$$

it follows that $I(U;X \mid S) = I(U;X)$ and therefore $I(U;Z) = I(U;X)$. Hence every $\Pi_q \in \mathcal{P}_q$ satisfies Definition 2, implying

$$\mathcal{P}_q \subseteq \mathcal{P}^*. \tag{59}$$

Conversely, let $\Pi^* \in \mathcal{P}^*$ and $Z = \Pi^*(X)$. By Definition 2 and $I(S;U;X) = 0$,

$$I(S;Z) = 0 \tag{60}$$
$$= I(U;X) - I(U;X \mid S) \tag{61}$$
$$= I(U;Z) - I(U;X \mid S). \tag{62}$$

Thus equality holds in Lemma 1, which implies $\Pi^* \in \mathcal{P}_q$ and therefore

$$\mathcal{P}^* \subseteq \mathcal{P}_q. \tag{63}$$

Combining both inclusions gives

$$\mathcal{P}_q = \mathcal{P}^*. \tag{64}$$

A.4 NEGATIVE INTERACTION

Assume $I(S;U;X) < 0$. If $\mathcal{P}_q \neq \emptyset$, Theorem 1 would imply $I(S;U;X) \geq 0$, a contradiction. Hence,

$$\mathcal{P}_q = \emptyset. \tag{65}$$

The necessary condition for ideality in Theorem 1, $I(S;U;X) \leq 0$, is compatible with negative interaction information and therefore does not rule out $\mathcal{P}^* \neq \emptyset$.

A.5 POSITIVE INTERACTION

By Theorem 1, $\mathcal{P}^* \neq \emptyset$ implies $I(S;U;X) \leq 0$. Therefore, $I(S;U;X) > 0$ implies $\mathcal{P}^* = \emptyset$. For a qualified mechanism, equality in Lemma 1 gives

$$I(U;Z) = I(U;X \mid S) + I(S;Z). \tag{66}$$

Let $t = I(S;Z)$. By Theorem 1, positive interaction implies

$$0 \leq t \leq I(S;U;X), \tag{67}$$

which yields the parameterization in equation 20. The lower extremal parameter value $t = 0$ corresponds to the coordinate

$$P_{\text{priv}} = \big(I(U;X \mid S), 0\big). \tag{68}$$

Furthermore, the interaction-information identity gives

$$I(S;U;X) = I(U;X) - I(U;X \mid S). \tag{69}$$

Hence, the upper extremal parameter value $t = I(S;U;X)$ corresponds to

$$I(U;X \mid S) + I(S;U;X) = I(U;X), \tag{70}$$

and therefore to the coordinate

$$P_{\text{util}} = \big(I(U;X), I(S;U;X)\big). \tag{71}$$

Since $I(S;U;X) > 0$, P_{priv} has utility coordinate $I(U;X \mid S) < I(U;X)$, while P_{util} has leakage coordinate $I(S;U;X) > 0$. Thus, any mechanism attaining either endpoint is not jointly ideal.

A.6 EXAMPLE: NON-ATTAINABLE ENDPOINT DESPITE POSITIVE INTERACTION INFORMATION

The extremal equality coordinates in Corollary 1 are not necessarily attained by a mechanism. To see this, let

$$U \sim \text{Bernoulli}\left(\frac{1}{2}\right), \tag{72}$$

$$N \sim \text{Bernoulli}(p), \qquad 0 < p < \frac{1}{2}, \tag{73}$$

$$S = U \oplus N, \tag{74}$$

where N is independent of U, and let $X = U$. Then,

$$I(U;X) = 1, \tag{75}$$

$$I(U;X \mid S) = H(U \mid S) = h_2(p), \tag{76}$$

where $h_2(\cdot)$ denotes the binary entropy function. Consequently,

$$I(S;U;X) = I(U;X) - I(U;X \mid S) \tag{77}$$

$$= 1 - h_2(p) > 0, \tag{78}$$

and the privacy endpoint from Corollary 1 is

$$P_{\text{priv}} = \left(h_2(p), 0\right). \tag{79}$$

Assume for contradiction that P_{priv} is attainable. Then there exists a mechanism such that

$$I(S;Z) = 0 \qquad \text{and} \qquad I(U;Z) = h_2(p) > 0. \tag{80}$$

Since $X = U$, every mechanism satisfies

$$S \to U \to Z. \tag{81}$$

For almost every z, let

$$q_z = \mathrm{P}(U = 1 \mid Z = z). \tag{82}$$

Using $S = U \oplus N$ gives

$$\mathrm{P}(S = 1 \mid Z = z) = p(1 - q_z) + (1 - p)q_z \tag{83}$$

$$= p + (1 - 2p)q_z. \tag{84}$$

Since U is uniform, S is uniform as well. Moreover, $I(S;Z) = 0$ implies statistical independence of S and Z, and therefore

