PRIVACY WITHOUT PRIVATIZATION: INHERENT MEMBERSHIP PRIVACY IN RIDGE REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Privacy in machine learning is commonly obtained by modifying the learning algorithm, for example to satisfy differential privacy (DP). We ask a complementary question: how much privacy may already be present in an ordinary, unmodified model? We study this in the context of membership inference risk in ridge regression models. To this end, we employ a recent information theoretic notion called *pointwise maximal cost* (PMC), which quantifies how much observing a particular released model can improve an adversary's inference relative to prior knowledge. In particular, we derive an exact characterization of membership inference risk under the PMC interpretation for ridge regression. In regimes where the numerical estimates are reliable, leakage is nonzero but generally small, typically decreases as the number of training records grows relative to dimension, and varies across released models and membership inference targets. The same qualitative behavior persists under a feature population derived from real-world data. Furthermore, at matched privacy levels, we show that two existing formal DP regression methods incur substantially greater prediction error than ordinary ridge regression. Thus, our results suggest that PMC can serve as a release-evaluation tool for quantifying inherent privacy before deciding whether additional privatization is necessary.

1 INTRODUCTION

A trained model is a compressed representation of its training data, but it does not necessarily reveal every property of those data equally well. In particular, observing a released model may provide evidence about whether specific records were used during training, as demonstrated by membership inference attacks (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022). A typical response is to modify the learning algorithm to satisfy differential privacy (DP), which requires similar output distributions on neighboring datasets (Dwork et al., 2006; Dwork & Roth, 2014). For regression, this has motivated objective perturbation, output perturbation, and private sufficient-statistic methods (Chaudhuri et al., 2011; Sheffet, 2017; Wang, 2018).

We make a distinction between two related but distinct privacy questions. The first concerns the *learning procedure*: does the algorithm provide a privacy guarantee uniformly over possible training datasets and outputs? Differential privacy, and much of the associated privacy auditing and certification literature, addresses this algorithm-level question. We instead study the privacy of the *realized model*: given an already trained model produced without any privacy mechanism, how much does its release reveal about a particular private fact? An unmodified learning algorithm does not generally satisfy DP, but this does not imply that every realized model reveals its training data equally well. Many different training datasets can produce similar models, and sampling variability, aggregation, regularization, and observation noise can limit what a particular release reveals. We refer to the privacy arising from these effects, without explicit privatization, as *inherent privacy*. This perspective is conceptually related to privacy certification, but shifts the object being evaluated from the learning procedure to the proposed model release. A trusted gatekeeper could therefore evaluate the inherent privacy of a model, with respect to a specified population and protected secret, before deciding whether additional privatization is required prior to releasing the model; Appendix A.5 discusses this setting in more detail.

We quantify this inherent privacy using pointwise maximal cost (PMC) (Saeidian et al., 2025). For a specified secret and a particular released model, PMC measures the largest multiplicative reduction

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.

in an adversary's optimal expected nonnegative inference cost after observing the release, relative to what is achievable from the adversary's prior knowledge alone. A small PMC means that observing the released model does not substantially improve the adversary's ability to infer the protected secret.

A central feature of this perspective is that inherent privacy is not a property of the learning algorithm alone. It depends jointly on the protected secret, the population model governing the training data, and the particular released model. Two releases produced by the same algorithm can therefore provide different amounts of privacy, and the same release can provide different privacy guarantees for different secrets. Our objective is to show that, once these ingredients are specified, the resulting inherent privacy can be characterized and evaluated in terms of the associated PMC.

We study this question for ordinary ridge regression and a *proximity-membership secret*, i.e., whether the training dataset contains at least one feature vector within a prescribed distance of a public target. Using a neighborhood rather than exact equality gives a nondegenerate membership event for continuous feature populations. We ask how much observing the released ridge coefficients improves an adversary's ability to infer this secret relative to its prior knowledge.

Our main technical contribution is an exact characterization of proximity-membership PMC for ordinary ridge regression. To this end, we consider the ridge estimator under a linear response model with Gaussian observation noise and show that, conditioned on the complete feature matrix, the released ridge coefficients are Gaussian. We then characterize exactly the feature-matrix distributions conditioned on membership and non-membership. Averaging the conditional Gaussian density over these two feature-matrix distributions gives the corresponding densities of the released ridge coefficients under membership and non-membership. The likelihood ratio between these densities determines the posterior membership probability and hence PMC. The secret-conditioning argument itself requires only i.i.d. feature vectors and therefore extends beyond Gaussian feature populations.

Using this characterization, we evaluate PMC across independently trained ridge models. In the regimes where the numerical estimates are reliable, ordinary ridge provides meaningful proximity-membership privacy: leakage is nonzero but generally small and typically decreases as the number of training records grows relative to dimension. Privacy nevertheless varies across released models and protected targets. Identical broad behavior persists when the Gaussian feature population is replaced by a smoothed empirical population derived from real-world data.

We also compare prediction utility with two DP regression methods. For each ordinary-ridge release, we map its PMC value to a corresponding symmetric likelihood-ratio level and evaluate the DP methods at that same numerical level. This provides a common scale for comparing utility without treating PMC and DP as equivalent guarantees. In the regimes studied, the DP methods incur substantially larger prediction error, while ordinary ridge retains much better predictive accuracy.

Our main contributions are fourfold. **First, inherent privacy without explicit privatization:** we show that ordinary ridge regression can provide meaningful proximity-membership privacy through the released model even without privacy-specific randomization. **Second, exact characterization of proximity-membership privacy:** we derive the secret-conditioned output distributions for ordinary ridge regression and show how their likelihood ratio determines the posterior membership probability and PMC. **Third, empirical evaluation across releases and targets:** we evaluate privacy across independently trained models and find that it varies with the release, number of training records, and protected target, while generally improving as the number of training records grows relative to dimension in regimes where numerical estimation is reliable. **Fourth, privacy–utility comparison with DP:** we compare ordinary ridge with two DP regression baselines at corresponding numerical likelihood-ratio levels and find substantially greater prediction error under the DP methods in the regimes studied.

2 Related Work

Membership inference. Membership inference studies whether a released model reveals whether particular records were used during training (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022). Much of this literature evaluates specific attacks or record-level membership. We instead consider *proximity membership*: whether the training dataset contains any feature vector near a public target. Rather than analyzing a particular attack, we characterize the output distributions under membership and non-membership and quantify the resulting disclosure about this secret.

Privacy auditing and certification. Privacy auditing and certification commonly assess the privacy properties of a learning procedure or implemented privacy mechanism. In the differential-privacy setting, empirical auditing methods seek lower bounds on the privacy loss of an implemented algorithm and can reveal loose analyses or implementation failures (Jagielski et al., 2020; Nasr et al., 2023). Membership inference and related hypothesis-testing procedures are frequently used as auditing tools since successful discrimination between neighboring training conditions provides evidence about the privacy of the learning mechanism.

Our objective is conceptually related but evaluates a different object. We do not certify that the learning procedure satisfies a uniform privacy guarantee. Instead, we evaluate the *inherent privacy of a realized model*: for a specified population model and protected secret, we quantify how much observing a particular ordinary-model release improves inference about that secret. PMC makes this evaluation independent of any particular membership inference attack or loss function. This model-level perspective suggests a complementary form of release evaluation in which a trusted gatekeeper assesses a proposed model before release and introduces additional privatization only when the specified privacy criterion is not met.

Information-theoretic measures of disclosure. Several information-theoretic notions quantify how an observation improves inference about private information. Maximal leakage measures multiplicative improvement in guessing (Issa et al., 2019), while pointwise variants localize disclosure to a realized output (Saeidian et al., 2023). Related posterior-to-prior measures include lift and local information privacy (Zarrabian et al., 2023). Most directly, pointwise maximal cost (PMC) characterizes the largest multiplicative reduction in optimal expected inference cost after observing a particular release (Saeidian et al., 2025). We use PMC to measure disclosure about proximity membership and derive secret-conditioned output distributions for ordinary ridge regression.

Differentially private regression. Differential privacy provides a distribution-free guarantee that controls changes in the output distribution over neighboring datasets (Dwork et al., 2006; Dwork & Roth, 2014). Private regression methods include objective perturbation, output perturbation, and perturbation of sufficient statistics (Chaudhuri et al., 2011; Sheffet, 2017; Wang, 2018); they introduce randomness specifically to obtain formal privacy guarantees. Our question is complementary: how much privacy is already present in an unmodified estimator for a specified secret under an assumed population model? We therefore compare ordinary ridge with DP regression at corresponding numerical likelihood-ratio levels, without treating the two privacy guarantees as equivalent.

3 Problem Setup

Data model and released estimator. Let $D_n = \{(X_i, Y_i)\}_{i=1}^n$ consist of n i.i.d. training records, where $X_i \in \mathbb{R}^p$ is a feature vector and $Y_i = X_i^\top w^\star + \epsilon_i$, with $w^\star \in \mathbb{R}^p$ fixed and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ independent of the features. Our primary synthetic model takes $X_i \sim \mathcal{N}(0, I_p)$; Section 6 also considers a non-Gaussian feature population derived from real data. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^n$ denote the stacked features and responses. The released model is the ordinary ridge estimator

$$W = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda I_p\right)^{-1} \frac{1}{n}\mathbf{X}^\top\mathbf{Y}, \qquad \lambda > 0. \tag{1}$$

No clipping, objective perturbation, or output noise is introduced for privacy.

Proximity-membership secret. Fix a public target $z \in \mathbb{R}^p$ and radius $r > 0$, and let $B(z, r) = \{x \in \mathbb{R}^p : \|x - z\|_2 \le r\}$. The private fact we wish to protect is the *proximity-membership secret*, i.e., $S_r = \mathbf{1}\{\text{at least one } X_i \in B(z, r)\}$. Thus, $S_r = 1$ means that the training dataset contains at least one feature vector within distance r of the target. For continuous feature populations, the positive-radius neighborhood avoids the probability-zero event of exact equality with a fixed target.

Let $\pi_r = \Pr(S_r = 1)$ denote the prior probability of proximity membership, which is available to the adversary prior to seeing the trained model. For any $i \in [n]$, let $a_{r,i} = \Pr(X_i \in B(z, r))$. Since the feature vectors are identically distributed, it follows that this probability does not actually depend on i, and we write $a_r = a_{r,i}$ in the sequel. Since the feature vectors are also independent, $\pi_r = 1 - (1 - a_r)^n$. In the experiments comparing target locations, we choose r so that π_r remains fixed, allowing changes in disclosure to be separated from changes in the prior probability of membership.

Pointwise maximal cost. For a finite-valued secret S and a realized output w, pointwise maximal cost (PMC) (Saeidian et al., 2025, Definition 11) is defined as

$$\Lambda_{P_{S,W}}(S \to w) = \log \max_s \frac{\Pr(S = s)}{\Pr(S = s \mid W = w)}, \tag{2}$$

where all logarithms are natural.

The operational characterization in (Saeidian et al., 2025, Theorem 3) shows that PMC equals the largest possible multiplicative reduction in an adversary's optimal expected nonnegative inference cost after observing w, relative to the optimal cost achievable from prior knowledge alone. PMC is thus not tied to a particular training algorithm, inference attack, or loss function. A small PMC implies the released model cannot substantially improve the adversary's inference about the secret, regardless of inference cost or attack mechanism, and regardless of adversarial computation power.

PMC is also suitable for our setting since our privacy question concerns a specified secret and a particular model release. Unlike an average information measure, PMC evaluates disclosure at the observed output w itself. This allows different trained models produced by the same learning algorithm to have different privacy values, which is essential for the release-evaluation perspective considered here. Appendix A.3 provides additional details about the definition and its relation to likelihood ratios.

For proximity membership, let $q_r(w) = \Pr(S_r = 1 \mid W = w)$. Since S_r is binary, the general PMC definition in (2) specializes to

$$\Lambda_r(w) = \log \max \left\{ \frac{\pi_r}{q_r(w)}, \frac{1 - \pi_r}{1 - q_r(w)} \right\}. \tag{3}$$

Hence, a small $\Lambda_r(w)$ implies that observing the released model cannot make either value of the secret substantially less plausible relative to its prior probability.

4 Exact Privacy Characterization

In what follows, we assume that $n \geq p$ and that the feature matrix $\mathbf{X}$ has full column rank almost surely. To evaluate PMC, we require the distribution of the released ridge coefficients under the two possible values of the proximity-membership secret. First, Section 4.1 characterizes the released coefficients conditional on a realized feature matrix. Second, Section 4.2 characterizes the feature-matrix distribution under membership and non-membership. Section 4.3 then combines the two to obtain the secret-conditioned output densities and the resulting PMC.

4.1 Ridge Output Conditional on the Feature Matrix

Fix a realized feature matrix $\mathbf{x}$, and define $S_{\mathbf{x}} = \mathbf{x}^\top \mathbf{x}/n$ and $H_{\mathbf{x}} = S_{\mathbf{x}} + \lambda I_p$. Substituting $Y_i = X_i^\top w^\star + \epsilon_i$ into (1) gives

$$W \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}), \qquad \mu_{\mathbf{x}} = H_{\mathbf{x}}^{-1} S_{\mathbf{x}} w^\star, \qquad \Sigma_{\mathbf{x}} = \frac{\sigma_\epsilon^2}{n} H_{\mathbf{x}}^{-1} S_{\mathbf{x}} H_{\mathbf{x}}^{-1}. \tag{4}$$

This conditional Gaussian distribution follows from the linear dependence of the responses on the features and the Gaussian observation noise; the feature distribution itself need not be Gaussian. A full proof is given in Appendix B.1.

Let $g_{\mathbf{x}}(w) = f_{W|\mathbf{X}}(w \mid \mathbf{x})$ denote the Gaussian density in (4). To obtain the distribution of the released model under each value of the secret, we average this conditional density over the corresponding distribution of the feature matrix $\mathbf{X}$. Section 4.2 therefore characterizes $P_{\mathbf{X}|S_r=0}$ and $P_{\mathbf{X}|S_r=1}$; Section 4.3 then averages $g_{\mathbf{X}}(w)$ under these two distributions to obtain the output densities needed to evaluate PMC.

