PERFORMATIVE PRIVACY: WHEN DIFFERENTIAL PRIVACY MAXIMIZES UTILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Privacy-preserving learning is often motivated by the idea that protecting users' data can preserve trust and thus participation, improving utility in the long term. However, this claim has not been formalized so far. In parallel, performative learning provides a framework for studying learning systems whose deployment affects the data they later observe. In this work, we bring these two perspectives together and introduce *performative privacy*, where data leakage reduces future participation. We study a simple model where agents repeatedly contribute data for mean estimation but may leave the system when their data is leaked. Privacy is implemented through differentially private mechanisms, creating a trade-off between estimation noise and future participation. We show, through a theoretical study of the dynamics and numerical experiments, that a finite privacy budget can outperform non-private estimation in the long term when the feedback loop between leakage and participation is sufficiently strong. We then demonstrate a similar behavior for learning tasks with a simple linear regression setting. This provides first evidence that differential privacy can be optimal not only as a protection mechanism, but also from the perspective of long-term utility.

1 INTRODUCTION

Protecting privacy in machine learning is motivated by moral, ethical, and practical reasons (Solove, 2008). Access to sensitive data, such as health records, is sometimes illegal without adequate protections (European Parliament & Council of the European Union). However, in many scenarios, privacy protection is motivated by users' willingness to participate in training, based on the risks it entails for their business or their own privacy if their data is leaked (Nissenbaum, 2004; IBM & Ponemon Institute, 2025). Users who generate training data can be seen as agents who decide whether to participate depending on the risk that their data is leaked. A natural feedback loop to consider is thus that, if a user has their data leaked, they may no longer be willing to participate in the future, harming future rounds of training. This is often claimed informally: privacy-preserving machine learning should optimize long-term utility by maintaining users' trust, and thus their data (Dwork & Roth, 2013). Surprisingly, this claim that private learning can provide optimal long-term utility has, to the best of our knowledge, never been formalized. We propose first results in this direction.

The gold standard for protecting data is differential privacy (DP)(Dwork et al., 2006). DP formalizes the following promise: a single data point should not influence the output of the algorithm too much, ensuring that any attacker with access to the output cannot gain significant information about the data point by observing the output. This protection, however, can only be achieved through randomization, often implemented by injecting noise into the algorithm. The level of privacy protection is then summarized by the privacy budget ϵ: the smaller the privacy budget, the stronger the privacy guarantee, and the larger the noise required, which decreases the accuracy of the model, creating a privacy-utility trade-off. The right value for ϵ is often debated in the community, based on ad hoc perceptions of data sensitivity and risk (Ponomareva et al., 2023; Cyffers, 2025; Lee & Clifton, 2011). Here, we take an orthogonal perspective, considering the optimal privacy budget to be the one that maximizes long-term utility.

Taking into account how models change the data over time is the motivation of *performative learning* (Perdomo et al., 2020), where the data distribution is parametrized by the same parameter as

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.

the model, and the goal is to optimize the overall system rather than find the best model for a fixed distribution, as usually done in supervised learning. Performative learning is an increasingly popular framework, motivated by examples where users may game the system by modifying their features to obtain more favorable outcomes, by settings involving generative models where data can be contaminated over time, or by scenarios where sampling can reinforce existing biases, raising long-term fairness concerns (Hardt & Mendler-Dünner, 2023). Feedback loops due to privacy leakage have not been studied so far, but they constitute a natural dynamic to study, as it is well known that data leakage can cause trust issues and reduce participation, or even lead participants to adopt malicious behaviors (IBM & Ponemon Institute, 2025). Relatedly, the term *performative privacy* was coined by legal scholars (Skinner-Thompson, 2017) to describe the actions agents may take to exercise their right to privacy, by making unavailable information that could otherwise be available by default.

In this work, we design and study a simple model that allows us to quantify regimes in which a finite privacy budget maximizes utility in a repeated mean estimation problem. Rather than using the privacy budget directly as a proxy for the decision to participate, we use whether the data point has been leaked in the feedback loop. This makes our setting applicable to non-differentially private mechanisms as well and avoids focusing on a meta-interpretation of the privacy budget, instead focusing on actual algorithm outputs. More precisely, our contributions are the following:

- We introduce *performative privacy*, the long-term utility of repeated learning tasks in which privacy drives participation.
- We study it in three settings. We warm up with binary private values, where agents withdraw if their value is used and new agents join at each step (Section 3). We then move to d-dimensional mean estimation, with leakage detected via membership inference attacks (Section 4), and finish with linear regression, where a noisy estimator is released at each step and leakage is detected in the same way (Section 5).
- We show that in all settings a *finite privacy budget can be optimal*, proving that privacy can maximize long-term utility under participation feedback loops.