$$\mathrm{P}(S = 1 \mid Z = z) = \mathrm{P}(S = 1) = \frac{1}{2}. \tag{85}$$

Hence,

$$p + (1 - 2p)q_z = \frac{1}{2}. \tag{86}$$

Since $p \neq 1/2$, it follows that

$$q_z = \frac{1}{2} = \mathrm{P}(U = 1). \tag{87}$$

Thus, U and Z are statistically independent, which implies

$$I(U;Z) = 0. \tag{88}$$

This contradicts equation 80, which requires $I(U;Z) = h_2(p) > 0$. Therefore, P_{priv} is not attainable despite $I(S;U;X) > 0$.

A.7 Lexicographic recovery of qualified endpoints

Assume the qualified privacy endpoint is attained by some $\Pi^{(q)}_{\text{priv}}$. Since mutual information is nonnegative and P_{priv} has $I(S;Z) = 0$, this mechanism attains the global minimum of the primary priLO objective. Restricting the secondary optimization to mechanisms with $I(S;Z) = 0$, Lemma 1 yields

$$I(U;Z) \leq I(U;X \mid S). \tag{89}$$

The assumed endpoint mechanism attains equality, so the secondary priLO optimum is $I(U;X \mid S)$. Hence, $\Pi^{(q)}_{\text{priv}}$ is a priLO solution, and priLO admits a solution. Furthermore, every priLO solution satisfies

$$I(S;Z) = 0 \quad \text{and} \quad I(U;Z) = I(U;X \mid S), \tag{90}$$

which is equality in Lemma 1. Hence every priLO solution is qualified and attains P_{priv}.

For utiLO, let $\Pi^{(q)}_{\text{util}}$ be a qualified mechanism attaining P_{util}. The data processing inequality gives

$$I(U;Z) \leq I(U;X). \tag{91}$$

By assumption, a qualified mechanism attains P_{util} and therefore attains equality, so $I(U;X)$ is the global optimum of the primary utiLO objective. Restricting the secondary optimization to mechanisms with $I(U;Z) = I(U;X)$, Lemma 1 yields

$$I(S;Z) \geq I(U;X) - I(U;X \mid S) \tag{92}$$

$$= I(S;U;X). \tag{93}$$

The assumed utility endpoint mechanism attains equality. Hence, $\Pi^{(q)}_{\text{util}}$ is a utiLO solution, and utiLO admits a solution. Furthermore, every utiLO solution satisfies

$$I(U;Z) = I(U;X) \quad \text{and} \quad I(S;Z) = I(S;U;X), \tag{94}$$

which is again equality in Lemma 1. Therefore every utiLO solution is qualified and attains P_{util}.

A.8 Relation to Weighted Scalarization

Let $\mathcal{R}$ denote the attainable region on the $I(U;Z) - I(S;Z)$ plane. For a more compact notation, we substitute $u = I(U;Z)$ and $s = I(S;Z)$. A scalarized privacy-utility objective function can be expressed as

$$L_\alpha(u,s) = \alpha s - (1-\alpha)u, \quad \alpha \in (0,1). \tag{95}$$

Further, suppose the privacy-first lexicographic extreme $(u_{\text{priv}}, 0) \in \mathcal{R}$ exists. For L_α to recover this point, it must satisfy

$$L_\alpha(u_{\text{priv}}, 0) \leq L_\alpha(u,s) \quad \forall (u,s) \in \mathcal{R}. \tag{96}$$

This is equivalent to

$$-(1-\alpha)u_{\text{priv}} \leq \alpha s - (1-\alpha)u. \tag{97}$$

Since points with $s = 0$ cannot jointly have $u > u_{\text{priv}}$ by definition of the privacy-first extreme, we consider the case $s > 0$, explicitly:

$$\frac{\alpha}{1-\alpha} \geq \frac{u - u_{\text{priv}}}{s}. \tag{98}$$

Since this inequality must hold for every attainable competitor, the scalarization ratio must be at least as large as the largest utility-gain-per-leakage ratio permitted by the attainable region. Thus,

$$\frac{\alpha}{1-\alpha} \geq \sup_{\substack{(u,s)\in\mathcal{R} \\ s>0}} \frac{u - u_{\text{priv}}}{s}, \tag{99}$$

which, however, depends on the attainable region $\mathcal{R}$. Hence, when the analytical bound endpoint is not attainable, recovering the privacy-first lexicographic extreme by weighted scalarization requires knowledge of the attainable region $\mathcal{R}$. The required α is therefore geometry-dependent, whereas the lexicographic priority is defined independently of this geometry. The utility-first case follows analogously.