4.2 Feature Distribution Under Membership and Non-Membership

Let $K_r = \sum_{i=1}^n \mathbf{1}\{X_i \in B(z, r)\}$ denote the number of training feature vectors inside the protected neighborhood. When $S_r = 0$, every feature vector lies outside $B(z, r)$. Since the X_i are independent

and the event $S_r = 0$ is the product event $\bigcap_{i=1}^{n}\{X_i \notin B(z,r)\}$, conditioning on $S_r = 0$ preserves independence across rows. Hence the n rows of $\mathbf{X}$ are independent draws from $P_X(\cdot \mid X \notin B(z,r))$. When $S_r = 1$, at least one row lies in the neighborhood, but neither the number of such rows nor their locations in the dataset are fixed. In this case, $K_r \sim \text{Binomial}(n, a_r) \mid K_r \geq 1$. Conditional on $K_r = k$, choose the k row indices uniformly among the $\binom{n}{k}$ possible subsets. Draw the selected rows independently from $P_X(\cdot \mid X \in B(z,r))$ and the remaining rows independently from $P_X(\cdot \mid X \notin B(z,r))$.

This procedure provides the distribution of $\mathbf{X}$ conditional on $S_r = 1$. Conditioning one predetermined row to lie in the neighborhood would not result in the correct distribution, since the secret specifies only that at least one feature vector lies in the neighborhood. It does not specify which row satisfies the event or how many rows do so.

The argument in this section uses only the assumption that the feature vectors are i.i.d. and therefore applies to feature populations beyond the Gaussian model used in the primary synthetic experiments. Appendix B.2 proves that these constructions give exactly $P_{\mathbf{X}|S_r=0}$ and $P_{\mathbf{X}|S_r=1}$.

4.3 Output Distributions and PMC

For $s \in \{0,1\}$, define

$$g_s(w) = f_{W|S_r=s}(w) = \mathbb{E}[g_{\mathbf{X}}(w) \mid S_r = s]\,. \tag{5}$$

That is, g_0 and g_1 are the output densities of the released ridge coefficients under non-membership and membership, respectively. Each is obtained by averaging the conditional Gaussian density (4) over the corresponding feature-matrix distribution. In what follows, let $L_r(w) = \log[g_1(w)/g_0(w)]$.

Theorem 1 (Exact proximity-membership PMC) *Suppose $0 < \pi_r < 1$ and $g_0(w), g_1(w) > 0$. Then*

$$\Lambda_r(w) = \log\max\left\{\pi_r + (1-\pi_r)e^{-L_r(w)},\ 1 - \pi_r + \pi_r e^{L_r(w)}\right\}. \tag{6}$$

In particular, for the balanced prior $\pi_r = 1/2$ we have that $\Lambda_r(w) = \log[(1 + e^{|L_r(w)|})/2]$.

The proof is given in Appendix B. Once the two output densities $g_0(w)$ and $g_1(w)$ are known, PMC follows exactly from Theorem 1. The purpose of the numerical approximation in Section 6 which follows is to evaluate the expectations in (5).

The characterization of PMC in Theorem 1 also explains how inherent privacy can arise. Although the feature-matrix distributions differ under membership and non-membership, they may induce similar distributions of the released coefficients. When $g_0(w)$ and $g_1(w)$ are close, observing w provides little evidence for distinguishing the two secret values.

5 Evaluating Inherent Privacy and Matching to DP

Section 4 showed PMC is determined by the output densities $g_0(w)$ and $g_1(w)$. These densities are expectations over the feature-matrix distributions under non-membership and membership, respectively. These expectations are generally unavailable analytically, so we evaluate them numerically.

5.1 Monte Carlo Evaluation of PMC

For each $s \in \{0,1\}$, we generate independent feature-matrix realizations $\mathbf{x}_s^{(1)}, \ldots, \mathbf{x}_s^{(M)}$ independently from $P_{\mathbf{X}|S_r=s}$ using the exact conditioning procedure in Section 4.2, and estimate $\widehat{g}_s(w) = \frac{1}{M}\sum_{m=1}^{M} g_{\mathbf{x}_s^{(m)}}(w)$, where $g_{\mathbf{x}}(w)$ is the conditional Gaussian density of the released ridge coefficients given $\mathbf{X} = \mathbf{x}$.

For fixed w, the estimator $\widehat{g}_s(w)$ is unbiased for $g_s(w)$. We compute $\widehat{L}_r(w) = \log[\widehat{g}_1(w)/\widehat{g}_0(w)]$ and substitute it into (6) to obtain $\widehat{\Lambda}_r(w)$. The resulting likelihood-ratio and PMC estimates are not generally unbiased, since they are nonlinear functions of the two mixture estimates.

The mixture estimate $\widehat{g}_s(w)$ is an average of likelihood contributions $g_{\mathbf{x}_s^{(m)}}(w)$. These contributions can be extremely uneven: many sampled feature matrices may contribute almost nothing,

while a small number dominate the average. In that situation, a nominal Monte Carlo budget of M can behave as though far fewer draws were informative. We therefore report an effective sample size (ESS), which here measures the effective number of feature-matrix draws contributing to the mixture average. If all M likelihood terms contribute equally, ESS equals M; if only one term dominates, ESS approaches one. Thus, ESS summarizes the concentration of the terms entering the mixture-density estimate. It is not an estimate of the sampling error of PMC and does not provide a finite-simulation confidence guarantee. We use it only as a numerical warning sign and examine independent-run stability when the contributions are strongly concentrated. The precise definition is given in Appendix B.6.

5.2 Matching PMC to DP Baselines

To compare prediction utility with DP regression methods, we map the PMC of each ordinary-ridge release to a corresponding symmetric likelihood-ratio level. This provides a common numerical scale for the comparison without treating PMC and DP as equivalent privacy guarantees.

Recall that $L_r(w) = \log[g_1(w)/g_0(w)]$. For a fixed prior π_r, consider a symmetric likelihood-ratio bound $|L_r(w)| \le \varepsilon$. By Theorem 1, the largest PMC compatible with this bound is

$$\log\left(\min\{\pi_r, 1-\pi_r\} + \max\{\pi_r, 1-\pi_r\}e^{\varepsilon}\right).$$

We define $\varepsilon_{\mathrm{match}}(w)$ by equating this worst-case value to the observed PMC value $\Lambda_r(w)$. Solving for ε gives

$$\varepsilon_{\mathrm{match}}(w) = \log \frac{e^{\Lambda_r(w)} - \min\{\pi_r, 1-\pi_r\}}{\max\{\pi_r, 1-\pi_r\}}.$$

Thus, $\varepsilon_{\mathrm{match}}(w)$ is the smallest symmetric likelihood-ratio radius whose worst-case PMC equals the observed PMC of the release. In particular, when $\pi_r = 1/2$, $\varepsilon_{\mathrm{match}}(w) = |L_r(w)|$. Appendix A.4 gives the full derivation and interpretation. We use $\varepsilon_{\mathrm{match}}$ only as a common numerical likelihood-ratio level for evaluating the DP baselines. In particular, this matching *does not* imply that ordinary ridge satisfies $(\varepsilon_{\mathrm{match}}, \delta_{\mathrm{DP}})$-DP.

For each ordinary-ridge release, we evaluate two DP regression methods at its corresponding $\varepsilon_{\mathrm{match}}$. The first clips the feature and response spaces, constrains the ridge solution, and adds Gaussian output noise calibrated to global sensitivity (Dwork & Roth, 2014). The second is AdaSSP (Wang, 2018), which perturbs regression sufficient statistics and adapts regularization using privately estimated spectral information. Full clipping bounds, sensitivity calculations, privacy-budget allocation, and implementation details are given in Appendix C.

6 Experiments

We evaluate inherent privacy under two feature populations. The primary synthetic experiments use isotropic Gaussian features, which allow controlled variation of the dimension, number of training records, membership prior, and protected target (see Appendix D.6 for details). We then repeat the analysis using a smoothed empirical feature population derived from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset Street et al. (1993) to test whether the same qualitative behavior persists under feature geometry derived from real-world measurements.

6.1 Evaluation Metrics

We use three complementary quantities. *Inherent privacy* is measured by PMC; smaller values indicate that observing the release provides less additional information about the proximity-membership secret. When aggregating across independently trained models, we report the mean PMC and its standard error, together with selected distributional summaries in the appendix. *Prediction utility* is measured by population mean squared error (MSE). For the isotropic Gaussian feature population, this quantity is available in closed form; for the WDBC-derived population, it is evaluated under the same semi-synthetic response model. Finally, we report *minESS* as a measure of the numerical concentration of the Monte Carlo likelihood estimates; this metric is used to identify regimes in which relatively few feature-matrix draws dominate an estimated mixture density and the resulting PMC estimate should therefore be interpreted more cautiously. The precise definition is given in

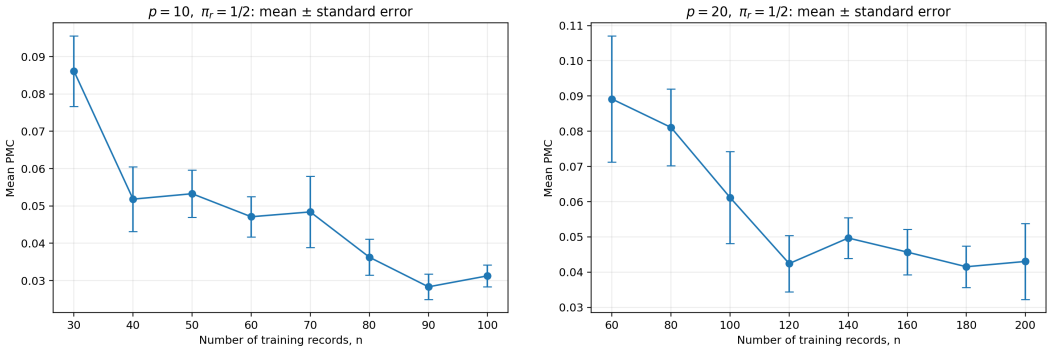


Figure 1: Proximity-membership PMC across independently trained ridge models. Left: $p = 10$. Right: $p = 20$. Leakage is nonzero but generally small in the reliable regimes and tends to decrease as n grows, although some releases are more revealing than others.

Table 1: Synthetic proximity-membership leakage at $\pi_r = 1/2$. minESS summarizes concentration of Monte Carlo likelihood estimates; smaller values indicate greater numerical difficulty.

p	n	Median PMC	90th percentile PMC	Median minESS
10	30	0.0419	0.1165	934.3
10	50	0.0411	0.0773	1400.3
10	100	0.0248	0.0561	2145.1
20	60	0.0713	0.4471	85.0
20	100	0.0284	0.1162	628.2
20	200	0.0190	0.0585	847.5

Appendix B.6, and the resulting values can be interpreted as follows: if the normalized likelihood contributions were equally spread across exactly k draws and negligible for the others, then ESS would equal k. Thus, an ESS of k can be interpreted as having approximately the same weight concentration as k equally influential feature-matrix draws. ESS does not provide a confidence interval or standard error for PMC. Nevertheless, it gives a useful scale for the numerical difficulty of an individual mixture-density estimate.

6.2 Synthetic Leakage and Privacy–Utility Comparison

For the synthetic experiments, we take $w^\star = e_1$, $\lambda = 0.1$, and $\sigma_\epsilon = 0.25$, with main target $z = 2e_1$. Since the Gaussian feature population is isotropic, it follows that setting $w^\star = e_1$ fixes a reference signal direction. We place the main target on the same direction at distance two from the origin so that the protected neighborhood is neither centered at the population mean nor orthogonal to the predictive direction. We consider $p = 10$ with $n \in \{30, 40, 50, 60, 70, 80, 90, 100\}$ and $p = 20$ with $n \in \{60, 80, 100, 120, 140, 160, 180, 200\}$. The main results use 60 independently trained ridge models per regime. For the Gaussian feature population, prediction risk is $R(w) = \|w - w^\star\|_2^2 + \sigma_\epsilon^2$. Complete simulation budgets, results for additional membership priors, and implementation details are given in Appendix D.

Figure 1 shows the mean PMC across independently trained ridge models, with error bars denoting one standard error of the mean. Across both values of p, mean leakage is nonzero but generally small and shows an overall tendency to decrease as n grows. The trend is not strictly monotone across the evaluated values of n, but the larger-n regimes generally exhibit lower mean PMC than the smaller-n regimes. The standard errors are small relative to the corresponding mean PMC values, indicating that the estimated means are reasonably stable across the independently trained models.

There is no universal ESS threshold that guarantees an accurate PMC estimate. We therefore interpret ESS together with independent-run stability. In Table 1, the $p = 10$ regimes have median minESS between approximately 934 and 2145, indicating that the mixture estimates are supported by many effective contributions. The $p = 20, n = 100$ and $n = 200$ regimes are also reasonably

Table 2: Paired synthetic privacy–utility comparison. The DP methods are evaluated at model-specific values of $\varepsilon_{\text{match}}$.

p	n	Median $\varepsilon_{\text{match}}$	Ridge MSE	Gaussian-DP MSE	AdaSSP MSE
10	30	0.0821	0.1045	5.1743	1.1050
10	50	0.0806	0.0877	5.1820	1.1100
10	100	0.0489	0.0804	5.0527	1.1156
20	60	0.1379	0.1091	5.0726	1.0854
20	100	0.0561	0.0896	5.0578	1.0908
20	200	0.0377	0.0787	5.1173	1.0861

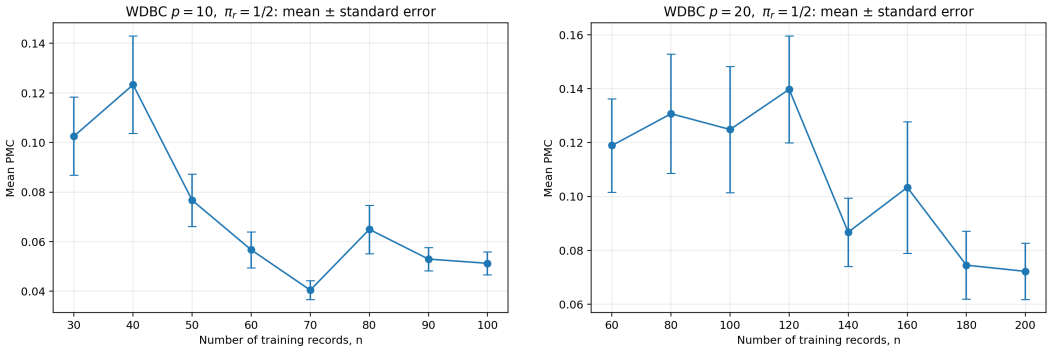


Figure 2: PMC under smoothed WDBC feature population. Left: $p = 10$. Right: $p = 20$.

well supported, with median minESS of approximately 628 and 848. In contrast, the $p = 20, n = 60$ regime has median minESS of only about 85, corresponding to a much more concentrated estimate. We therefore regard this regime, and especially its upper-tail PMC values, as numerically less reliable and interpret it cautiously.