2 Related work

Privacy and participation While a direct link between data leakage and future participation has not yet been formalized, differential privacy originates from Warner (1965), where randomization aims to reduce bias in data collection, and thus improve data quality. Inspirations from sociology, notably through the contextual integrity framework, have also been numerous (Nissenbaum, 2004; Cyffers, 2025; Cummings & Durfee, 2020). Considering data producers as privacy-aware agents is a popular framework, sometimes formalized as a data market (Ghosh & Roth, 2011; Pai & Roth, 2013; Fallah et al., 2022), in particular for deciding participation in surveys (Roth & Schoenebeck, 2012; Ligett & Roth, 2012). The choice of the privacy budget itself has been analyzed through economic and mechanism-design lenses (Hsu et al., 2014; Nissim et al., 2012; Cummings et al., 2015; Ligett et al., 2017), possibly personalizing it across users (Jorgensen et al., 2015). However, privacy awareness is often translated into a supposedly known optimal privacy budget, computed from expected utility and privacy preferences (Cummings et al., 2022; Chaudhuri et al., 2025), rather than modeled as a dynamical system. Participation has also been studied as a strategic decision in collaborative and federated learning (Donahue & Kleinberg, 2021; Blum et al., 2021; Karimireddy et al., 2022), without the link to privacy leakage that we consider. On the empirical side, data breaches measurably affect customer behavior and firm value (IBM & Ponemon Institute, 2025), and the way privacy guarantees are communicated changes users' willingness to share their data (Xiong et al., 2020; Cummings et al., 2021), supporting the feedback loop we model.

Performativity in Trustworthy Machine Learning. Performative learning (Perdomo et al., 2020) is motivated by the social impact of machine learning deployments (Hardt & Mendler-Dünner, 2023; Hardt et al., 2022), with dedicated optimization algorithms (Mendler-Dünner et al., 2020) and regret analyses (Jagadeesan et al., 2022; Zhu et al., 2025), and with links to robustness (Cyffers et al., 2024) and fairness (Somerstep et al., 2024). Closest to our setting are retention dynamics, where user groups shrink after receiving poor utility (Hashimoto et al., 2018; Zhang et al., 2019): we study the privacy analogue, where users leave after data leakage. As the population size depends on the whole history of deployments, our model is an instance of stateful performativity (Brown et al., 2022). Self-reinforcing feedback loops have also been documented in predictive policing (Ensign

et al., 2018), dataset bias amplification (Taori & Hashimoto, 2023), and generative model collapse (Shumailov et al., 2024). Interestingly, participants can also contribute to steering the model toward desirable properties, as studied in collective action (Hardt et al., 2023). To the best of our knowledge, there is no theoretical work, nor simulated feedback-loop model, tackling privacy concerns from this perspective so far.

Membership Inference Attacks. Membership inference attacks (MIAs) were introduced in the context of genomic studies by Homer et al. (2008), who demonstrated that aggregate statistics can reveal an individual's participation in a dataset. MIAs have played an important role in motivating DP and establishing lower bounds for private data analysis (Bun et al., 2014; Nasr et al., 2021). In machine learning, numerous MIAs have been proposed under different threat models and access assumptions (Shokri et al., 2017; Sablayrolles et al., 2019; Even et al., 2026; Carlini et al., 2021; Yeom et al., 2017; Ye et al., 2022). These attacks are also widely used to empirically audit machine-learning systems and evaluate privacy defenses (Jagielski et al., 2020; Aerni et al., 2024; Jayaraman & Evans, 2019; Thudi et al., 2022). MIA is naturally cast in the hypothesis-testing formulation of DP (Wasserman & Zhou, 2010; Kairouz et al., 2015), culminating in Gaussian DP (Dong et al., 2022), whose privacy-loss random variable is the leakage statistic we analyze in Section 4.3. Beyond measuring average privacy leakage, recent work has shown that privacy risk can vary substantially across individuals (Leemann et al., 2024; Azize & Basu, 2025), motivating per-instance guarantees and individual privacy accounting (Wang, 2019; Feldman & Zrnic, 2021; Yu et al., 2023; Koskela et al., 2023), and that worst-case privacy guarantees may be overly pessimistic in finite-sample regimes (Haghifam et al., 2025). Most closely related to our setting, recent work has studied membership inference under repeated interactions (Michel et al., 2026).