A.9 Interpolation on the qualified bound

Proposition 2 (Interpolation on the qualified bound)**.** *Let $I(S;U;X) > 0$ and assume qualified mechanisms $\Pi^{(q)}_{\text{priv}}$ and $\Pi^{(q)}_{\text{util}}$ attain the endpoints P_{priv} and P_{util} from Corollary 1. Following the standard time-sharing construction in information theory (Tian et al., 2008), for any $\alpha \in [0,1]$, let*

$$B_\alpha \sim \text{Bernoulli}(\alpha), \qquad B_\alpha \perp\!\!\!\perp (S,U,X), \tag{100}$$

and define

$$Z_\alpha := \big(B_\alpha, Z_{B_\alpha}\big), \tag{101}$$

$$Z_{B_\alpha} := \begin{cases} \Pi^{(q)}_{\text{priv}}(X), & B_\alpha = 0, \\ \Pi^{(q)}_{\text{util}}(X), & B_\alpha = 1. \end{cases} \tag{102}$$

Then the induced privacy mechanism $\Pi^{(q)}_\alpha$ is qualified for every $\alpha \in [0,1]$, and its point in the $I(U;Z)$-$I(S;Z)$ plane is

$$P_\alpha := \big(I(U;Z_\alpha), I(S;Z_\alpha)\big) \tag{103}$$

$$= (1-\alpha)P_{\text{priv}} + \alpha P_{\text{util}}, \tag{104}$$

with

$$P_{\text{priv}} = (I(U;X \mid S), 0) \text{ and } P_{\text{util}} = (I(U;X), I(S;U;X)). \tag{105}$$

Consequently, if both endpoints are attainable, every point on the segment between them is attainable by a qualified privacy mechanism.

Let $Z_{\text{priv}} = \Pi^{(q)}_{\text{priv}}(X)$ and $Z_{\text{util}} = \Pi^{(q)}_{\text{util}}(X)$. By the chain rule of mutual information, it holds that

$$\begin{aligned} I(S; Z_\alpha) &= I(S; B_\alpha, Z_{B_\alpha}) && (106)\\ &= I(S; B_\alpha) + I(S; Z_{B_\alpha} \mid B_\alpha) && (107)\\ &= I(S; Z_{B_\alpha} \mid B_\alpha) && (108)\\ &= \sum_{b_\alpha \in \{0,1\}} P(B_\alpha = b_\alpha) I(S; Z_{B_\alpha} \mid B_\alpha = b_\alpha) && (109)\\ &= (1-\alpha) I(S; Z_{B_\alpha} \mid B_\alpha = 0) + \alpha I(S; Z_{B_\alpha} \mid B_\alpha = 1). && (110) \end{aligned}$$

By equation 102, however, we can write $(Z_{B_\alpha} \mid B_\alpha = 0) = Z_{\text{priv}}$ and $(Z_{B_\alpha} \mid B_\alpha = 1) = Z_{\text{util}}$. From this, we can simplify the I-term as

$$I(S; Z_\alpha) = (1-\alpha) I(S; Z_{\text{priv}}) + \alpha I(S; Z_{\text{util}}). \tag{111}$$

An equivalent derivation for $I(U; Z_\alpha)$ yields

$$I(U; Z_\alpha) = (1-\alpha) I(U; Z_{\text{priv}}) + \alpha I(U; Z_{\text{util}}). \tag{112}$$

Because $B_\alpha \perp\!\!\!\perp (S, U, X)$,

$$\begin{aligned} I(S; Z_\alpha \mid U) &= I(S; B_\alpha, Z_{B_\alpha} \mid U) && (113)\\ &= I(S; B_\alpha \mid U) + I(S; Z_{B_\alpha} \mid U, B_\alpha) && (114)\\ &= (1-\alpha) I(S; Z_{\text{priv}} \mid U) + \alpha I(S; Z_{\text{util}} \mid U) && (115)\\ &= 0, && (116) \end{aligned}$$

and similarly

$$\begin{aligned} I(U; Z_\alpha \mid S) &= I(U; B_\alpha, Z_{B_\alpha} \mid S) && (117)\\ &= (1-\alpha) I(U; Z_{\text{priv}} \mid S) + \alpha I(U; Z_{\text{util}} \mid S) && (118)\\ &= I(U; X \mid S). && (119) \end{aligned}$$

Thus, $\forall \alpha \in [0,1] :\ \Pi^{(q)}_\alpha \in \mathcal{P}_q$. Combining both marginal coordinates proves equation 104.