We next compare prediction utility with Gaussian output perturbation and AdaSSP. For each ordinary-ridge release, we compute its corresponding $\varepsilon_{\text{match}}$ and evaluate both DP methods at that same numerical privacy level. An important feature of Table 2 is that these matched numerical levels are very small. Across the six regimes, the median $\varepsilon_{\text{match}}$ ranges from approximately 0.038 to 0.138. On a symmetric likelihood-ratio scale, these values correspond to multiplicative factors $e^{\varepsilon_{\text{match}}}$ of only about 1.04–1.15. Thus, the DP baselines are being evaluated in a stringent privacy regime. This distinction is important when interpreting the utility results. In practical differentially private machine learning, substantially larger privacy budgets are often considered in order to obtain usable predictive performance. For example, ref. Ponomareva et al. (2023) describes $\varepsilon \leq 1$ as a strong privacy target and $\varepsilon \leq 10$ as a commonly used practical range in which meaningful privacy–utility tradeoffs may still be attainable. By comparison, every matched level in Table 2 is well below 0.15. At these small matched numerical levels, Table 2 shows a substantial difference in prediction utility. Ordinary ridge has population MSE of approximately 0.08–0.11. AdaSSP substantially improves over direct Gaussian output perturbation, but its MSE remains near 1.1, while Gaussian output perturbation has MSE near 5. The comparison therefore illustrates the privacy–utility difficulty of imposing a uniform neighboring-dataset DP guarantee at numerical privacy levels comparable to the small release-specific likelihood ratios observed for ordinary ridge.

6.3 Feature Population Derived from Real-World Data

We construct a smoothed empirical feature population from the WDBC dataset. Selected features are rank-Gaussianized, centered, and whitened. A feature vector is generated by choosing one whitened data vector uniformly and adding $0.05Z$, where $Z \sim \mathcal{N}(0, I_p)$. The response remains semi-synthetic, $Y = X^\top w^\star + \epsilon$; the original diagnostic label is not used as the prediction target.

Figure 2 shows that the broad synthetic pattern persists under the WDBC-derived feature population: mean PMC generally decreases with n, although the trend is less smooth than under isotropic

Table 3: Results under the smoothed empirical WDBC feature population at $\pi_r = 1/2$. WDBC supplies the feature geometry; the response is semi-synthetic.

p	n	Median PMC	90th percentile PMC	Ridge MSE	AdaSSP MSE	Gaussian-DP MSE	Median minESS
10	30	0.0884	0.1626	0.1181	1.0860	4.9198	482.0
10	50	0.0540	0.1212	0.1030	1.0873	5.0986	546.9
10	100	0.0269	0.0899	0.0819	1.1017	5.0175	587.0
20	60	0.0740	0.2624	0.1156	1.0862	5.0810	38.1
20	100	0.0322	0.1036	0.0978	1.0708	5.0350	276.3
20	200	0.0305	0.0708	0.0807	1.0691	5.1620	332.4

Gaussian features. This additional heterogeneity is consistent with the non-Gaussian joint geometry retained after rank-Gaussianization and whitening.

Table 3 gives selected distributional summaries and utility results. The $p = 20, n = 60$ regime has notably smaller minESS and is therefore interpreted more cautiously. In the better-conditioned regimes, ordinary ridge retains substantially lower prediction error than both AdaSSP and Gaussian output perturbation, so the qualitative privacy–utility separation is not eliminated when the feature geometry is derived from real-world measurements.

6.4 Numerical Scaling Limits

The analytical characterization in Section 4 remains valid in higher dimensions, but evaluating the secret-conditioned mixture densities becomes more difficult. At $p = 50$, direct Monte Carlo produces likelihood-ratio estimates that are not stable across independent repetitions, and the alternative sampling procedures we tested do not reliably resolve this concentration. We therefore do not report the corresponding PMC values as substantive privacy measurements. Appendix D.7 gives the supporting numerical results. These findings separate the exact analytical characterization from its numerical evaluation: the former remains valid under the stated model, while high-dimensional mixture-likelihood estimation is a computational limitation of the present work.

7 Discussion and Conclusion

In this paper we study how much privacy may already be present in ordinary ridge regression before privacy-specific randomization is introduced. In regimes where the numerical estimates are reliable, proximity-membership leakage is nonzero but generally small, tends to decrease as n grows relative to p, and varies across released models and protected targets. The same broad behavior persists under a feature population derived from real-world data. At corresponding numerical likelihood-ratio levels, both DP regression baselines incur substantially greater prediction error than ordinary ridge, although PMC and DP provide different guarantees: PMC is specific to a population, protected secret, and realized release, whereas DP provides uniform neighboring-dataset protection.

The main limitation is numerical evaluation of the secret-conditioned mixture likelihoods. Although the underlying output distributions are characterized exactly, their Monte Carlo estimates become increasingly concentrated as the dimension grows, and we therefore do not report higher-dimensional PMC values when repeated likelihood-ratio estimates are unstable. More scalable likelihood-evaluation methods, such as improved adaptive importance sampling or sequential Monte Carlo procedures, are a natural direction for extending the framework to higher dimensions.

More broadly, the results suggest a release-evaluation approach to privacy certification. Given a population model and protected secret, a proposed model release can be evaluated before additional perturbation is introduced. If the resulting release-specific privacy is sufficient for the intended objective, further privatization may be unnecessary for that privacy goal; if uniform neighboring-dataset protection is required, DP remains appropriate.

AI Use Disclosure

In this work, we used generative AI tools to generate synthetic datasets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical

claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results.

Additionally, we used generative AI tools to edit the paper to improve readability, including correcting grammar and improving phrasing. All scientific content, arguments, and section drafts were written by the authors, and all AI-suggested edits were reviewed by the authors before inclusion.

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Ibrahim Issa, Aaron B Wagner, and Sudeep Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2019.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33: 22205–22216, 2020.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1631–1648, 2023.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.

Sara Saeidian, Giulia Cervia, Tobias J Oechtering, and Mikael Skoglund. Pointwise maximal leakage on general alphabets. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 388–393. IEEE, 2023.

Sara Saeidian, Leonhard Grosse, Parastoo Sadeghi, Mikael Skoglund, and Tobias J Oechtering. Information density bounds for privacy. *IEEE Transactions on Information Theory*, 2025.

Or Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pp. 3105–3114. PMLR, 2017.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pp. 861–870. SPIE, 1993.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Mohammad Amin Zarrabian, Ni Ding, and Parastoo Sadeghi. On the lift, related privacy measures, and applications to privacy–utility trade-offs. *Entropy*, 25(4):679, 2023.

ADDITIONAL DERIVATIONS, NUMERICAL METHODS, AND EXPERIMENTS

The appendix provides the derivations and additional results supporting the main paper. We give a detailed formulation of the privacy problem, prove the exact conditional distribution of the ridge estimator, derive the secret-conditioned output distributions, describe the numerical procedures used to evaluate their likelihoods, specify the DP baselines, and report additional experiments on membership priors, target locations, numerical reliability, and higher-dimensional settings.

A PRELIMINARIES AND DETAILED PROBLEM FORMULATION

A.1 NOTATION

For a positive integer m, let $[m] = \{1, \dots, m\}$. We distinguish random quantities from their realizations throughout. Uppercase symbols denote random quantities and lowercase symbols their realizations. Thus, $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ denote a random feature vector and response, while x_i and y_i denote their realizations.

Bold uppercase symbols denote random matrices or stacked random vectors. In particular, $\mathbf{X} \in \mathbb{R}^{n\times p}$ denotes the random feature matrix and $\mathbf{x}$ a realization, while $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ denotes the random response vector and $\mathbf{y}$ its realization. The released ridge estimator is the random vector W, and w denotes a realized release.

The Euclidean norm of a vector is denoted by $\|\cdot\|_2$. We use the same notation for the spectral norm of a matrix when there is no ambiguity. For a symmetric matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalues, and let $\mathrm{Tr}(A)$ denote its trace. We write I_m for the $m \times m$ identity matrix.

For random quantities U and V, let P_U and $P_{U,V}$ denote their marginal and joint distributions. Conditional distributions such as $P_{U|V=v}$ are understood as regular conditional distributions. When densities with respect to the relevant dominating measures exist, we denote them by f_U, $f_{U,V}$, and $f_{U|V}$. All logarithms are natural unless stated otherwise.

A.2 PROXIMITY MEMBERSHIP AS THE PROTECTED SECRET

Fix a public target $z \in \mathbb{R}^p$ and radius $r > 0$, and define $B(z, r) = \{x \in \mathbb{R}^p : \|x - z\|_2 \le r\}$. The protected binary secret is

$$S_r = \mathbf{1}\left\{\bigcup_{i=1}^{n}\{X_i \in B(z,r)\}\right\} = \mathbf{1}\left\{\min_{i\in[n]} \|X_i - z\|_2 \le r\right\}.$$

Thus, $S_r = 1$ means that the training dataset contains at least one feature vector within distance r of the public target. For a continuously distributed feature population, exact equality with a fixed target has probability zero, so a positive-radius neighborhood gives a nondegenerate membership event.

For a single feature vector, let $a_r(z) = \Pr\{X_i \in B(z, r)\}$. Under the Gaussian feature population, we have $\|X_i - z\|_2^2 \sim \chi^2_p(\|z\|_2^2)$, and therefore $a_r(z) = F_{\chi^2_p(\|z\|_2^2)}(r^2)$. Since the feature vectors are i.i.d., the prior probability of proximity membership is

$$\pi_r(z) = \Pr(S_r = 1) = 1 - \big(1 - a_r(z)\big)^n.$$

When z is fixed, we suppress this dependence and write a_r and π_r.

Appendix D.3 varies the membership prior π_r. In the target-location experiments of Appendix D.6, we instead recalibrate r for each target so that π_r remains fixed. This separates the effect of the protected target from changes in the prior probability of membership.

The privacy question is how much observing the released model W changes what an adversary can infer about S_r. Although the events $\{X_i \in B(z, r)\}$ are independent before observing W, they generally become dependent after conditioning on the common released model. Consequently, the posterior probability of S_r cannot in general be obtained by treating the training records independently after observing W.

A.3 Pointwise Maximal Cost

Let S be a finite-valued secret and W an observable random output. For values with positive probability or density, the information density is

$$\imath_{S;W}(s;w) = \log \frac{dP_{S,W}}{d(P_S \times P_W)}(s,w) = \log \frac{\Pr(S=s \mid W=w)}{\Pr(S=s)}.$$

Pointwise maximal cost (PMC) at a realized output w is (Saeidian et al., 2025)

$$\Lambda_{P_{S,W}}(S \to w) = -\min_s \imath_{S;W}(s;w) = \log \max_s \frac{\Pr(S=s)}{\Pr(S=s \mid W=w)}.$$

Crucially, PMC has an operational interpretation. By Saeidian et al. (2025, Theorem 3), it characterizes the largest multiplicative reduction in an adversary's optimal expected nonnegative inference cost after observing $W=w$, relative to the optimal cost achievable from prior knowledge alone. This interpretation is the basis for using PMC as a measure of membership disclosure in this work. PMC is therefore not tied to a particular training algorithm, inference task, decision rule, or loss function.

For the binary proximity-membership secret, let $q_r(w) = \Pr(S_r = 1 \mid W = w)$, and abbreviate $\Lambda_{P_{S_r,W}}(S_r \to w)$ as $\Lambda_r(w)$, i.e.,

$$\Lambda_r(w) = \log \max \left\{ \frac{\pi_r}{q_r(w)}, \frac{1-\pi_r}{1-q_r(w)} \right\}.$$

Equivalently, $\Lambda_r(w) \le \eta$ if and only if

$$e^{-\eta}\pi_r \le q_r(w) \le 1 - e^{-\eta}(1-\pi_r).$$

Hence, a small PMC value means that observing the released model cannot make either value of the protected secret substantially less plausible relative to its prior probability.

For a concrete interpretation, consider the WDBC-derived $p=10$, $n=100$ setting at the balanced prior $\pi_r = 1/2$. The median PMC across released models is 0.0269; see Figure 13. For a release with $\Lambda_r(w) = 0.0269$, the posterior bound above gives

$$\frac{1}{2}e^{-0.0269} \le q_r(w) \le 1 - \frac{1}{2}e^{-0.0269},$$

or numerically,

$$0.4867 \le q_r(w) \le 0.5133.$$

Thus, starting from prior membership probability $1/2$, observing such a release can shift the posterior membership probability only by about 0.0133 in either direction. This illustrates the operational meaning of a small PMC value: the released model provides only limited evidence for changing the adversary's prior belief about the protected secret.

A.4 Relation to Likelihood Ratios and Differential Privacy

The PMC guarantee should be distinguished from a uniform likelihood-ratio guarantee. Define the log-likelihood ratio between the output distributions under membership and non-membership by

$$L_r(w) = \log \frac{f_{W|S_r=1}(w)}{f_{W|S_r=0}(w)}.$$

Bayes' rule gives $L_r(w) = \operatorname{logit} q_r(w) - \operatorname{logit} \pi_r$, where $\operatorname{logit}(u) = \log\{u/(1-u)\}$.

Substituting this relation into the binary PMC expression gives

$$\Lambda_r(w) = \log \max \left\{ \pi_r + (1-\pi_r)e^{-L_r(w)},\ 1 - \pi_r + \pi_r e^{L_r(w)} \right\}.$$

In particular, for the balanced prior $\pi_r = 1/2$, we have

$$\Lambda_r(w) = \log \left(\frac{1 + e^{|L_r(w)|}}{2} \right).$$

In this case, PMC depends only on the magnitude of the log-likelihood ratio.