3 Estimation of Bernoulli parameters

In this section, we study the estimation of a Bernoulli parameter when the ground truth changes at each time step, from i.i.d. samples locally privatized using randomized response (RR). An answer is considered leaked whenever the true answer is preserved by RR. More precisely, over T time steps, at each step t the learner computes an estimate $\hat{\mu}_t$ of the true parameter μ_t of an underlying Bernoulli distribution, from privatized responses $(Y_{i,t})_{i=1}^{N_t}$ of the users' samples $(X_{i,t})_{i=1}^{N_t} \in \{0,1\}^{N_t}$, drawn i.i.d. from $\mathcal{B}(\mu_t)$. The learner aims to minimize the mean squared error at the final horizon, $\mathcal{E} = \mathbb{E}\big[(\mu_T - \hat{\mu}_T)^2\big]$, or alternatively, as developed in the next section, the error over the whole trajectory, $\mathcal{R}_T = \sum_{t=1}^{T} \mathbb{E}(\mu_t - \hat{\mu}_t)^2$.

Randomized response We consider local differential privacy, where anyone can access $(Y_{i,t})_{i=1}^{N_t}$, and privacy is ensured at the user level via randomized response (Dwork & Roth, 2013; Ioannidis et al., 2024; Wang et al., 2016). We assume that all users have the same privacy budget ϵ, constant over time, and define

$$Y_{i,t} \triangleq \begin{cases} X_{i,t}, & \text{with prob. } \frac{1}{2}+\beta \\ 1-X_{i,t}, & \text{with prob. } \frac{1}{2}-\beta \end{cases} \qquad \text{with} \quad \beta = \frac{1}{2}\tanh\left(\frac{\epsilon}{2}\right) \in \left[0, \tfrac{1}{2}\right), \tag{1}$$

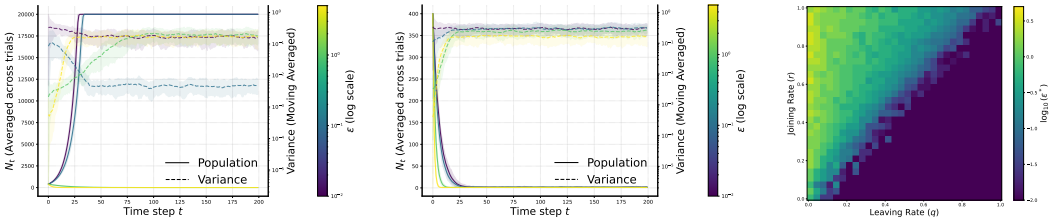
which guarantees ϵ-local differential privacy (ϵ-LDP). The learner's best unbiased estimator is then

$$\hat{\mu}_t \triangleq \frac{\hat{\nu}_t - \frac{1}{2} + \beta}{2\beta} \quad \text{with} \quad \hat{\nu}_t \triangleq \frac{1}{N_t}\sum_{i=1}^{N_t} Y_{i,t}. \tag{2}$$

Both claims are proved in Appendix E.

Population evolution The population changes as participating users leave in reaction to data leakage, defined as a response remaining unperturbed ($X_{i,t-1} = Y_{i,t-1}$), and as new users join. A leaked participant leaves with a fixed probability q, independent of time and of the user (the model has no memory of past leakages, which minimizes the number of parameters), and each participant whose data is not leaked recruits a new participant with probability r.[1] The expected one-step change of

[1] Growth proportional to the total number of users, regardless of leakage, is studied in Appendix C.



(a) $q = 0.3, r = 0.6$ (b) $q = 0.6, r = 0.3$ (c) Heatmap

Figure 1: Randomized response simulations. (a)–(b): population size N_t (clipped to $[1, 20000]$) and moving-averaged variance of $\hat{\mu}_t$ (window 15) for 10 values of ϵ (color intensity), in the low- ($q < r$) and high-attrition ($q > r$) regimes, averaged over 30 trials. (c): utility-maximizing ϵ as a function of (q, r), where utility is the squared error at the final horizon.

the population is then summarized by a single growth factor (Lemma 6 in Appendix E):

$$C(\beta) \triangleq \frac{\mathbb{E}[N_{t+1}|N_t]}{N_t} = 1 - q\left(\frac{1}{2} + \beta\right) + r\left(\frac{1}{2} - \beta\right) = 1 + \frac{r-q}{2} - \beta(q+r). \tag{3}$$

The expected population is stable or growing if $C(\beta) \geq 1$ and declines otherwise, which bounds how little private the mechanism can be:

Theorem 1 (Threshold for Nonnegative Expected Growth)**.** *For $r > q$, nonnegative expected growth, $C(\beta) \geq 1$, holds exactly when $\beta \leq \beta_c$, or equivalently $\epsilon \leq \epsilon_c \triangleq \ln(r/q)$, where:*

$$\beta_c \triangleq \frac{r-q}{2(q+r)} \tag{4}$$

We prove this theorem in Appendix E. From this formula, we can see two regimes: if $q > r$, the population always collapses, regardless of the privacy protection. However, when $r > q$, the population can grow or shrink depending on the choice of β, and $C(\beta) \geq 1$ gives a hard operational constraint on β. Simulations confirm this behavior (Figure 1): the population collapses for $q > r$, whereas for $r > q$ it reaches a stable equilibrium or grows to the maximum size, and a finite privacy budget optimally balances privacy noise against user retention. Moreover, where a finite optimum exists, the optimal ϵ decreases monotonically with the leaving rate q: stronger privacy is required when users are more sensitive to data leakage. Beyond expected growth, (N_t) is exactly a Galton–Watson branching process, whose extinction probability $e^{-N_0(\epsilon_c - \epsilon)_+}$ and fluctuations around $N_0 C(\beta)^t$ are given in Theorem 4 (Appendix E.3).

4 Mean Estimation with Membership Inference Detection

4.1 Setting

The previous section examined performative population dynamics under randomized response, where data is binary and leakage is defined as the exact release of the true bit. In that setting, utility and privacy risk were strictly opposed, as leakage was defined as sharing the single bit of information. Fortunately, richer data decouples a record's informational utility from its membership detection, as we show in this section by estimating the mean of continuous vectors in $\mathbb{R}^d$. A record located exactly at the population mean is statistically difficult to detect, but valuable for the estimation, whereas an outlier is easy to detect and hurts utility. Finally, this high-dimensional continuous setting is closer to realistic machine learning deployments, allowing us to formalize user vulnerability through standard Membership Inference Attacks (MIAs) based on an individual's geometric footprint.

More precisely, at each round t, the goal is to estimate the mean μ of a distribution $\mathcal{D}$ over $\mathbb{R}^d$ with covariance Σ from a set of N_t samples $Z_t = \{z_1, \ldots, z_{N_t}\}$ drawn independently from $\mathcal{D}$. To ensure bounded sensitivity, each record is clipped to a maximum ℓ_2-norm R: $\bar{z}_i = z_i \min\left(1, \frac{R}{\|z_i\|_2}\right)$

denotes the clipped vector z_i. In our theoretical utility results, we assume that R is large enough to neglect the clipping bias. The output o_t is an empirical mean protected by the Gaussian Mechanism:

$$o_t = \mathcal{M}_\gamma(Z_t) = \frac{1}{N_t}\sum_{i=1}^{N_t} \bar{z}_i + \eta_t, \quad \eta_t \sim \mathcal{N}\left(0, \frac{\gamma^2}{N_t} I_{d\times d}\right) \tag{5}$$

This corresponds to each user adding an independent noise $\mathcal{N}(0, \gamma^2 I_d)$ to their clipped record, whose ℓ_2-sensitivity is $2R$. To satisfy a given zero-concentrated differential privacy (zCDP) budget ρ (Bun & Steinke, 2016) for each individual contribution, the platform must thus set $\gamma^2 = \frac{2R^2}{\rho}$. Utility is measured through the expected squared Mahalanobis distance between the private release and the true mean μ (Biswas et al., 2020; Azize & Basu, 2025; Leemann et al., 2024), which, conditioned on N_t, is (Lemma 7 in Appendix F)

$$\mathbb{E}\left[\|o_t - \mu\|^2_{\Sigma^{-1}} \mid N_t\right] = \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_t}. \tag{6}$$