A.10 Finite-Support endpoint recovery

Neural approximation. We construct independent $C, A, B \sim \text{Bernoulli}(1/2)$ and $E \sim \text{Bernoulli}(q)$, and define $D = C \oplus E$, $S = (C, A)$, $U = (C, B)$, and $X = (D, A, B)$. For $q = 1/2$, the interaction information is zero and the privacy and utility endpoints coincide at $(1, 0)$, whereas for $q = 0$, the interaction information is one bit and the endpoints are $(1, 0)$ and $(2, 1)$, respectively. The privacy mechanism is parameterized by a four-dimensional diagonal-Gaussian conditional distribution $q_\theta(z \mid x)$ whose mean and log-variance are produced by an MLP with hidden-layer widths 128 and 64. priLO minimizes $I(S; Z)$ as its primary objective and subsequently maximizes $I(U; Z)$, whereas utiLO maximizes $I(U; Z)$ as its primary objective and subsequently minimizes $I(S; Z)$. We use vCLUB for MI minimization and BA for MI maximization, together with the adaptive gradient-surgery procedure described in Section 3.1. Each configuration is trained for 3000 optimization steps with batch size 256 and evaluated over 15 random seeds. The training, validation, and test sets contain 8192, 2048, and 4096 samples, respectively. For evaluation, we enumerate the exact finite distribution $P(X, S, U)$ and Monte-Carlo integrate over the learned Gaussian channel $Z \mid X$ using 4096 draws per support component. The resulting Gaussian-mixture density ratios are evaluated directly and therefore do not reuse the BA/vCLUB critics employed during training.

Exact enumeration. For the first synthetic experiment, we construct finite distributions for which the privacy-first and utility-first endpoints are known exactly and attained by deterministic privacy mechanisms. Let C, U_p and S_p be mutually independent random variables with

$$C, U_p, S_p \sim \text{Bernoulli}(1/2), \tag{120}$$

and define

$$U = (C, U_p) \quad \text{and} \quad S = (C, S_p). \tag{121}$$

Here, C represents information shared by the utility and sensitive variables, while U_p and S_p represent their respective private components. The observation is given by

$$X = (X_C, X_U, X_S), \tag{122}$$

where

$$X_C = C \oplus E_C, \tag{123}$$
$$X_U = U_p \oplus E_U, \tag{124}$$
$$X_S = S_p \oplus E_S, \tag{125}$$

with mutually independent $E_i \sim \mathrm{Bernoulli}(q_i)$ and $q_i \in [0, 1/2]$. Thus, each component is observed through an independent binary symmetric channel. With

$$r_i := 1 - h_2(q_i), \tag{126}$$

where h_2 denotes the binary entropy function, the independence of the individual components gives

$$I(U; X) = r_C + r_U, \tag{127}$$
$$I(U; X \mid S) = r_U, \tag{128}$$
$$I(S; U; X) = r_C. \tag{129}$$

Consequently, the two endpoint coordinates from Corollary 1 are

$$P_{\mathrm{priv}} = (r_U, 0), \tag{130}$$
$$P_{\mathrm{util}} = (r_C + r_U, r_C). \tag{131}$$

Both endpoints are attainable in this construction. For the privacy-first endpoint, consider

$$Z_{\mathrm{priv}} = X_U. \tag{132}$$

Since X_U is independent of S, it follows that

$$I(S; Z_{\mathrm{priv}}) = 0, \tag{133}$$
$$I(U; Z_{\mathrm{priv}}) = r_U, \tag{134}$$

and therefore Z_{priv} attains P_{priv}. For the utility-first endpoint, let

$$Z_{\mathrm{util}} = (X_C, X_U). \tag{135}$$

Then

$$I(U; Z_{\mathrm{util}}) = r_C + r_U = I(U; X), \tag{136}$$
$$I(S; Z_{\mathrm{util}}) = r_C = I(S; U; X), \tag{137}$$

such that Z_{util} attains P_{util}. In contrast, X_S contains information about S_p but no information about U and is therefore not contained in either endpoint mechanism.

The interaction information is directly controlled by the crossover probability of the shared channel:

$$I(S; U; X) = 1 - h_2(q_C). \tag{138}$$

For $q_C = 0$, the shared component is observed without error and $I(S; U; X) = 1$ bit. For $q_C = 1/2$, X_C is independent of C and the interaction information vanishes. Hence, varying q_C controls the separation between the two endpoint coordinates.