To compare prediction utility with differentially private regression methods, we place each realized PMC value on a symmetric likelihood-ratio scale. For a fixed prior π_r, suppose that a log-likelihood ratio satisfies $|L_r(w)| \leq \varepsilon$. The largest PMC compatible with this symmetric bound is

$$\log\left(\min\{\pi_r, 1-\pi_r\} + \max\{\pi_r, 1-\pi_r\}e^{\varepsilon}\right).$$

We therefore define $\varepsilon_{\text{match}}(w)$ by equating this worst-case value to the observed $\Lambda_r(w)$, which gives

$$\varepsilon_{\text{match}}(w) = \log \frac{e^{\Lambda_r(w)} - \min\{\pi_r, 1-\pi_r\}}{\max\{\pi_r, 1-\pi_r\}}.$$

Thus, $\varepsilon_{\text{match}}(w)$ is the smallest symmetric likelihood-ratio radius whose worst-case PMC equals the observed PMC of the release. For the balanced prior, this simplifies to $\varepsilon_{\text{match}}(w) = |L_r(w)|$.

We use $\varepsilon_{\text{match}}$ only to establish a common numerical scale for the privacy–utility experiments. It does not imply that ordinary ridge satisfies differential privacy. PMC concerns disclosure about a specified secret under an assumed population and for a particular realized release, whereas DP controls changes in the output distribution uniformly over neighboring datasets.

A.5 Release-Evaluation Setting

We interpret PMC as a tool for evaluating the *inherent privacy of a realized model* before release. This is distinct from evaluating the privacy of the learning procedure itself. The latter asks whether an algorithm provides a privacy guarantee uniformly over possible training datasets and outputs, as in differential privacy. Here, instead, a trusted gatekeeper operating inside a protected computing environment evaluates how much a particular proposed release reveals about a specified secret. The gatekeeper has access to the training dataset, the learning procedure, and the information needed to specify the privacy question, including the population model, protected secret, and public parameters defining that secret. The external recipient of the model does not require access to the training dataset; the gatekeeper's role is to decide whether the proposed model release satisfies a specified privacy criterion before the model leaves the protected environment.

For the proximity-membership problem studied here, the privacy specification consists of the population model, public target z, radius r, and prior probability π_r. We assume that a potential external observer knows this public information and the learning rule, and would observe the released coefficients w. Accordingly, quantities $q_r(w)$ and $\Lambda_r(w)$ measure disclosure about S_r against an observer with access to all of this public information.

A possible release procedure is therefore as follows. First, the data holder trains the ordinary model inside the protected environment. Second, the gatekeeper evaluates the proposed release with respect to the specified secret and population model. If the resulting disclosure is below the release criterion, the model may be approved without adding privacy-specific randomization for that particular privacy objective. If the criterion is not met, the release can instead be rejected, retrained, modified, or passed through an additional privacy mechanism before being reconsidered. Thus, the privacy decision is made at the level of the proposed release rather than by requiring in advance that every trained model be produced by a fixed privacy mechanism.

For ridge regression, the secret-conditioned output distributions needed for this evaluation can be characterized analytically, as shown in Appendix B. For more general learning procedures, the same gatekeeping principle could in principle be applied if the relevant secret-conditioned output distributions can be characterized or estimated with sufficient reliability.

This model-level release evaluation provides a different and more specific guarantee than algorithm-level differential privacy. A PMC value certifies disclosure about the specified secret under the assumed population model and for the particular proposed release; it does not provide secret-uniform or neighboring-dataset-uniform protection. If the application instead requires such uniform protection, then a dedicated privacy mechanism such as DP remains appropriate.

B Exact Inherent Privacy of Ridge Regression

Theorem 1 characterizes proximity-membership PMC through the two secret-conditioned output densities $f_{W|S_r=0}$ and $f_{W|S_r=1}$. This section derives these densities and the resulting PMC expres-

sion. We first characterize the released ridge coefficients conditional on the complete feature matrix, then derive the feature-matrix distributions under the two values of the proximity-membership secret, and finally average the conditional output density over these distributions to obtain the densities g_0 and g_1 and compute PMC.

Throughout this section, assume $n \geq p$ and that the feature matrix $\mathbf{X}$ has full column rank almost surely. The principal analytical model is

$$X_i \sim \mathcal{N}(0, I_p), \qquad Y_i = X_i^\top w^\star + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2),$$

with the observation-noise variables independent of the features. Let $\lambda > 0$ denote the ridge regularization parameter. For a realized feature matrix $\mathbf{x}$, define $S_\mathbf{x} = \mathbf{x}^\top \mathbf{x}/n$ and $H_\mathbf{x} = S_\mathbf{x} + \lambda I_p$.

B.1 Conditional Distribution of the Released Ridge Estimator

Proposition 1 *Under the linear response model above,*

$$W \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\mu_\mathbf{x}, \Sigma_\mathbf{x}),$$

where

$$\mu_\mathbf{x} = H_\mathbf{x}^{-1} S_\mathbf{x} w^\star = \left(I_p - \lambda H_\mathbf{x}^{-1}\right) w^\star, \qquad \Sigma_\mathbf{x} = \frac{\sigma_\epsilon^2}{n} H_\mathbf{x}^{-1} S_\mathbf{x} H_\mathbf{x}^{-1}.$$

Proof of Proposition 1. Substituting $\mathbf{Y} = \mathbf{X}w^\star + \boldsymbol{\epsilon}$ into the ridge estimator gives

$$\begin{aligned} W &= H_\mathbf{X}^{-1} \frac{1}{n} \mathbf{X}^\top \left(\mathbf{X} w^\star + \boldsymbol{\epsilon}\right) \\ &= H_\mathbf{X}^{-1} S_\mathbf{X} w^\star + H_\mathbf{X}^{-1} \frac{1}{n} \mathbf{X}^\top \boldsymbol{\epsilon}, \end{aligned} \tag{7}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 I_n)$.

Conditional on $\mathbf{X} = \mathbf{x}$, the first term on the right-hand side of (7) is deterministic, and the second term on the right-hand side of (7) is a linear transformation of a zero-mean Gaussian vector. Hence $W \mid \mathbf{X} = \mathbf{x}$ is Gaussian with mean $H_\mathbf{x}^{-1} S_\mathbf{x} w^\star$ and covariance

$$\begin{aligned} \mathrm{Cov}(W \mid \mathbf{X} = \mathbf{x}) &= H_\mathbf{x}^{-1} \frac{1}{n} \mathbf{x}^\top \left(\sigma_\epsilon^2 I_n\right) \frac{1}{n} \mathbf{x} H_\mathbf{x}^{-1} \\ &= \frac{\sigma_\epsilon^2}{n} H_\mathbf{x}^{-1} S_\mathbf{x} H_\mathbf{x}^{-1}. \end{aligned}$$

Finally, $H_\mathbf{x}^{-1} S_\mathbf{x} = I_p - \lambda H_\mathbf{x}^{-1}$, which provides the alternative expression for the mean. □

Since $\mathbf{x}$ has full column rank, it follows that both $S_\mathbf{x}$ and $\Sigma_\mathbf{x}$ are positive definite. Therefore $W \mid \mathbf{X} = \mathbf{x}$ has the Gaussian density

$$g_\mathbf{x}(w) = \frac{\exp\left[-\frac{1}{2}(w - \mu_\mathbf{x})^\top \Sigma_\mathbf{x}^{-1} (w - \mu_\mathbf{x})\right]}{(2\pi)^{p/2} \det(\Sigma_\mathbf{x})^{1/2}}. \tag{8}$$

The Gaussian feature assumption is not needed for the conditional Gaussian distribution in Proposition 1. Conditional Gaussianity follows from the linear response model and Gaussian observation noise. The feature population instead determines the distribution over which $g_\mathbf{X}(w)$ is averaged to obtain the secret-conditioned output densities.

B.2 Feature-Matrix Distribution Under the Secret

Let $B = B(z, r)$ and define $K_r = \sum_{i=1}^n \mathbf{1}\{X_i \in B\}$. Then $S_r = \mathbf{1}\{K_r \geq 1\}$ and, before conditioning on the secret, $K_r \sim \mathrm{Binomial}(n, a_r)$.

The conditional feature-matrix distributions can be generated as follows. When $S_r = 0$, draw all n rows independently from $P_X(\cdot \mid X \notin B)$. When $S_r = 1$, first draw $K_r \sim \mathrm{Binomial}(n, a_r) \mid K_r \geq 1$. Conditional on $K_r = k$, choose a subset $J \subset [n]$ of size k uniformly among the $\binom{n}{k}$ possible subsets. For $i \in J$, draw X_i independently from $P_X(\cdot \mid X \in B)$; for $i \notin J$, draw X_i independently from $P_X(\cdot \mid X \notin B)$. These conditional feature-matrix distributions are used directly in the Monte Carlo evaluation below.

Proposition 2 *The procedures above generate exactly $P_{\mathbf{X} \mid S_r = 0}$ and $P_{\mathbf{X} \mid S_r = 1}$, respectively.*

Proof of Proposition 2. Let $I_i = \mathbf{1}\{X_i \in B\}$. Since the feature vectors are i.i.d., it follows that $I_1, \ldots, I_n$ are i.i.d. Bernoulli(a_r) and $K_r = \sum_{i=1}^n I_i$.

First consider $S_r = 0$. This event is equivalent to $I_i = 0$ for every i, or $\bigcap_{i=1}^n \{X_i \notin B\}$. For measurable sets $A_1, \ldots, A_n$,

$$\Pr\left(\bigcap_{i=1}^n \{X_i \in A_i\} \,\middle|\, S_r = 0\right) = \frac{\prod_{i=1}^n \Pr(X_i \in A_i \cap B^c)}{\prod_{i=1}^n \Pr(X_i \in B^c)} = \prod_{i=1}^n \Pr(X_i \in A_i \mid X_i \notin B).$$

Hence conditioning on $S_r = 0$ preserves independence across rows, with each row distributed according to $P_X(\cdot \mid X \notin B)$. This proves the stated construction for $P_{\mathbf{X}|S_r=0}$.

Now consider $S_r = 1$, which is equivalent to $K_r \geq 1$. Since $K_r \sim \text{Binomial}(n, a_r)$ before conditioning, its conditional distribution is $\text{Binomial}(n, a_r) \mid K_r \geq 1$. Conditional on $K_r = k$, every binary indicator vector $(I_1, \ldots, I_n)$ containing exactly k ones has the same probability, so the set $J = \{i \in [n] : I_i = 1\}$ is uniformly distributed over the $\binom{n}{k}$ subsets of $[n]$ of size k.

Finally, conditional on a fixed indicator vector, the original independence of $X_1, \ldots, X_n$ is preserved. For each i with $I_i = 1$, the conditional distribution is $P_X(\cdot \mid X \in B)$, while for each i with $I_i = 0$ it is $P_X(\cdot \mid X \notin B)$. Combining the conditional distribution of K_r, the uniform distribution of J given K_r, and these row-wise conditional distributions gives exactly $P_{\mathbf{X}|S_r=1}$. □

The argument uses only that the feature vectors are i.i.d. and therefore applies beyond the Gaussian feature population whenever the two conditional feature distributions can be generated or otherwise evaluated.

B.3 Output Distributions Under Membership and Non-Membership

For $s \in \{0, 1\}$, let $g_s(w) = f_{W|S_r=s}(w)$.

Lemma 1 *For each $s \in \{0, 1\}$,*

$$g_s(w) = \mathbb{E}[g_{\mathbf{X}}(w) \mid S_r = s], \tag{9}$$

where $g_{\mathbf{x}}(w)$ is the conditional Gaussian density defined in (8)*.*

Proof of Lemma 1. Since S_r is a deterministic function of $\mathbf{X}$, it follows that conditioning on S_r provides no additional information about W once $\mathbf{X}$ is known. Hence $f_{W|\mathbf{X},S_r}(w \mid \mathbf{x}, s) = f_{W|\mathbf{X}}(w \mid \mathbf{x}) = g_{\mathbf{x}}(w)$. Therefore,

$$\begin{aligned} f_{W|S_r=s}(w) &= \int f_{W|\mathbf{X},S_r}(w \mid \mathbf{x}, s)\, dP_{\mathbf{X}|S_r=s}(\mathbf{x}) \\ &= \int g_{\mathbf{x}}(w)\, dP_{\mathbf{X}|S_r=s}(\mathbf{x}) \\ &= \mathbb{E}[g_{\mathbf{X}}(w) \mid S_r = s], \end{aligned}$$

which proves the claim. □

Thus, the output distributions under membership and non-membership are mixtures of the conditional Gaussian density (8), averaged under the exact secret-conditioned feature-matrix distributions derived in Section B.2.

B.4 Exact Posterior and PMC

Define $L_r(w) = \log[g_1(w)/g_0(w)]$. We provide a complete proof of Theorem 1.

Proof of Theorem 1. Since $L_r(w) = \log \frac{g_1(w)}{g_0(w)}$, it follows that $g_1(w) = e^{L_r(w)} g_0(w)$. By Bayes' rule, we have that

$$\begin{aligned} q_r(w) &= \Pr(S_r = 1 \mid W = w) \\ &= \frac{\Pr(S_r = 1) f_{W|S_r=1}(w)}{\Pr(S_r = 1) f_{W|S_r=1}(w) + \Pr(S_r = 0) f_{W|S_r=0}(w)} \\ &= \frac{\pi_r g_1(w)}{\pi_r g_1(w) + (1 - \pi_r) g_0(w)}. \end{aligned}$$

Substituting $g_1(w) = e^{L_r(w)} g_0(w)$ and cancelling the common factor $g_0(w) > 0$ gives

$$q_r(w) = \frac{\pi_r e^{L_r(w)}}{1 - \pi_r + \pi_r e^{L_r(w)}}. \tag{10}$$

Recall from (3) that $\Lambda_r(w) = \log \max \left\{ \frac{\pi_r}{q_r(w)}, \frac{1-\pi_r}{1-q_r(w)} \right\}$. From (10) we have

$$\begin{aligned} \frac{\pi_r}{q_r(w)} &= \pi_r \frac{1 - \pi_r + \pi_r e^{L_r(w)}}{\pi_r e^{L_r(w)}} \\ &= \frac{1 - \pi_r + \pi_r e^{L_r(w)}}{e^{L_r(w)}} \\ &= \pi_r + (1 - \pi_r) e^{-L_r(w)}. \end{aligned}$$

We also have

$$1 - q_r(w) = 1 - \frac{\pi_r e^{L_r(w)}}{1 - \pi_r + \pi_r e^{L_r(w)}} = \frac{1 - \pi_r}{1 - \pi_r + \pi_r e^{L_r(w)}},$$

which means that $\frac{1-\pi_r}{1-q_r(w)} = 1 - \pi_r + \pi_r e^{L_r(w)}$. This completes the proof. □

For the balanced prior, $\Lambda_r(w) = \log[(1 + e^{|L_r(w)|})/2]$.