While Section 3 defined leakage via the discrete event $X_{i,t} = Y_{i,t}$, exposure in continuous aggregates ($o_t \in \mathbb{R}^d$) is quantified by a record's statistical influence on o_t. By the Neyman-Pearson lemma, the optimal test statistic for the hypothesis $z \in Z_t$ is the log-likelihood ratio:

$$\Lambda(o_t|z) \triangleq \ln\left(\frac{\Pr[\mathcal{M}_\gamma(Z_t) = o_t \mid z \in Z_t]}{\Pr[\mathcal{M}_\gamma(Z_t \setminus \{z\}) = o_t]}\right) = \ln\left(\frac{P^{\mathcal{M}}_{\text{in}}(o_t)}{P^{\mathcal{M}}_{\text{out}}(o_t)}\right) \tag{7}$$

where $\Pr[\cdot]$ denotes the density of the output, and $P^{\mathcal{M}}_{\text{in}}$ and $P^{\mathcal{M}}_{\text{out}}$ the factual and counterfactual output distributions. In the leave-one-out comparison, the round's noise covariance is held fixed; only the empirical mean is recomputed. The magnitude of $\Lambda(o_t|z)$ indicates how much skew z exerts on the noisy sample mean o_t given we observe the set Z_t. Practically, an individual cannot compute $\Lambda(o_t|z)$ without access to Z_t. However, it is the quantity approximated by external adversaries and auditing tools in MIAs (Shokri et al., 2017; Carlini et al., 2021). For simplicity, we assume this approximation to be exact, which provides a worst-case upper bound against any adversary. The privacy tolerance is captured via a constant threshold $\tau > 0$, such that a record is compromised if $\Lambda(o_t|z) > \tau$. From a hypothesis testing perspective, maintaining a fixed false positive rate would require scaling the threshold with the dimension d and the population size N_t over time, but we keep it constant for simplicity.

At round t, each participant $z \in Z_t$ updates their status for the subsequent round: if $\Lambda(o_t|z) > \tau$, the individual departs with probability q; otherwise ($\Lambda(o_t|z) \leq \tau$), the individual remains and recruits one additional participant with probability r, as described in Algorithm 1.

Algorithm 1 Perf. MIA Mean Estimation

Input: N_0, $\mathcal{D}$, bounds R, τ, q, r, T, γ
Set $N_1 = N_0$
for $t = 1, \ldots, T$ **do**
  if $N_t \leq 1$ **then**
    break
  Sample $Z_t \sim \mathcal{D}^{N_t}$, clip $\|z_i\|_2 \leq R$
  Compute exact mean $\mu_t = \frac{1}{N_t}\sum_{z_i \in Z_t} z_i$
  Share private output $o_t \sim \mathcal{N}(\mu_t, \frac{\gamma^2}{N_t} I_d)$
  Initialize $N_{t+1} \leftarrow N_t$
  for each $z_i \in Z_t$ **do**
    $\mu_t^{-i} \leftarrow \frac{N_t \mu_t - z_i}{N_t - 1}$
    $\Lambda_i \leftarrow \ln\left(\frac{\Pr[o_t|\mu_t]}{\Pr[o_t|\mu_t^{-i}]}\right)$
    if $\Lambda_i > \tau$ **then**
      $N_{t+1} \leftarrow N_{t+1} - 1$ w.p. q
    else
      $N_{t+1} \leftarrow N_{t+1} + 1$ w.p. r
Return $(N_1, \ldots, N_T)$

Algorithm 2 Perf. MIA Ridge Regression

Input: N_0, $\mathcal{D}$, bounds R, Y, reg λ, τ, q, r, T, γ
Set $N_1 = N_0$
for $t = 1, \ldots, T$ **do**
  if $N_t \leq d$ **then**
    break
  Sample $Z_t \sim \mathcal{D}^{N_t}$, clip $\|x_i\|_2 \leq R, |y_i| \leq Y$
  $A_t \leftarrow X_t^\top X_t + \lambda I_d$
  Compute exact ridge $\theta_t = A_t^{-1} X_t^\top y_t$
  Share private output $\tilde{\theta}_t \sim \mathcal{N}(\theta_t, \frac{\gamma^2}{N_t} I_d)$
  Initialize $N_{t+1} \leftarrow N_t$
  for each $(x_i, y_i) \in Z_t$ **do**
    $\theta_t^{-i} \leftarrow \theta_t - \frac{y_i - x_i^\top \theta_t}{1 - h_i} A_t^{-1} x_i$
    $\Lambda_i \leftarrow \ln\left(\frac{\Pr[\tilde{\theta}_t|\theta_t]}{\Pr[\tilde{\theta}_t|\theta_t^{-i}]}\right)$
    if $\Lambda_i > \tau$ **then**
      $N_{t+1} \leftarrow N_{t+1} - 1$ w.p. q
    else
      $N_{t+1} \leftarrow N_{t+1} + 1$ w.p. r
Return $(N_1, \ldots, N_T)$

4.2 General Bound on the Leakage Probability

Computing the exact trajectory of the population size requires the knowledge of the marginal leakage probability $p_\tau^{(t)} = \Pr(\Lambda(o_t|z) > \tau)$. For arbitrary data distributions, deriving the exact probability density function of the log-likelihood ratio is intractable. Instead, we provide an upper bound on the leakage probability that involves the Rényi divergence, and thus makes the link with Rényi DP (Mironov, 2017), via Markov's inequality.