In addition, we consider a separate zero-interaction control with independent $S, U \sim \mathrm{Bernoulli}(1/2)$ and

$$X = (X_U, X_S), \tag{139}$$

where $X_U = U \oplus E_U$ and $X_S = S \oplus E_S$. In this case,

$$I(S; U; X) = 0 \tag{140}$$

and both endpoints coincide at

$$P_{\mathrm{priv}} = P_{\mathrm{util}} = (I(U; X), 0). \tag{141}$$

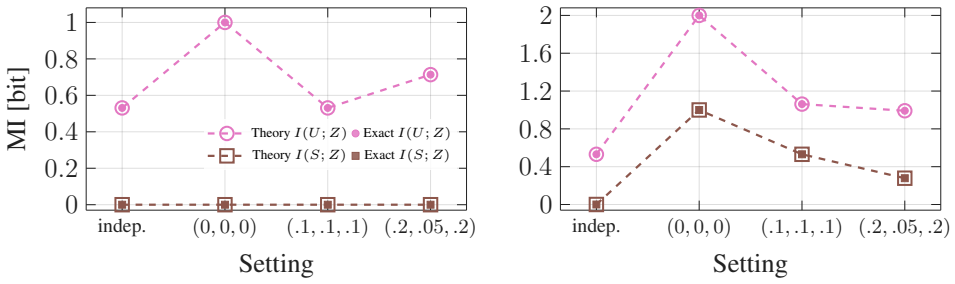


Figure 5: Exact finite-support recovery of the theoretical endpoints by priLO (left) and utiLO (right), respectively. Filled markers show the exact lexicographic solutions obtained by exhaustive deterministic-mechanism enumeration. Dashed open-marker curves show the analytical endpoint coordinates. The markers coincide up to numerical precision. The independent setting is the zero-interaction control. The remaining settings use the shared-private construction with (q_C, q_U, q_S) shown on the horizontal axis.

For the reported experiments, the independent control uses $q_U = q_S = 0.1$. For the shared-private construction, we evaluate

$$(q_C, q_U, q_S) \in \{(0,0,0), (0.1, 0.1, 0.1), (0.2, 0.05, 0.2)\}, \tag{142}$$

covering a noise-free, a symmetric noisy and an asymmetric noisy setting.

Since all variables have finite support, all mutual information quantities are computed directly from the corresponding joint probability mass functions using Shannon entropies. For each setting, deterministic mechanisms $Z = f(X)$ are enumerated up to permutations of the output labels, with at most four representation states for the shared-private construction and two states for the independent control. The objectives of all investigated methods are then solved exactly over this finite mechanism class. Therefore, this experiment is independent of variational mutual information estimation, finite-sample estimation and neural optimization. Figure 5 visualizes the resulting exact endpoint coordinates and their exhaustive-enumeration solutions.

A.11 Synthetic Dimensional Scaling

For the dimensional-scaling experiment, let $C, V \sim \text{Bernoulli}(1/2)$ be independent and define

$$S = C \qquad \text{and} \qquad U = (C, V). \tag{143}$$

The unrotated observation consists of two informative Gaussian channels and $d - 2$ independent nuisance dimensions. In particular,

$$X_C = a_C(2C - 1) + N_C, \tag{144}$$
$$X_V = a_V(2V - 1) + N_V, \tag{145}$$

where $N_C, N_V \sim \mathcal{N}(0, 1)$ are independent. The remaining coordinates are independent standard Gaussian random variables and contain no information about either S or U. Throughout the experiment, we use $a_C = a_V = 1$. The complete observation is subsequently transformed by a fixed orthogonal rotation. Since this transformation is invertible, it preserves all mutual-information quantities while preventing the informative components from remaining trivially axis-aligned.

Let

$$J(a) = I(B; a(2B - 1) + N), \tag{146}$$

for $B \sim \text{Bernoulli}(1/2)$ and $N \sim \mathcal{N}(0, 1)$. The corresponding mutual-information quantities are

$$I(S; X) = J(a_C), \tag{147}$$
$$I(U; X \mid S) = J(a_V), \tag{148}$$
$$I(U; X) = J(a_C) + J(a_V), \tag{149}$$
$$I(S; U; X) = J(a_C). \tag{150}$$

For $a_C = a_V = 1$, numerical Gauss-Hermite quadrature gives

$$J(1) \approx 0.48594 \text{ bit.} \tag{151}$$

The analytical endpoint coordinates are therefore

$$P_{\text{priv}} = (0.48594, 0)\,, \tag{152}$$

$$P_{\text{util}} = (0.97189, 0.48594)\,. \tag{153}$$

Importantly, these coordinates do not depend on the number of nuisance dimensions.

We evaluate input dimensions

$$d \in \{2, 8, 32, 128, 512, 1024\}, \tag{154}$$

where $d - 2$ coordinates contain independent Gaussian nuisance information. priLO and utiLO use the same marginal BA and vCLUB objectives as in the remaining neural experiments. The mechanism maps the observation to a four-dimensional stochastic representation. For evaluation, independent synthetic populations are generated and fresh variational posterior models are trained such that the reported mutual-information estimates do not reuse the critics employed during mechanism optimization.

Figure 3 reports the mean and standard deviation of the resulting held-out mutual-information estimates. Since the analytical endpoints remain constant with increasing d, any deviation from the horizontal reference curves is attributable to the practical neural realization rather than to a change in the underlying privacy-utility problem.