Theorem 1 reduces the privacy calculation to the two output densities $g_0(w)$ and $g_1(w)$. Once these densities are known, the posterior probability and PMC follow exactly. Numerical approximation is required only to evaluate the mixture expectations in (9).

B.5 Direct Monte Carlo Evaluation

Theorem 1 reduces the privacy calculation for a realized release w to the two mixture densities $g_0(w)$ and $g_1(w)$. Our primary numerical procedure evaluates these densities by direct Monte Carlo sampling from the exact secret-conditioned feature-matrix distributions derived in Section B.2.

The computation proceeds as follows. For each secret value $s \in \{0, 1\}$, we draw feature matrices from $P_{\mathbf{X}|S_r=s}$, evaluate the conditional Gaussian density $g_{\mathbf{x}}(w)$ from (8) at each realized matrix, and average these values to estimate $g_s(w)$. The two estimated densities determine the estimated log-likelihood ratio, which is then substituted into Theorem 1 to obtain the estimated PMC. The empirical effective contribution count introduced in Section B.6 is reported only as a numerical concentration measure. Section D.7 subsequently shows how direct Monte Carlo can become computationally difficult.

More precisely, for each $s \in \{0, 1\}$, let $\mathbf{X}_s^{(1)}, \ldots, \mathbf{X}_s^{(M)}$ be independent random feature matrices distributed according to $P_{\mathbf{X}|S_r=s}$, and let $\mathbf{x}_s^{(1)}, \ldots, \mathbf{x}_s^{(M)}$ denote their respective realized values. Using the sampling procedure in Section B.2, we compute

$$\widehat{g}_s(w) = \frac{1}{M} \sum_{m=1}^{M} g_{\mathbf{x}_s^{(m)}}(w).$$

For fixed w, the expectation of $\widehat{g}_s(w)$ over the Monte Carlo draws equals $g_s(w)$, so each mixture-density estimator is unbiased.

Algorithm 1 Monte Carlo Evaluation of Proximity-Membership PMC

Require: Realized ridge release w; target z; radius r; membership prior π_r; Monte Carlo budget M

Ensure: Estimated PMC $\widehat{\Lambda}_r(w)$ and empirical concentration measure $\widehat{\mathrm{minESS}}(w)$

1: **for** $s \in \{0,1\}$ **do**
2: Define $\mathbf{X}_s^{(1)}, \ldots, \mathbf{X}_s^{(M)} \overset{\text{i.i.d.}}{\sim} P_{\mathbf{X}|S_r=s}$ using the exact construction in Section B.2.
3: **for** $m = 1, \ldots, M$ **do**
4: Let $\mathbf{x}_s^{(m)}$ denote the realized feature matrix.
5: Evaluate $g_{\mathbf{x}_s^{(m)}}(w)$ using the conditional Gaussian density in (8).
6: **end for**
7: Compute $\widehat{g}_s(w) = \frac{1}{M}\sum_{m=1}^{M} g_{\mathbf{x}_s^{(m)}}(w)$.
8: Compute $\widehat{\mathrm{ESS}}_s(w)$ from (11).
9: **end for**
10: Compute $\widehat{L}_r(w) = \log \widehat{g}_1(w) - \log \widehat{g}_0(w)$.
11: Compute $\widehat{\Lambda}_r(w)$ by substituting $\widehat{L}_r(w)$ into Theorem 1.
12: Set $\widehat{\mathrm{minESS}}(w) = \min\{\widehat{\mathrm{ESS}}_0(w), \widehat{\mathrm{ESS}}_1(w)\}$.
13: **return** $\widehat{\Lambda}_r(w)$ and $\widehat{\mathrm{minESS}}(w)$.

We then form $\widehat{L}_r(w) = \log \widehat{g}_1(w) - \log \widehat{g}_0(w)$ and substitute $\widehat{L}_r(w)$ into the expression in Theorem 1 to obtain $\widehat{\Lambda}_r(w)$. Although $\widehat{g}_0(w)$ and $\widehat{g}_1(w)$ are unbiased, $\widehat{L}_r(w)$ and $\widehat{\Lambda}_r(w)$ are not generally unbiased, since they are nonlinear functions of the two mixture-density estimates.

All likelihood calculations are performed on the logarithmic scale. Writing $\ell_{s,m}(w) = \log g_{\mathbf{x}_s^{(m)}}(w)$, we evaluate

$$\log \widehat{g}_s(w) = \mathrm{LSE}\big(\ell_{s,1}(w), \ldots, \ell_{s,M}(w)\big) - \log M,$$

where LSE denotes log-sum-exp.

Algorithm 1 summarizes the numerical procedure used to evaluate PMC for a fixed realized ridge release. The only approximation is the Monte Carlo evaluation of the two secret-conditioned mixture densities in line 7; once these are estimated, the likelihood ratio and PMC are obtained directly from the exact characterization in Theorem 1.

B.6 Concentration of the Monte Carlo Contributions

The likelihood terms $g_{\mathbf{x}_s^{(m)}}(w)$ entering the Monte Carlo estimate of $g_s(w)$ can be highly uneven. In particular, a small number of sampled feature matrices may account for most of the estimated mixture density, while many others contribute negligibly. In that situation, a nominal budget of M feature-matrix draws can behave as though substantially fewer draws were informative, which can lead to appreciable Monte Carlo variability.

For a realized set of Monte Carlo draws, define the normalized likelihood contributions

$$\widehat{\omega}_{s,m}(w) = \frac{g_{\mathbf{x}_s^{(m)}}(w)}{\sum_{j=1}^{M} g_{\mathbf{x}_s^{(j)}}(w)}, \qquad m = 1, \ldots, M.$$

Since each conditional density is positive, it follows that $\widehat{\boldsymbol{\omega}}_s(w) = (\widehat{\omega}_{s,1}(w), \ldots, \widehat{\omega}_{s,M}(w))$ is a probability vector: its entries are nonnegative and sum to one. It can therefore be viewed as describing how the total estimated mixture density is allocated among the M Monte Carlo draws. A nearly uniform vector means that many feature matrices contribute comparably, whereas a vector concentrated on a few coordinates means only a small number of draws account for most of the estimate.

We summarize their concentration using the empirical effective contribution count

$$\widehat{\mathrm{ESS}}_s(w) = \frac{1}{\sum_{m=1}^{M} \widehat{\omega}_{s,m}(w)^2} = \frac{\left(\sum_{m=1}^{M} g_{\mathbf{x}_s^{(m)}}(w)\right)^2}{\sum_{m=1}^{M} g_{\mathbf{x}_s^{(m)}}(w)^2}. \tag{11}$$

This inverse-squared-norm form gives a simple interpretation. If the likelihood mass is spread uniformly across all M draws, then $\widehat{\omega}_{s,m}(w) = 1/M$ for every m and $\widehat{\mathrm{ESS}}_s(w) = M$. More generally, if the likelihood mass is divided equally among exactly k draws and negligible elsewhere, then $\widehat{\mathrm{ESS}}_s(w) = k$. At the opposite extreme, if one draw dominates the mixture estimate, the value approaches one. Thus, the quantity $\widehat{\mathrm{ESS}}_s(w)$ can be interpreted as the effective number of feature-matrix draws contributing materially to $\widehat{g}_s(w)$.

Since both secret-conditioned densities are needed to form the likelihood ratio, we also report the more conservative quantity $\widehat{\mathrm{minESS}}(w) = \min\{\widehat{\mathrm{ESS}}_0(w), \widehat{\mathrm{ESS}}_1(w)\}$. This quantity should not be interpreted as a confidence interval, a standard error, or a finite-simulation guarantee for PMC. We use it only as a numerical warning sign: a small value indicates that the mixture-density estimate is being determined by relatively few feature-matrix draws. In such regimes, we also examine stability across independent Monte Carlo repetitions. In the main synthetic regimes at $\pi_r = 1/2$, the median minESS values are generally in the hundreds or thousands, with the smallest reported value occurring in the more difficult $p = 20$ regime. The WDBC experiments are more heterogeneous and produce smaller effective contribution counts, especially for $p = 20$. We therefore interpret regimes with low minESS more cautiously and, when necessary, examine stability across independent Monte Carlo repetitions. The numerical scaling experiments in Appendix D.7 further illustrate how this concentration becomes a limiting factor as dimension increases.

C Comparison with Differential Privacy

Differential privacy (DP) is the typical framework for controlling how much the output distribution of an algorithm can change when one training record is replaced. In particular, pure ε-DP can be interpreted as a uniform bound on output likelihood ratios over all neighboring datasets, while (ε, δ)-DP relaxes this requirement by allowing a small exceptional probability (Dwork et al., 2006; Dwork & Roth, 2014).

This makes DP a natural comparison point for our likelihood-ratio characterization of proximity-membership privacy. The guarantees are nevertheless different. PMC measures disclosure about a specified secret under an assumed population model and for a particular realized release, whereas DP provides a uniform neighboring-dataset guarantee for the learning algorithm. Thus, a small value of $\Lambda_r(w)$ does not imply that ordinary ridge regression satisfies DP.

Our goal is therefore not to equate the two guarantees. Instead, for each ordinary-ridge release, we map its PMC value to a corresponding symmetric likelihood-ratio level and evaluate DP regression methods at that same numerical value. This gives a common numerical scale for comparing prediction utility while preserving the distinction between the underlying privacy notions.

C.1 Matched Numerical Privacy Level

Recall from Theorem 1 that

$$\Lambda_r(w) = \log \max\left\{\pi_r + (1-\pi_r)e^{-L_r(w)},\ 1 - \pi_r + \pi_r e^{L_r(w)}\right\},$$

where $L_r(w) = \log[g_1(w)/g_0(w)]$ is the log-likelihood ratio between the released model distributions under membership and non-membership.

To place an observed PMC value on a symmetric likelihood-ratio scale, consider the requirement $|L_r(w)| \le u$. For fixed prior π_r, the largest PMC compatible with this requirement is

$$\log\left(\min\{\pi_r, 1-\pi_r\} + \max\{\pi_r, 1-\pi_r\}e^u\right).$$

Indeed, the maximum over $L_r(w) \in [-u, u]$ is attained at the endpoint that multiplies the larger of π_r and $1 - \pi_r$ by e^u.

We define $\varepsilon_{\mathrm{match}}(w)$ as the value of u for which this worst-case PMC equals the observed $\Lambda_r(w)$. Solving gives

$$\varepsilon_{\mathrm{match}}(w) = \log \frac{e^{\Lambda_r(w)} - \min\{\pi_r, 1-\pi_r\}}{\max\{\pi_r, 1-\pi_r\}}. \tag{12}$$

Thus, $\varepsilon_{\mathrm{match}}(w)$ is the smallest symmetric likelihood-ratio radius whose worst-case PMC equals the PMC of the realized ordinary-ridge release.

For the balanced prior $\pi_r = 1/2$, we have,

$$\varepsilon_{\mathrm{match}}(w) = \log\Big(2e^{\Lambda_r(w)} - 1\Big) = |L_r(w)|.$$

We compute $\varepsilon_{\mathrm{match}}$ separately for each independently trained ordinary-ridge release, so the matched numerical level varies across releases together with PMC.

The matching in (12) is only numerical. In particular, it does not imply that ordinary ridge regression satisfies $(\varepsilon_{\mathrm{match}}, \delta_{DP})$-DP. It allows us to ask instead: what prediction error is incurred by DP regression methods when their privacy parameter is set to the same numerical likelihood-ratio level as the observed PMC of a particular ordinary-ridge release?

C.2 Gaussian Output Perturbation

Our first DP baseline is Gaussian output perturbation. The ordinary-ridge analysis permits unbounded feature vectors and responses, so the unconstrained ridge estimator does not have a finite global sensitivity over the full data domain. For this DP baseline, we therefore clip the features and responses and constrain the coefficient vector.

For each training record (X_i, Y_i), define

$$\widetilde{X}_i = X_i \min\left\{1, \frac{R_x}{\|X_i\|_2}\right\}, \qquad \widetilde{Y}_i = \mathrm{sign}(Y_i)\min\{|Y_i|, B_y\}.$$

Hence $\|\widetilde{X}_i\|_2 \le R_x$ and $|\widetilde{Y}_i| \le B_y$. Let $\widetilde{D} = \{(\widetilde{X}_i, \widetilde{Y}_i)\}_{i=1}^n$ denote the clipped training dataset. We define the constrained ridge estimator

$$\widehat{w}_{\mathrm{clip}}(\widetilde{D}) = \arg\min_{\|w\|_2 \le B_w} \left\{ \frac{1}{2n}\sum_{i=1}^n \left(\widetilde{Y}_i - \widetilde{X}_i^\top w\right)^2 + \frac{\lambda}{2}\|w\|_2^2 \right\}.$$

We next bound the global ℓ_2 sensitivity under replacement adjacency. For a single clipped record (x, y), the gradient contribution of the squared-loss term is $x(x^\top w - y)$. For every feasible $\|w\|_2 \le B_w$,

$$\|x(x^\top w - y)\|_2 \le \|x\|_2\left(|x^\top w| + |y|\right) \le R_x(R_x B_w + B_y).$$

If neighboring clipped datasets $\widetilde{D}$ and $\widetilde{D}'$ differ in one training record, replacement can remove one bounded gradient contribution and replace it by another. Therefore, for every feasible w,

$$\left\|\nabla F_{\widetilde{D}}(w) - \nabla F_{\widetilde{D}'}(w)\right\|_2 \le \frac{2R_x(R_x B_w + B_y)}{n}.$$

The factor 2 arises from replacement adjacency: the two differing per-record gradients can point in opposite directions.

Since the constrained objectives are λ-strongly convex over the same convex feasible set, the standard stability bound for strongly convex empirical risk minimization gives

$$\left\|\widehat{w}_{\mathrm{clip}}(\widetilde{D}) - \widehat{w}_{\mathrm{clip}}(\widetilde{D}')\right\|_2 \le \frac{1}{\lambda}\sup_{\|w\|_2 \le B_w}\left\|\nabla F_{\widetilde{D}}(w) - \nabla F_{\widetilde{D}'}(w)\right\|_2.$$

Thus, the global ℓ_2 sensitivity is bounded by

$$\Delta_2 = \frac{2R_x(R_x B_w + B_y)}{n\lambda}.$$

No additional factor of $\sqrt{p}$ is required: the bound is directly on the Euclidean norm of the full coefficient vector.