Lemma 1 (Tail Bound on the Leakage Probability)**.** *With the above notations, for any threshold $\tau > 0$ and order $\alpha > 1$, the tail probability of leakage is bounded by:*

$$\Pr_{o_t \sim P_{in}^{\mathcal{M}}}(\Lambda(o_t|z) > \tau) \leq \exp\Big(-(\alpha-1)\big(\tau - D_\alpha(P_{in}^{\mathcal{M}} \parallel P_{out}^{\mathcal{M}})\big)\Big)$$

where $D_\alpha(P_{in}^{\mathcal{M}} \parallel P_{out}^{\mathcal{M}})$ is the Rényi divergence of order α.

In Appendix F.3, we show that, when applied to the Gaussian mechanism, this general bound is tight up to the Φ-versus-$\exp$ gap.

4.3 The Gaussian Mechanism Case

We now exploit the fact that the noise $\eta_t \sim \mathcal{N}(0, \frac{\gamma^2}{N_t} I_{d\times d})$ is Gaussian, ignoring clipping for simplicity. Let $\mu_{z,t} \triangleq \frac{1}{N_t}\sum_{i=1}^{N_t} z_i$ denote the sample mean at round t. By direct computation (Lemma 9 in Appendix F), the log-likelihood ratio $\Lambda(o_t|z)$ conditioned on Z_t is distributed as

$$\Lambda(o_t|z) \mid Z_t \sim \mathcal{N}\left(\frac{N_t\|z-\mu_{z,t}\|^2}{2(N_t-1)^2\gamma^2},\ \frac{N_t\|z-\mu_{z,t}\|^2}{(N_t-1)^2\gamma^2}\right). \tag{8}$$

From this, we deduce the exact leakage probability:

Lemma 2. *Given the aggregated empirical mean mechanism, the exact probability that a specific individual data point z is exposed (i.e., its log-likelihood ratio breaches the threshold τ) conditioned on Z_t is given in closed form by:*

$$\Pr(\Lambda(o_t|z) > \tau \mid Z_t) = 1 - \Phi\left(\frac{\tau(N_t-1)\gamma}{\sqrt{N_t}\|z-\mu_{z,t}\|} - \frac{\sqrt{N_t}\|z-\mu_{z,t}\|}{2(N_t-1)\gamma}\right)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

This probability increases with the distance of z to the empirical mean, so outliers are the most exposed, and decreases with N_t, reflecting privacy amplification by aggregation: a growing population becomes less exposed, creating a virtuous circle.

Remark 1 (Connection to Gaussian DP)**.** *With $\mu_{z',t} \triangleq \sqrt{N_t}\|z-\mu_{z,t}\|/(\gamma(N_t-1)) \approx \|z-\mu_{z,t}\|/(\gamma\sqrt{N_t})$, equation 8 reads $\Lambda(o_t|z) \mid Z_t \sim \mathcal{N}(\mu_{z',t}^2/2, \mu_{z',t}^2)$, which is the privacy loss of a $\mu_{z',t}$-Gaussian-DP pair (Dong et al., 2022): for the record z, membership inference is exactly as hard as testing $\mathcal{N}(0,1)$ against $\mathcal{N}(\mu_{z',t},1)$. Lemma 2 is then the power of the Neyman–Pearson test at likelihood-ratio threshold e^τ — equivalently, of the Bayes-optimal attack with prior odds of membership $e^{-\tau}$. Thus, maintaining a fixed FPR ν corresponds to the record-dependent threshold $\tau = \mu_{z',t}\Phi^{-1}(1-\nu) - \mu_{z',t}^2/2$, as discussed in Section 4.1, and since $\|\bar{z}\|_2, \|\mu_{\bar{z},t}\|_2 \leq R$, one recovers the worst-case $(2R\sqrt{N_t}/(\gamma(N_t-1)))$-Gaussian-DP guarantee.*

4.4 The Gaussian Data Case

We now further assume that the data is drawn from a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ which allows us to characterize precisely the dynamics.