A.12 Detailed Description of the Datasets

Adult. The Adult dataset (Becker & Kohavi, 1996) is derived from U.S. census data and contains demographic and employment-related attributes such as age, work class, education, marital status, occupation, relationship, race, capital gain and loss, working hours and country of origin. We use annual income above \$50k as the binary utility variable U and sex as the sensitive variable S. The original training and test portions are combined and samples containing missing values are removed, resulting in $45{,}222$ observations. Numerical attributes are standardized and categorical attributes are one-hot encoded, yielding a 102-dimensional input representation after excluding the utility and sensitive labels.

COMPAS. The COMPAS dataset (Larson et al., 2016) contains demographic and criminal-history information used in recidivism-risk assessment. We define the binary utility variable U by whether an individual recidivates within two years and use the indicator for African-American ethnicity as the sensitive variable S. The number of prior offenses is treated as a numerical covariate, while the remaining input attributes are represented categorically. After preprocessing, the dataset contains $6{,}172$ observations with a 19-dimensional input representation.

Dutch Census. The Dutch Census dataset (van der Laan, 2000) contains demographic and socioeconomic information from the 2001 census of the Netherlands. Following the benchmark formulation, the utility variable U is a binary occupation label indicating whether the recorded occupation belongs to category `5_4_9`, while sex is used as the sensitive variable S. The remaining census attributes form the input X, with age encoded separately and the categorical attributes represented by one-hot encodings. The resulting dataset comprises $60{,}420$ observations with 48 input dimensions.

Credit. The Default of Credit Card Clients dataset (Yeh, 2016; Yeh & Lien, 2009) contains records of Taiwanese credit-card customers, including credit limits, demographic variables, repayment status, monthly bill amounts and previous payments. The utility variable U indicates whether the client defaults in the following month, while sex defines the sensitive variable S. The record identifier is discarded, education and marital status are treated as categorical variables, and the remaining financial attributes are standardized numerical features. This results in $30{,}000$ observations and a 31-dimensional input representation.

HAR. The Human Activity Recognition dataset (Reyes-Ortiz et al., 2013) contains smartphone inertial measurements recorded from 30 participants performing six activities: walking, walking upstairs, walking downstairs, sitting, standing and lying. Each observation consists of nine inertial channels over 128 consecutive samples, giving $X \in \mathbb{R}^{9\times 128}$. We use the activity class as the utility

variable U and participant identity as the sensitive variable S. Thus, unlike the four tabular datasets, HAR provides a high-dimensional temporal setting in which utility corresponds to human activity while identity information constitutes the sensitive attribute.

AudioMNIST. The AudioMNIST dataset (Becker et al., 2024) contains $30{,}000$ recordings of English spoken digits from 60 different speakers. Each speaker utters every digit from zero to nine 50 times, resulting in a balanced set of 500 recordings per speaker and $3{,}000$ recordings per digit. The recordings are provided as single-channel audio signals and speaker-level metadata are additionally available. We transform each recording into a fixed-size log-Mel spectrogram with 64 Mel-frequency bins and 98 time frames, yielding $X \in \mathbb{R}^{64\times 98}$. In our privacy-preserving representation-learning setting, we use the spoken digit as the utility variable U and speaker identity as the sensitive variable S. Hence, the task is to retain the linguistic information required for digit recognition while suppressing speaker-specific information contained in the same speech signal.

A.13 Real-data Benchmark Protocol

For every real-data benchmark, all methods use identical train, validation and test partitions and the same training budget within the respective dataset. Competing methods are evaluated for the hyperparameter values reported in Table 1, whereas priLO and utiLO do not require a privacy-utility trade-off coefficient. Each configuration is evaluated over 15 random seeds and the reported operating points are averaged over these runs. The identity mechanism $Z = X$ is included as a reference.

For the conventional baselines, information terms involving the continuous observation X, such as $I(X;Z)$, $I(X;Z \mid S)$ and $I(X;Z \mid S,U)$, are optimized directly from the known stochastic encoder density $q_\phi(z \mid x)$. Given samples $z_i \sim q_\phi(\cdot \mid x_i)$, we estimate these terms using differentiable finite Gaussian-mixture MI surrogates in latent space, with the mixture restricted to the corresponding conditioning stage for conditional mutual information. Label-information terms are optimized using BA lower bounds for maximization and vCLUB for minimization. These training surrogates are distinct from the fresh held-out BA evaluation used for the marginal coordinates reported in Figure 4.