For privacy parameters $(\varepsilon_{\mathrm{match}}(w), \delta_{\mathrm{DP}})$, the Gaussian output-perturbation baseline releases

$$\widetilde{W}_G = \widehat{w}_{\mathrm{clip}}(\widetilde{D}) + \xi, \qquad \xi \sim \mathcal{N}\left(0, \sigma_{\mathrm{DP}}(w)^2 I_p\right),$$

where

$$\sigma_{\mathrm{DP}}(w) = \Delta_2 \frac{\sqrt{2\log(1.25/\delta_{\mathrm{DP}})}}{\varepsilon_{\mathrm{match}}(w)}.$$

This is the standard Gaussian-mechanism calibration associated with the global sensitivity bound above (Dwork & Roth, 2014).

The resulting $\sigma_{\mathrm{DP}}(w)$ should not be interpreted as the minimum amount of randomization required for differentially private regression. It is the noise standard deviation prescribed by this particular sensitivity-based Gaussian output-perturbation mechanism at the matched numerical privacy level.

C.3 AdaSSP

Direct output perturbation is transparent but can have poor prediction utility. We therefore also evaluate AdaSSP (Wang, 2018), a method developed specifically for differentially private linear regression. AdaSSP perturbs regression sufficient statistics and uses privately estimated spectral information to adapt the amount of regularization.

Unlike the Gaussian output-perturbation baseline, AdaSSP does not simply add isotropic noise to the final coefficient vector. It perturbs quantities derived from $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top\mathbf{Y}$ and uses private spectral information in constructing the estimator. There is therefore no single coefficient-noise standard deviation directly analogous to $\sigma_{\mathrm{DP}}(w)$.

For training replicate b, let $W^{(b)}$ denote the corresponding ordinary-ridge release. We compute $\varepsilon^{(b)}_{\mathrm{match}} = \varepsilon_{\mathrm{match}}(W^{(b)})$ and run AdaSSP with privacy parameters $(\varepsilon^{(b)}_{\mathrm{match}}, \delta_{\mathrm{DP}})$. Thus, the comparison is paired at the level of the training dataset and its corresponding numerical privacy level.

The two DP baselines play complementary roles. Gaussian output perturbation makes the effect of a simple global-sensitivity mechanism explicit in coefficient space. AdaSSP provides a regression-specific DP baseline for testing whether the observed prediction-utility difference persists under a method designed specifically for private linear regression.

C.4 Implementation of the Paired Comparison

For each experimental replicate, we generate one training dataset and fit ordinary ridge regression to obtain a release $W^{(b)}$. We estimate the PMC of that release and compute $\varepsilon^{(b)}_{\mathrm{match}}$ using (12). The Gaussian output-perturbation and AdaSSP estimators for the same replicate are then trained using the same underlying training dataset and the same replicate-specific numerical privacy level $\varepsilon^{(b)}_{\mathrm{match}}$.

This pairing reduces variation due solely to differences among training datasets when comparing prediction error across the three estimators. Prediction error is evaluated using the same population-risk criterion for ordinary ridge, Gaussian output perturbation, and AdaSSP.

The comparison does not make their privacy guarantees equivalent. Ordinary ridge is evaluated through PMC for the specified proximity-membership secret and realized release, whereas the two DP methods satisfy neighboring-dataset guarantees at $(\varepsilon^{(b)}_{\mathrm{match}}, \delta_{\mathrm{DP}})$. The clipping bounds, value of δ_{DP}, AdaSSP privacy-budget allocation, and other implementation details are given in Section D.1.

D Additional Experiments

This section gives the full experimental setup and additional results supporting Section 6. We examine how proximity-membership leakage varies across independently trained ordinary ridge models, how prediction utility compares with DP regression at corresponding numerical privacy levels, whether the same behavior persists under a feature population derived from real data, and how the protected target, membership prior, and dimension affect the numerical privacy estimates.

D.1 Experimental Setup and Evaluation Protocol

We first consider the Gaussian linear model

$$X_i \sim \mathcal{N}(0, I_p), \qquad Y_i = X_i^\top w^\star + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2),$$

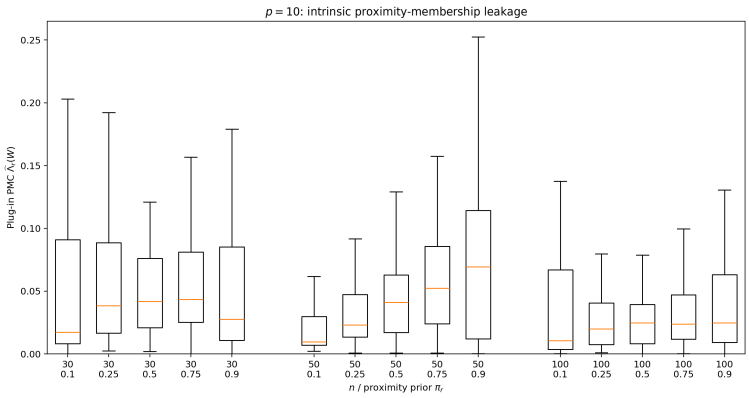


Figure 3: Proximity-membership PMC across independently trained ordinary-ridge models for $p = 10$. Boxplots show $\widehat{\Lambda}_r(W^{(b)})$ over 60 released models for $n \in \{30, 50, 100\}$, with $M = 10{,}000$ secret-conditioned feature-matrix draws per secret value.

with $w^\star = e_1$, $\lambda = 0.1$, and $\sigma_\epsilon = 0.25$. Unless otherwise stated, the protected target is $z = \rho e_1$ with $\rho = 2$.

For each (p, n, π_r), we choose r so that $\Pr(S_r = 1) = \pi_r$. Since $\pi_r = 1 - (1 - a_r)^n$, where $a_r = \Pr(\|X_i - z\|_2 \le r)$, the required single-record probability is $a_r = 1 - (1 - \pi_r)^{1/n}$. For Gaussian features, $\|X_i - z\|_2^2 \sim \chi^2_p(\|z\|_2^2)$, so the corresponding radius r is obtained from $F_{\chi^2_p(\|z\|_2^2)}(r^2) = a_r$.

We study $p = 10$ with $n \in \{30, 50, 100\}$ and $p = 20$ with $n \in \{50, 100, 200\}$. To examine dependence on the membership prior, we use $\pi_r \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. The balanced prior $\pi_r = 1/2$ is the primary reference setting for the summary tables and paired DP comparisons.

For each regime, we independently generate complete training datasets and train ordinary ridge regression. For every released model $W^{(b)}$, we estimate $g_0(W^{(b)})$ and $g_1(W^{(b)})$ using independent feature-matrix draws from the corresponding secret-conditioned distributions, and then compute $L_r(W^{(b)})$, $\Lambda_r(W^{(b)})$, and $\varepsilon_{\text{match}}(W^{(b)})$.

The distribution across b describes variation in disclosure across independently trained released models. This is distinct from the Monte Carlo variation involved in evaluating the privacy quantities for a fixed release.

For the principal synthetic experiments, we use 60 independently trained models in each (p, n, π_r) regime. The secret-conditioned Monte Carlo budget is $M = 10{,}000$ for $p = 10$. For $p = 20$, we use $M = 50{,}000$, 30,000, and 15,000 for $n = 50, 100, 200$, respectively.

To assess concentration of the Monte Carlo averages, we report

$$\text{minESS}(w) = \min\{\text{ESS}_0(w), \text{ESS}_1(w)\},$$

where $\text{ESS}_s(w)$ is the effective number of likelihood contributions defined in Section B.6. This is a numerical concentration measure rather than a finite-simulation confidence guarantee. Unless otherwise stated, all reported PMC values are plug-in Monte Carlo estimates.

For the isotropic Gaussian feature population, the prediction risk of an estimator $\widehat{w}$ is

$$R(\widehat{w}) = \mathbb{E}\left[(Y_{\text{new}} - X_{\text{new}}^\top \widehat{w})^2\right] = \|\widehat{w} - w^\star\|_2^2 + \sigma_\epsilon^2.$$

We therefore evaluate population prediction error directly rather than using a finite test set.

D.2 Inherent Privacy Across Independently Trained Models

We first examine proximity-membership disclosure across independently trained ordinary-ridge releases. Figures 3 and 4 show the distributions of $\widehat{\Lambda}_r(W^{(b)})$ for the balanced prior $\pi_r = 1/2$.

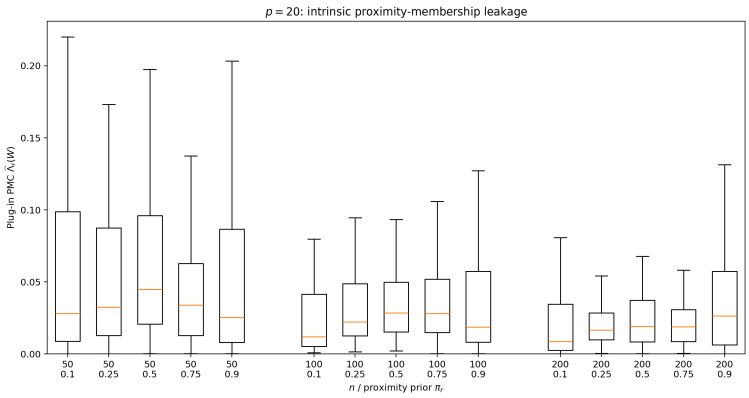


Figure 4: Proximity-membership PMC across independently trained ordinary-ridge models for $p = 20$. Monte Carlo budgets are $M = 50{,}000$, $30{,}000$, and $15{,}000$ for $n = 50, 100, 200$, respectively.

Table 4: Synthetic proximity-membership estimates and numerical concentration measures at $\pi_r = 1/2$. PMC statistics are computed across 60 independently trained ridge models. minESS is the smaller effective number of likelihood contributions for the two secret-conditioned mixture estimates.

p	n	M	Median PMC	q_{90}	q_{95}	Median minESS	q_{10} minESS
10	30	10,000	0.0419	0.1165	0.1236	934.3	125.7
10	50	10,000	0.0411	0.0773	0.0951	1400.3	413.2
10	100	10,000	0.0248	0.0561	0.0752	2145.1	571.9
20	50	50,000	0.0448	0.2011	0.3217	196.8	11.1
20	100	30,000	0.0284	0.1162	0.1439	628.2	49.5
20	200	15,000	0.0190	0.0585	0.0690	847.5	94.5

For $p = 10$, median PMC values are 0.0419, 0.0411, and 0.0248 for $n = 30, 50, 100$, with corresponding 90th percentiles 0.1165, 0.0773, and 0.0561. For $p = 20$, the medians are 0.0448, 0.0284, and 0.0190, while the 90th percentiles are 0.2011, 0.1162, and 0.0585.

Thus, proximity-membership leakage is nonzero but generally small in the regimes where the numerical estimates are reliable, and its typical value tends to decrease as n grows relative to p. The $p = 20, n = 50$ regime is less numerically reliable, so its upper-tail values should be interpreted more cautiously.

The experiments also illustrate the release-specific nature of PMC. Even within a fixed (p, n) regime, independently trained releases can have noticeably different privacy values.

D.3 Dependence on the Membership Prior

PMC depends explicitly on the prior probability of the protected secret, so we repeat the synthetic evaluation for $\pi_r \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, recalibrating r in each regime.

Changing π_r affects both the prior odds of membership and the protected neighborhood through the corresponding radius. The exact expression in (6) accounts for both effects.

These experiments test whether the qualitative low-leakage behavior observed at $\pi_r = 1/2$ persists beyond the balanced-prior setting. We retain $\pi_r = 1/2$ as the primary reference elsewhere since in that case $\varepsilon_{\text{match}} = |L_r|$, which gives the most direct numerical interpretation of the likelihood ratio.

D.4 Privacy–Utility Comparison with Differential Privacy

We next compare prediction utility with the DP regression methods in Section C. For each ordinary-ridge release, we compute its model-specific $\varepsilon_{\text{match}}$ and evaluate both DP methods at that numerical

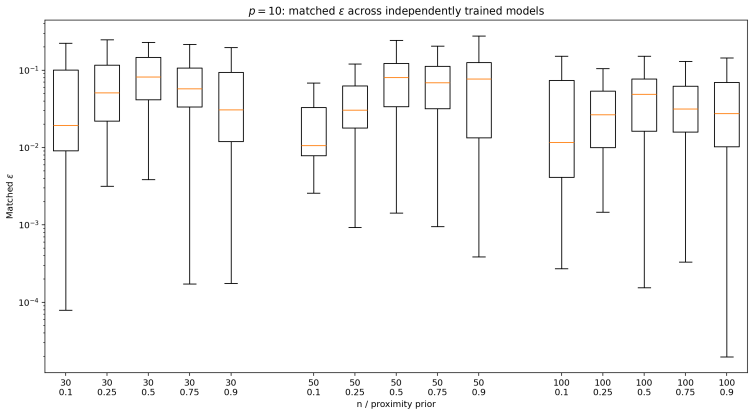


Figure 5: Model-specific $\varepsilon_{\mathrm{match}}$ for the synthetic $p = 10$ experiments.

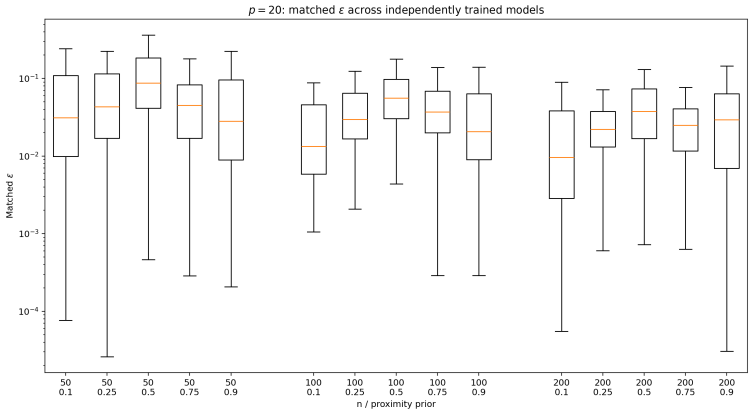


Figure 6: Model-specific $\varepsilon_{\mathrm{match}}$ for the synthetic $p = 20$ experiments.

privacy level. At the balanced prior, $\varepsilon_{\mathrm{match}}(W) = |L_r(W)|$. As emphasized in Section C.1, this matching gives a common numerical likelihood-ratio scale but does not imply that ordinary ridge satisfies differential privacy.