Lemma 3 (Expected Population Evolution)**.** *We define the marginal leakage probability, $p_\tau^{(t)}$, as the expected fraction of the population that breaches the privacy threshold τ at round t, i.e., $p_\tau^{(t)} = \mathbb{E}[\mathbb{1}(\Lambda(o_t|z) > \tau)]$. Letting $\bar{z}$ denote the record subjected to the clipping and $\mu_{\bar{z},t}$ denote the sample mean of clipped vectors, the marginal leakage probability evaluates to:*

$$p_\tau^{(t)} = \int_{\mathbb{R}^{dN_t}} \left[1 - \Phi\left(\frac{\tau(N_t-1)\gamma}{s\sqrt{N_t}} - \frac{s\sqrt{N_t}}{2(N_t-1)\gamma}\right)\right] \prod_{j=1}^{N_t} f_{\mathcal{N}}(z_j)\, dz_{1:N_t}, \tag{9}$$

where $s \triangleq \|\bar{z}_1 - \mu_{\bar{z},t}\|_2$ *and* $dz_{1:N_t} \triangleq dz_1 \cdot \ldots dz_{N_t}$. *The expected population size at round* $t+1$ *in Algorithm 1 is then given by:*

$$\mathbb{E}[N_{t+1} \mid N_t] = N_t\Big(1 + r - (q+r)p_\tau^{(t)}\Big) \tag{10}$$

Lemma 3 gives an exact characterization of the user population evolution. As in Lemma 6, the expected population is the previous population multiplied by a growth factor. However, unlike the previous case, this factor is determined by the marginal leakage probability p_τ, whose integral form captures the dependence on several parameters: the dimension d, the threshold τ, the clipping bound R and the noise scale γ, instead of relying only on the privacy parameter β.

To determine whether the expected user population grows or declines, we denote similarly as before $C_t(\gamma)$ the expected ratio of the population size from round t to round $t+1$:

$$C_t(\gamma) \triangleq \frac{\mathbb{E}[N_{t+1} \mid N_t]}{N_t} = 1 + r - (q+r)p_\tau^{(t)}(\gamma) \tag{11}$$

The continuous growth factor $C_t(\gamma) = 1 + r - (q+r)p_\tau^{(t)}(\gamma)$ mirrors the discrete model of Lemma 6 by substituting the leakage probability $\frac{1}{2} + \beta$ of randomized response with the marginal leakage probability $p_\tau^{(t)}(\gamma)$. This substitution preserves the system's monotonicity: increasing the noise γ strictly decreases $p_\tau^{(t)}(\gamma)$, improving population stability. However, because $p_\tau^{(t)}(\gamma)$ requires evaluating a non-linear integral over a multivariate data distribution, the boundary condition $C_t(\gamma) \geq 1$ cannot be inverted analytically in general, but it can easily be computed numerically. Moreover, unlike in Section 3, C_t is now time-dependent, introducing an additional layer of performativity.

Theorem 2 (Sufficient Noise Threshold for Expected Growth)**.** *The condition* $C_t(\gamma) \geq 1$ *is guaranteed as soon as the mechanism's noise scale satisfies the explicit lower bound* $\gamma \geq \gamma_c$. *For* $q, r > 0$, $N_t > 1$ *and clipping bound* R, *a sufficient threshold* γ_c *is defined as:*

$$\gamma_c \triangleq \frac{2R\sqrt{N_t}}{N_t - 1}\left(\frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau}\right) \tag{12}$$

where $\Phi_{q,r}^{-1} \triangleq \Phi^{-1}\left(\frac{q}{q+r}\right)$ *is the probit function (the inverse of the standard normal cumulative distribution function) evaluated at the performative survival threshold.*

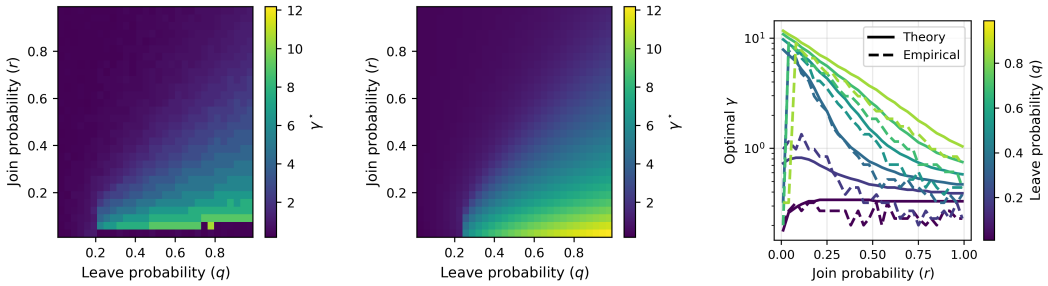
Here $C_t(\gamma) < 1$ indicates expected population decline, which can reinforce leakage and further decline. Conversely, as the marginal leakage probability decreases with N_t (proof of Lemma 11), small populations leak more and shrink while large ones leak less and grow. This phenomenon is known as the Allee effect in population ecology (Allee & Bowen, 1932; Courchamp et al., 1999). For $R = \Theta(\sqrt{d})$, the sufficient threshold scales as $\sqrt{d/N_t}$, separating the data geometry from the behavioral parameters.