For the four tabular datasets, numerical attributes are standardized and categorical attributes are encoded using transformations fitted only on the training subset. All variational posteriors used for label-information terms are parameterized by three-layer MLPs with 256 hidden units per layer and LeakyReLU activations with negative slope 0.01. Conditional posteriors additionally receive a one-hot encoding of the conditioning variable.

For the post-hoc classification evaluation, the learned representations are frozen and separate classifiers are trained for utility prediction $Z \to U$ and sensitive-attribute prediction $Z \to S$. Within each dataset, the utility classifier and sensitive-attribute attacker use the same architecture and training protocol, differing only in the dimensionality of their output layers. These classifiers are used exclusively for post-hoc evaluation and are independent of the variational classifiers employed during representation learning. The classifiers are fitted on the training split, selected using the validation split, and evaluated once on the held-out test split.

For the operational comparison, baseline configurations are selected from their MI operating points in Figure 4, independently of the post-hoc classification results. For CPFSI, we retain two representative configurations per dataset from the high-utility, low-leakage region. CPFSI-P denotes the configuration with lower sensitive leakage, whereas CPFSI-U denotes the configuration with higher retained utility. For CFB, we use $\beta = 1024$, which is the evaluated configuration that remains non-collapsed across all datasets while consistently lying near the high-utility, low-leakage region.

A.14 Complete Post-hoc Classifier Results

For completeness, we report the post-hoc macro-F_1 scores for all datasets and methods with varied hyperparameters as described in Section A.13.

Table 3: Post-hoc utility macro-F_1 (in %) on frozen representations for all evaluated methods and hyperparameter configurations. Higher F_1^U is better.

		Utility $F_1^U \uparrow$					
Method	Parameters	Adult	COMPAS	Dutch	Credit	HAR	AudioMNIST
$Z = X$	–	78.7	68.1	81.8	67.8	93.3	99.6
priLO	–	60.2	49.0	77.0	65.5	94.4	31.6
utiLO	–	78.6	67.0	81.6	68.1	98.6	99.9
IB	$\beta = 0.1$	42.93	47.68	44.39	43.78	15.08	8.96
	$\beta = 1$	42.93	47.67	44.39	43.78	15.10	8.94
	$\beta = 10$	77.37	62.05	80.76	63.95	98.14	99.87
PF	$\beta = 0.1$	55.72	57.34	66.37	54.41	56.47	54.33
	$\beta = 1$	56.73	59.50	66.90	57.12	53.02	47.69
	$\beta = 10$	56.62	57.98	68.57	57.27	49.20	42.25
CPF	$\gamma = 1$	55.91	57.80	67.62	57.46	42.39	36.29
	$\gamma = 64$	55.71	57.80	68.19	57.26	38.08	27.57
	$\gamma = 1024$	55.36	58.84	68.42	56.23	42.62	27.18
CFB	$\beta = 1$	42.93	47.73	44.46	43.78	95.82	99.85
	$\beta = 64$	78.47	66.89	81.21	67.88	97.14	99.86
	$\beta = 1024$	78.8	67.6	81.5	68.1	96.8	99.8
CPFSI	$\beta = 1,\ \gamma = 1$	76.1	65.43	80.80	67.79	97.18	99.75
	$\beta = 1,\ \gamma = 64$	76.15	65.70	80.79	67.8	90.91	77.12
	$\beta = 1,\ \gamma = 1024$	76.08	65.78	80.7	67.52	89.7	73.68
	$\beta = 64,\ \gamma = 1$	78.89	67.79	81.48	67.95	96.8	99.8
	$\beta = 64,\ \gamma = 64$	78.74	67.82	81.50	67.58	96.19	99.79
	$\beta = 64,\ \gamma = 1024$	78.87	67.6	81.44	67.68	95.38	99.45
	$\beta = 1024,\ \gamma = 1$	78.9	67.8	81.50	67.96	96.80	99.9
	$\beta = 1024,\ \gamma = 64$	78.86	67.79	81.5	67.69	96.67	99.9
	$\beta = 1024,\ \gamma = 1024$	78.87	67.58	81.53	68.0	96.12	99.83

Table 4: Post-hoc leakage macro-F_1 (in %) on frozen representations for all evaluated methods and hyperparameter configurations. Lower F_1^S is better.