The first baseline is Gaussian output perturbation applied to the clipped and constrained ridge problem in Section C.2. The second is AdaSSP Wang (2018). Within each replicate, ordinary ridge and both DP methods use the same underlying training dataset.

The median matched numerical privacy levels are 0.0821, 0.0806, and 0.0489 for $p = 10$, and 0.0877, 0.0561, and 0.0377 for $p = 20$ as n increases over the corresponding regimes.

D.4.1 Gaussian Output Perturbation

Figures 7 and 8 show the coefficient-noise standard deviation for Gaussian output perturbation.

At $\pi_r = 1/2$, the median coefficient-noise standard deviations are approximately 2329, 1423, and 1174 for $p = 10$, and 2022, 1579, and 1177 for $p = 20$. These values are not lower bounds on the randomization required by DP; they are the noise levels prescribed by the particular sensitivity-based output-perturbation mechanism considered here.

Since σ_{DP} scales as $\Delta_2/\varepsilon_{\mathrm{match}}$, the reduction in sensitivity with increasing n can be partly offset by a simultaneous decrease in the matched numerical privacy level.

Figures 9 and 10 show the resulting prediction errors. For the synthetic model, $\|w^\star\|_2 = 1$, $B_w = 2$, and $\sigma_\epsilon^2 = 0.0625$. After adding Gaussian noise, we project the perturbed coefficient vector

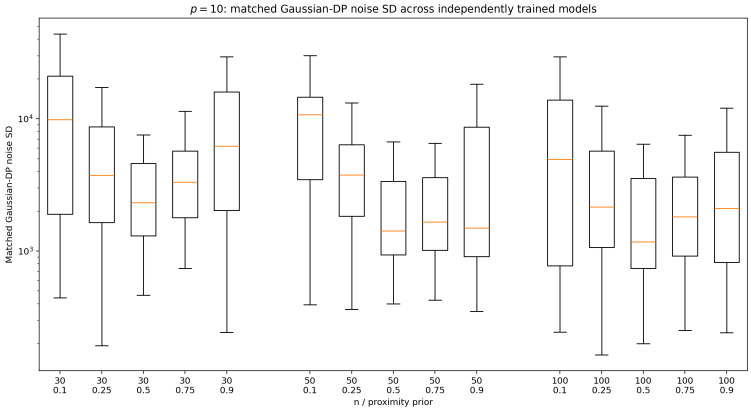


Figure 7: Gaussian output-perturbation noise standard deviation for synthetic $p = 10$, calibrated to the model-specific $\varepsilon_{\text{match}}$.

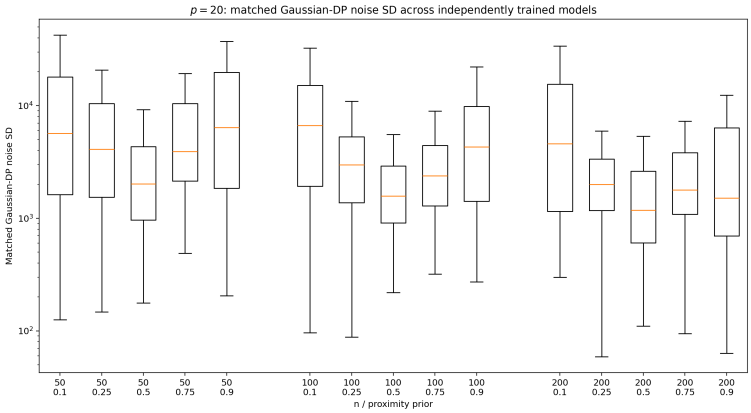


Figure 8: Gaussian output-perturbation noise standard deviation for synthetic $p = 20$, calibrated to the model-specific $\varepsilon_{\text{match}}$.

onto the feasible Euclidean ball $\{w : \|w\|_2 \le B_w\}$. When the added noise is large, this projection often places the released coefficient vector near the boundary $\|w\|_2 = B_w$ with a direction that is approximately uninformative about $w^\star$. In that regime, the prediction risk is naturally on the scale

$$B_w^2 + \|w^\star\|_2^2 + \sigma_\epsilon^2 = 5.0625.$$

The observed Gaussian-DP risks near 5 are consistent with this projection-dominated regime.

D.4.2 AdaSSP

We also evaluate AdaSSP at the same replicate-specific numerical privacy levels. Figures 11 and 12 show the resulting prediction errors.

AdaSSP substantially improves prediction utility relative to direct Gaussian output perturbation. Its median MSE is near 1.1 across the six regimes, rather than near 5. For reference, the zero predictor has prediction risk $R(0) = \|w^\star\|_2^2 + \sigma_\epsilon^2 = 1.0625$. Thus, AdaSSP operates near this reference scale in the regimes studied, whereas ordinary ridge retains much smaller prediction error. This comparison does not imply that AdaSSP converges to the zero predictor; the value 1.0625 is used only as a utility reference.

A clipped nonprivate control shows that clipping itself accounts for little of the observed utility difference: clipped ridge has prediction error close to ordinary ridge across the synthetic regimes.

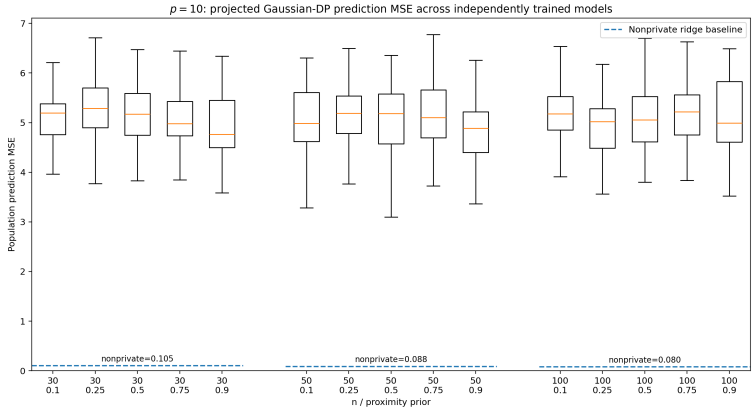


Figure 9: Population prediction MSE for Gaussian output perturbation in the synthetic $p = 10$ experiments.

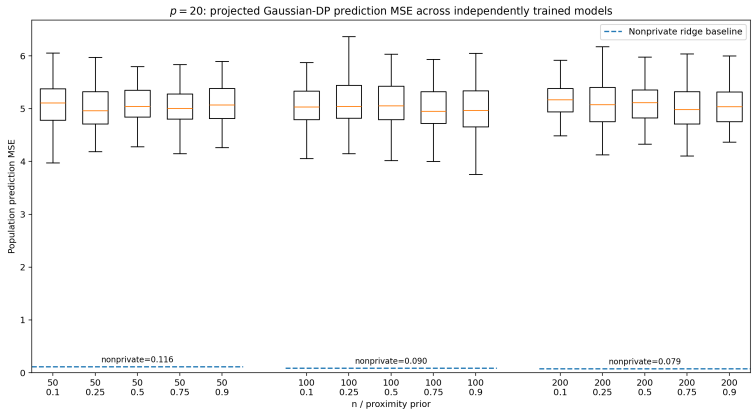


Figure 10: Population prediction MSE for Gaussian output perturbation in the synthetic $p = 20$ experiments.

Table 5 summarizes the paired comparison. Ordinary ridge has MSE between approximately 0.079 and 0.116, while clipping alone changes these values only slightly. Gaussian output perturbation has MSE near 5, and AdaSSP reduces the error to approximately 1.1. Thus, the utility separation persists under the regression-specific DP baseline.

D.5 Feature Population Derived from WDBC

We next replace the independent Gaussian feature population with one derived from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset Street et al. (1993). The experiment remains semi-synthetic: WDBC supplies the feature geometry, while responses are generated from the same linear Gaussian model as above. The original diagnostic label is not used as the prediction target.

Let $\mathbf{X}^{\mathrm{raw}} \in \mathbb{R}^{N\times p}$ denote the selected WDBC feature matrix. For each selected feature, define

$$U_{ij} = \frac{\mathrm{rank}(X_{ij}^{\mathrm{raw}}) - \frac{1}{2}}{N}, \qquad \widetilde{X}_{ij} = \Phi^{-1}(U_{ij}).$$

After centering and whitening the transformed columns, let $\bar{x}_1, \ldots, \bar{x}_N$ denote the resulting feature vectors.

We define the smoothed empirical feature population by

$$X = \bar{x}_J + hZ, \qquad J \sim \mathrm{Unif}\{1, \ldots, N\}, \qquad Z \sim \mathcal{N}(0, I_p), \qquad h = 0.05. \tag{13}$$

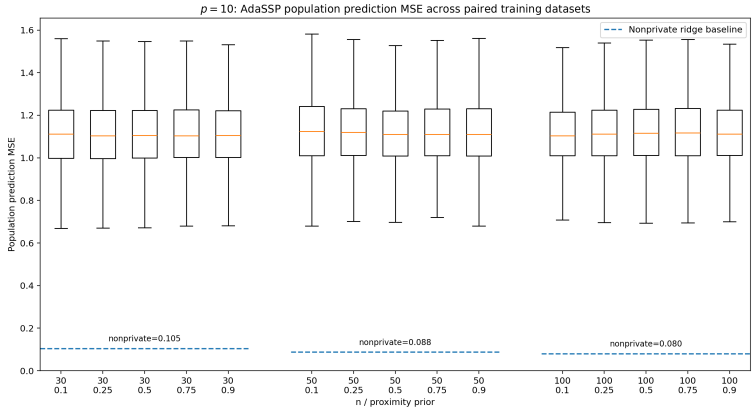


Figure 11: Population prediction MSE for AdaSSP in the synthetic $p = 10$ experiments. AdaSSP uses the same training datasets as ordinary ridge and the corresponding model-specific values of $\varepsilon_{\text{match}}$.

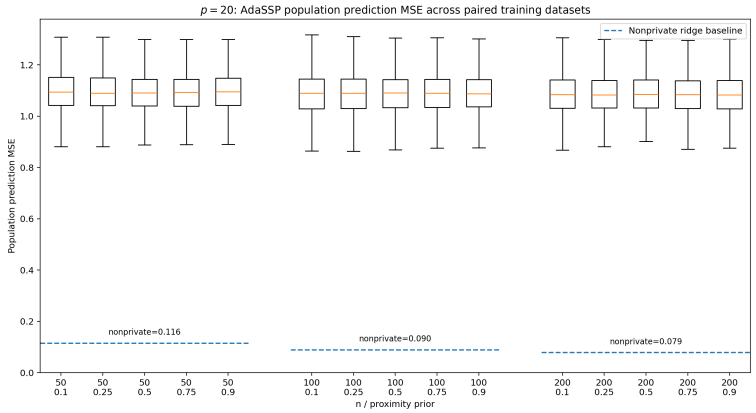


Figure 12: Population prediction MSE for AdaSSP in the synthetic $p = 20$ experiments.

This gives a continuous feature population while retaining geometry inherited from the transformed WDBC measurements.

For numerical evaluation, we approximate (13) using a large pre-generated pool and draw the secret-conditioned feature vectors from the corresponding in-ball and out-of-ball subsets. Thus, the WDBC likelihood calculation includes an additional finite-pool approximation that is absent from the Gaussian experiments.

Responses are generated as $Y = X^\top w^\star + \epsilon$, with $w^\star = e_1$ and $\epsilon \sim \mathcal{N}(0, 0.25^2)$.

For $p = 10$, we use the ten WDBC mean measurements and $n \in \{30, 50, 100\}$. For $p = 20$, we use the ten mean and ten worst measurements and $n \in \{50, 100, 200\}$. We take $z = 2e_1$, choose r so that $\pi_r = 1/2$, and train 30 independent ridge models in each regime.

For consistency with the synthetic experiments, we report

$$\widetilde{R}(\widehat{w}) = \|\widehat{w} - w^\star\|_2^2 + \sigma_\epsilon^2.$$

Since whitening gives identity empirical covariance and the smoothing adds $h^2 I_p$, the exact prediction risk under (13) differs only by the factor $1 + h^2 = 1.0025$ on the coefficient-error term.

Table 5: Paired synthetic privacy–utility comparison at $\pi_r = 1/2$. All estimators within a replicate use the same underlying training dataset. The DP methods are evaluated at the model-specific $\varepsilon_{\text{match}}$ obtained from the ordinary-ridge release. σ_{DP} is the median Gaussian output-perturbation coefficient-noise standard deviation.

p	n	Median $\varepsilon_{\text{match}}$	Median σ_{DP}	Ridge MSE	Clipped MSE	Gaussian-DP MSE	AdaSSP MSE
10	30	0.0821	2329	0.1045	0.1047	5.1743	1.1050
10	50	0.0806	1423	0.0877	0.0884	5.1820	1.1100
10	100	0.0489	1174	0.0804	0.0814	5.0527	1.1156
20	50	0.0877	2022	0.1156	0.1169	5.0444	1.0916
20	100	0.0561	1579	0.0896	0.0907	5.0578	1.0908
20	200	0.0377	1177	0.0787	0.0795	5.1173	1.0861

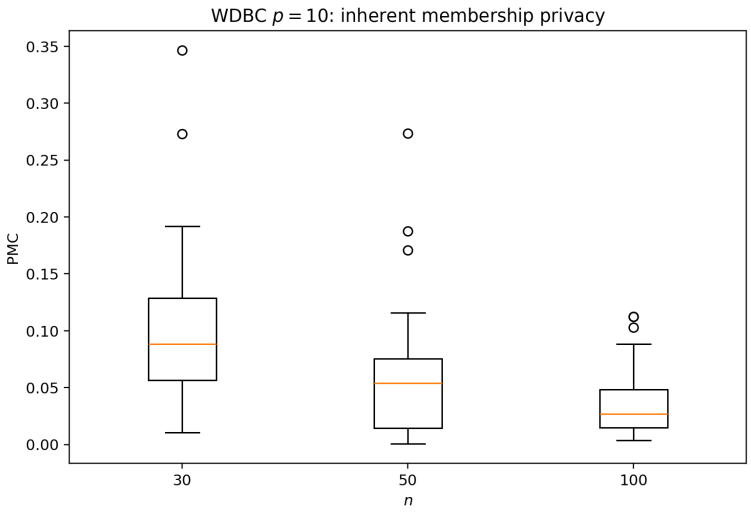


Figure 13: Proximity-membership PMC under the WDBC-derived feature population for $p = 10$ at $\pi_r = 1/2$, across 30 independently trained ordinary-ridge models.