		Attacker $F_1^S \downarrow$					
Method	Parameters	Adult	COMPAS	Dutch	Credit	HAR	AudioMNIST
$Z = X$	–	83.0	69.1	67.7	56.1	36.0	98.0
priLO	–	40.3	48.8	51.1	38.5	2.8	1.9
utiLO	–	52.9	61.1	60.0	39.0	2.0	1.7
IB	$\beta = 0.1$	40.30	48.82	49.09	38.03	2.70	1.20
	$\beta = 1$	40.30	48.83	49.09	38.02	2.70	1.20
	$\beta = 10$	49.27	56.92	58.91	38.54	2.16	1.01
PF	$\beta = 0.1$	40.87	51.00	50.78	38.39	3.87	5.01
	$\beta = 1$	40.55	53.92	51.41	38.67	3.66	3.92
	$\beta = 10$	40.39	53.14	52.40	38.46	4.19	4.17
CPF	$\gamma = 1$	40.45	52.12	51.63	38.87	3.19	2.89
	$\gamma = 64$	40.52	53.35	52.10	38.94	3.04	2.59
	$\gamma = 1024$	40.36	55.97	52.48	38.69	3.22	2.47
CFB	$\beta = 1$	40.30	48.80	49.10	38.02	2.33	1.13
	$\beta = 64$	50.04	59.97	57.61	38.41	2.93	0.95
	$\beta = 1024$	52.8	62.9	59.4	38.7	4.8	2.1
CPFSI	$\beta = 1,\ \gamma = 1$	41.2	56.53	56.72	39.23	12.35	19.93
	$\beta = 1,\ \gamma = 64$	41.02	58.85	56.96	39.2	4.62	16.55
	$\beta = 1,\ \gamma = 1024$	41.16	58.98	56.9	39.21	4.8	13.70
	$\beta = 64,\ \gamma = 1$	51.85	62.70	59.56	42.61	10.9	8.6
	$\beta = 64,\ \gamma = 64$	53.18	62.95	59.81	43.17	9.78	16.56
	$\beta = 64,\ \gamma = 1024$	54.30	65.6	59.52	43.51	7.43	20.20
	$\beta = 1024,\ \gamma = 1$	53.4	62.7	60.35	42.31	9.90	11.1
	$\beta = 1024,\ \gamma = 64$	56.32	65.19	60.3	46.24	12.45	11.1
	$\beta = 1024,\ \gamma = 1024$	59.04	65.64	60.95	42.6	11.73	12.91

A.15 Additional Real-Data Experiment: MotionSense

As an auxiliary non-tabular real-world experiment, we evaluate MotionSense (Malekzadeh et al., 2019), a smartphone inertial-sensor dataset containing 50-Hz accelerometer and gyroscope-derived time series from 24 participants. We use 128-sample windows of the 12 raw DeviceMotion channels as X, the four dynamic activities downstairs, upstairs, walking and jogging as the utility variable U, and gender as the sensitive variable S. Complete recordings are split before windowing, and all learned methods use the same end-to-end temporal stochastic encoder. The reported learned-method coordinates average three optimization seeds on one fixed recording-disjoint train/validation/test split; the identity mechanism $Z = X$ is deterministic and is evaluated once.

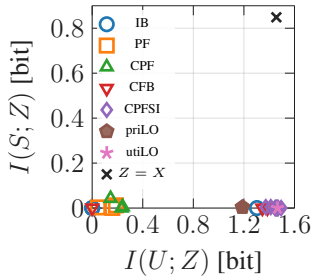


Figure 6: MotionSense results in the marginal utility-privacy plane. Colors and markers match Figure 4. Each learned-method point is the mean over optimization seeds $0, 1, 2$ for the fixed dataset split; $Z = X$ is deterministic and shown once.

The architecture-matched held-out BA estimates for the raw signal are

$$\widehat{I}(U;X) = 1.4565 \text{ bit}, \qquad \widehat{I}(U;X \mid S) = 1.4457 \text{ bit}, \tag{155}$$

$$\widehat{I}(S;X) = 0.8486 \text{ bit}, \qquad \widehat{I}(S;X \mid U) = 0.8288 \text{ bit}. \tag{156}$$

Accordingly, the two equivalent interaction-information decompositions yield the empirical differences

$$\widehat{I}(U;X) - \widehat{I}(U;X \mid S) = 0.0108 \text{ bit}, \tag{157}$$

$$\widehat{I}(S;X) - \widehat{I}(S;X \mid U) = 0.0198 \text{ bit}. \tag{158}$$

Because the four quantities are separately fitted variational lower bounds, neither difference is interpreted as a direct estimator of $I(S;U;X)$. Instead, their small magnitude provides empirical evidence that their small magnitude is consistent with a near-zero-interaction regime at the resolution of the evaluator. This interpretation is consistent with the independently trained post-hoc probes: raw X yields utility and sensitive balanced accuracies of 0.9312 and 0.9636, respectively, whereas utiLO retains utility balanced accuracy 0.9261 while reducing the sensitive balanced accuracy to 0.4991.

Figure 6 shows substantial sensitive leakage for the identity reference, while multiple learned mechanisms approach zero leakage with utility close to the raw-signal utility level. We therefore use MotionSense primarily as a real-world illustration of the near-zero-interaction regime, rather than as a benchmark intended to establish superiority of one optimization method over another.