D.5.1 Proximity-Membership Leakage

Figures 13 and 14 show the PMC distributions under balanced prior $\pi_r = 1/2$. For $p = 10$, median PMC decreases from 0.0884 at $n = 30$ to 0.0540 at $n = 50$ and 0.0269 at $n = 100$. For $p = 20$, the corresponding medians are 0.1016, 0.0322, and 0.0305 for $n = 50, 100, 200$, with 90th percentiles 0.3592, 0.1036, and 0.0708.

The larger-n regimes therefore reproduce the broad synthetic behavior: proximity-membership leakage remains nonzero but generally small and tends to decrease as n grows relative to dimension. The $p = 20, n = 50$ regime is numerically less stable and is discussed further in Section D.7.

D.5.2 Paired DP Utility Comparison

For each WDBC-based training dataset, we compute $\varepsilon_{\text{match}}$ from the ordinary-ridge release and evaluate Gaussian output perturbation and AdaSSP at that same numerical privacy level. For $p = 10$, the median values of $\varepsilon_{\text{match}}$ are 0.1696, 0.1052, and 0.0531 for $n = 30, 50, 100$. For $p = 20$, they are 0.1936, 0.0634, and 0.0600 for $n = 50, 100, 200$.

Figures 15 and 16 show the resulting prediction errors. The boxplots show the prediction error of the corresponding DP estimator across independently generated training datasets. We additionally plot two nonprivate reference curves. The blue dashed curve shows the mean prediction error of ordinary ridge regression. The orange dashed curve shows the mean prediction error of the clipped and constrained ridge estimator used as the deterministic input to the DP mechanism, before any privacy-specific noise is added. The difference between these two curves therefore isolates the

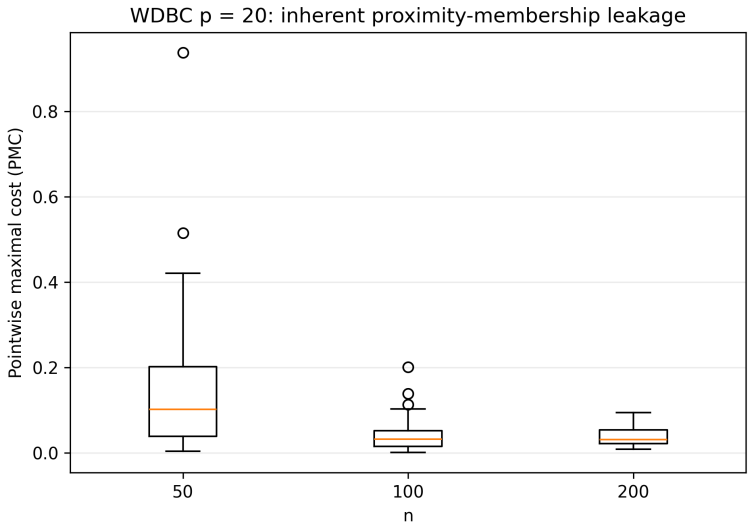


Figure 14: Proximity-membership PMC under the WDBC-derived feature population for $p = 20$. The $n = 50$ regime has substantially lower numerical reliability than the larger-n regimes.

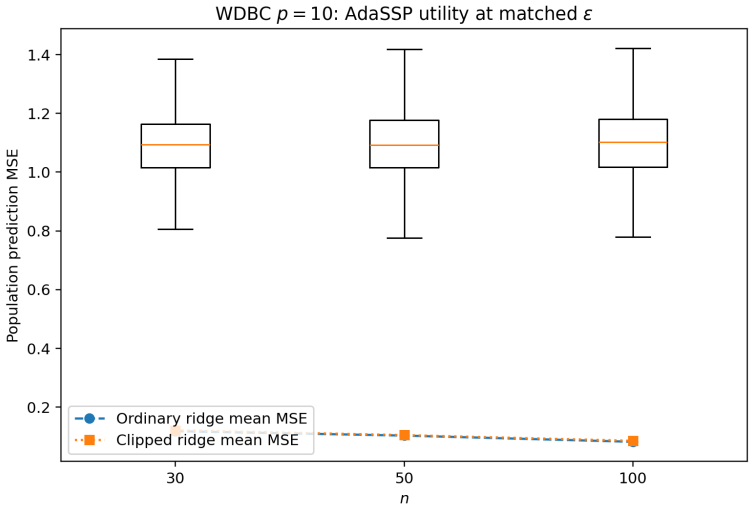


Figure 15: Prediction-error proxy for AdaSSP under the WDBC-derived feature population with $p = 10$, evaluated at model-specific matched numerical privacy levels.

utility cost of clipping and constraining the estimator, while the gap between the orange curve and the DP boxplots reflects the additional loss associated with the privacy mechanism.

Table 6 reports additional results using WDBC at balanced prior $\pi_r = 1/2$. Specifically, ordinary ridge has prediction error 0.1181, 0.1030, and 0.0819 for $p = 10$, while the corresponding AdaSSP values are 1.0860, 1.0873, and 1.1017, and the Gaussian-DP values are 4.9198, 5.0986, and 5.0175. For $p = 20$, ordinary-ridge errors are 0.1351, 0.0978, and 0.0807; AdaSSP gives 1.0762, 1.0708, and 1.0691; and Gaussian output perturbation gives 5.0574, 5.0350, and 5.1620.

Thus, the privacy–utility separation observed under Gaussian features persists under this real-data-derived feature population.

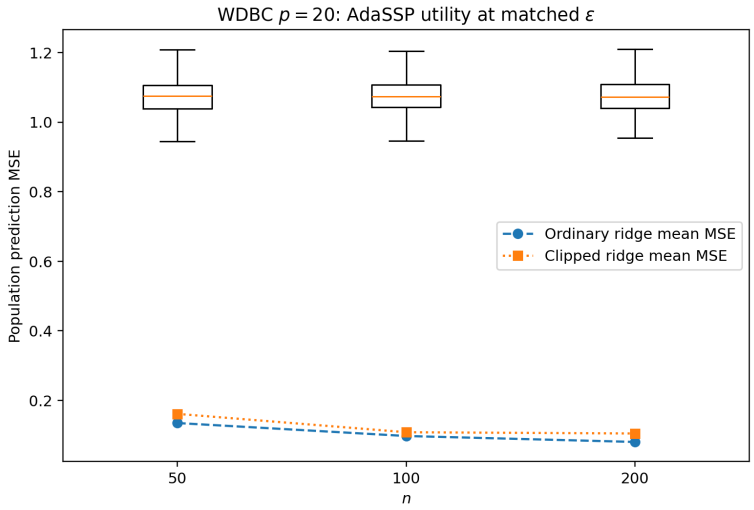


Figure 16: Prediction-error proxy for AdaSSP under the WDBC-derived feature population with $p = 20$.

Table 6: WDBC semi-synthetic results at $\pi_r = 1/2$. Columns 6-9 report MSE. Results are computed across 30 independently trained models. DP methods use the same underlying training datasets and the corresponding model-specific values of $\varepsilon_{\text{match}}$.

p	n	Median PMC	q_{90} PMC	Median $\varepsilon_{\text{match}}$	Ridge	Clipped Ridge	AdaSSP	Gaussian DP	Median minESS
10	30	0.0884	0.1626	0.1696	0.1181	0.1210	1.0860	4.9198	482.0
10	50	0.0540	0.1212	0.1052	0.1030	0.1047	1.0873	5.0986	546.9
10	100	0.0269	0.0899	0.0531	0.0819	0.0853	1.1017	5.0175	587.0
20	50	0.1016	0.3592	0.1936	0.1351	0.1614	1.0762	5.0574	73.8
20	100	0.0322	0.1036	0.0634	0.0978	0.1086	1.0708	5.0350	276.3
20	200	0.0305	0.0708	0.0600	0.0807	0.1050	1.0691	5.1620	332.4

D.6 Dependence on the Protected Target

To examine target dependence separately from prior membership probability, we vary $z = \rho e_1$ for $\rho \in \{0, 1, 2, 3\}$ and recalibrate the radius r at each target so that $\pi_r = 1/2$ remains fixed. Thus, differences in PMC across ρ cannot be attributed simply to changes in the prior probability of the protected event. Table 7 shows that target location can materially affect disclosure even when the prior probability of membership is unchanged. For example, at $p = 10, n = 50$, median PMC is 0.0373 at $\rho = 0$, decreases to 0.0201 at $\rho = 1$, and then increases to 0.0889 at $\rho = 3$; the corresponding 90th-percentile PMC at $\rho = 3$ is 0.1644. Hence, greater distance from the population mean does not systematically imply greater or smaller privacy. This nonmonotonicity is expected from the release-specific nature of the problem. Changing ρ changes not only the location of the protected region but also its relationship to the predictive direction $w^\star = e_1$. Moreover, because r is recalibrated at every ρ to keep $\pi_r = 1/2$, moving the target changes the geometry of the entire protected neighborhood rather than merely translating a fixed-radius ball. These changes alter the feature-matrix distributions under $S_r = 0$ and $S_r = 1$, and therefore can change how distinguishable the corresponding ridge-output distributions are. The results therefore indicate that prior membership probability and distance from the population mean are not sufficient to determine privacy. Disclosure depends on how the protected region interacts with the feature population, the regression problem, and the particular released coefficients.

Table 7: Dependence on target location at $\pi_r = 1/2$. The target is $z = \rho e_1$, and r is recalibrated for each ρ to preserve the same membership prior.

p	n	ρ	Median PMC	q_{90} PMC	Median minESS
10	50	0	0.0373	0.0635	2584.8
10	50	1	0.0201	0.0372	2435.9
10	50	2	0.0318	0.0845	2457.1
10	50	3	0.0889	0.1644	2458.3
10	100	0	0.0196	0.0446	1851.2
10	100	1	0.0147	0.0423	1924.8
10	100	2	0.0308	0.0834	1800.7
10	100	3	0.0563	0.1496	1921.5
20	100	0	0.0297	0.0846	393.8
20	100	1	0.0251	0.0897	435.8
20	100	2	0.0279	0.0854	440.8
20	100	3	0.0384	0.0899	444.6
20	200	0	0.0186	0.0757	573.3
20	200	1	0.0265	0.0721	459.4
20	200	2	0.0275	0.0707	436.0
20	200	3	0.0430	0.1506	513.8

Table 8: Selected concentration measures for PMC estimation. The final column gives the number of released models with minESS below 200. This threshold is used only as a descriptive numerical flag, not as a formal reliability criterion.

Setting	p	n	Median minESS	q_{10} minESS	# below 200
Synthetic	10	30	934.3	125.7	11
Synthetic	10	50	1400.3	413.2	3
Synthetic	10	100	2145.1	571.9	5
Synthetic	20	50	196.8	11.1	31
Synthetic	20	100	628.2	49.5	19
Synthetic	20	200	847.5	94.5	10
WDBC	10	30	482.0	289.3	0
WDBC	10	50	546.9	273.4	1
WDBC	10	100	587.0	292.3	2
WDBC	20	50	73.8	17.3	22
WDBC	20	100	276.3	151.3	6
WDBC	20	200	332.4	213.5	0

D.7 Numerical Reliability and Dimensional Scaling

The exact characterization reduces privacy evaluation to mixture expectations over secret-conditioned feature matrices. The main numerical difficulty is that the likelihood contributions can become highly concentrated as the dimension increases. We therefore distinguish the numerical stability of the estimator from the underlying privacy quantity.

Table 8 summarizes concentration measures for the main synthetic and WDBC regimes. For synthetic $p = 10$, median minESS ranges from approximately 934 to 2145. For $p = 20$, the larger-n regimes are better behaved, while $n = 50$ has median minESS 196.8 and 10th-percentile minESS 11.1.

The same effect is more pronounced under the WDBC-derived population. For $p = 20$, median minESS is 73.8, 276.3, and 332.4 for $n = 50, 100, 200$. Increasing the Monte Carlo budget improves many low-contribution-count releases, but $p = 20, n = 50$ remains difficult. We therefore interpret its upper-tail PMC estimates more cautiously than those of the larger-n regimes.

We also examined $p = 50$. Direct secret-conditioned Monte Carlo (Algorithm 1) produced very small effective contribution counts for many releases. We additionally explored alternative numeri-

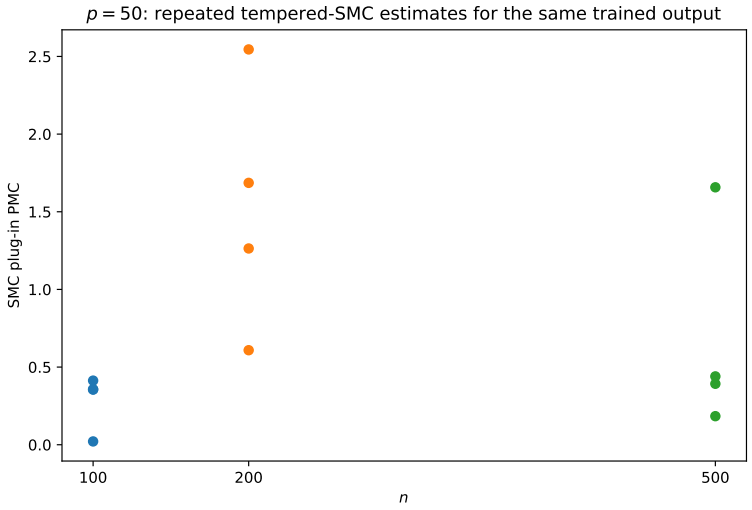


Figure 17: Repeated tempered-SMC likelihood-ratio estimates in a synthetic $p = 50$ regime. Although intermediate tempering steps can maintain favorable local particle-concentration and acceptance measures, independent repetitions produce materially different accumulated likelihood-ratio estimates. We therefore do not report the corresponding PMC values as substantive privacy measurements.

cal procedures, including importance sampling with adapted Gaussian densities, defensive mixtures, and tempered sequential Monte Carlo, whose result is shown in Figure 17, but these did not provide sufficiently stable repeated likelihood-ratio estimates either. Therefore, we report our results up to $p = 20$, and leave higher dimensional estimations for future development of richer computational methodologies.