In Appendix, theorem 5 gives expected-error bounds of the same order, proportional to $(\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d)/(C_{t-1}(\gamma)N_{t-1})$ for large N_{t-1} and $q < 1$. The unconditional error $\mathcal{E}_t(\gamma) \triangleq \mathbb{E}[\|o_t - \mu\|_{\Sigma^{-1}}^2]$ depends on $\mathbb{E}[1/N_t]$, which we approximate by the inverse mean-field population. Empirically accurate (Figure 2), this approximation is further justifiable globally via sequential chaining on Lemma 13 (Appendix F). A fixed noise level sufficient for growth throughout the trajectory can be calibrated at its minimum population size, since $\gamma_c \propto 1/\sqrt{N_{\min}}$.

Remark 2 (Finite-Horizon Optimization)**.** *In the mean-field dynamics, low noise can initiate a feedback loop of leakage, population decline, and increasing estimation error. Conversely, sufficiently strong privacy can sustain population growth, eventually offsetting its noise penalty (Corollaries 1 and 2). At a finite horizon, however, excessive noise remains costly, motivating the trade-off studied below.*

Optimization of the Global Noise Scale (γ^*). In the growth regime $C_1(\gamma) > 1$, the cumulative error $\mathcal{E}_{\mathrm{total}}(\gamma) \triangleq \sum_{t=1}^{T} \mathcal{E}_t(\gamma)$ is upper bounded, under the mean-field approximation, by (Lemma 12 in Appendix F.5)

$$\hat{\mathcal{E}}_{\mathrm{total}}(\gamma) \triangleq \frac{C_1(\gamma)(\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d)}{N_0(C_1(\gamma) - 1)}\left(1 - \frac{1}{[C_1(\gamma)]^T}\right). \tag{13}$$



(a) Empirical heatmap (b) Theoretical heatmap (c) γ^* line plot comparison

Figure 2: Comparison of the optimal noise scale γ^* obtained from theory and simulations in mean estimation for $\tau = 0.02$. (c) compares the theoretical and empirical γ^* as a function of the join probability. Other values of τ are reported in Figure 4.

This surrogate captures the noise–growth trade-off and depends on the initial growth factor $C_1(\gamma)$, avoiding tracking the sequence $(C_t(\gamma))_t$. The closed-form objective equation 13 makes explicit the trade-off between providing better utility at a given round and a larger population at the next round. Increasing γ penalizes the numerator via the quadratic variance term (γ^2), but simultaneously reduces the error by increasing the growth factor $C(\gamma)$ in the denominator. We cannot compute a closed form of $\gamma^* = \arg\min_{\gamma>\gamma_c} \hat{\mathcal{E}}_{\text{total}}(\gamma)$. Indeed, the marginal leakage $p_\tau^{(1)}$ nested within $C_1(\gamma)$ is governed by the standard normal cumulative distribution function $\Phi(\cdot)$, so setting the derivative of this objective to zero does not yield a closed-form algebraic solution. However, $\hat{\mathcal{E}}_{\text{total}}(\gamma)$ is a smooth, one-dimensional objective function and γ^* can easily be computed with numerical solvers.

We thus compute the theoretical γ^* heatmap over the population's recruitment rate r and leaving rate q. As reported in Figure 2 (and Figure 4 in Appendix A), the theoretical γ^* closely aligns with the optimal empirical parameters discovered via grid search.

5 Linear Regression With Membership Inference Detection

Our framework extends beyond mean estimation to supervised learning, which we illustrate with differentially private ridge regression. At round t, each of the N_t records contains a feature vector $x_i \in \mathbb{R}^d$, $x_i \sim \mathcal{N}(0, \Sigma)$, and a label $y_i \mid x_i \sim \mathcal{N}(x_i^\top \theta_t^*, \sigma^2)$, where θ_t^* is the true parameter at round t. Features are clipped to $\|x_i\|_2 \le R$ and labels to $|y_i| \le Y$. The mechanism releases the output-perturbed ridge estimator

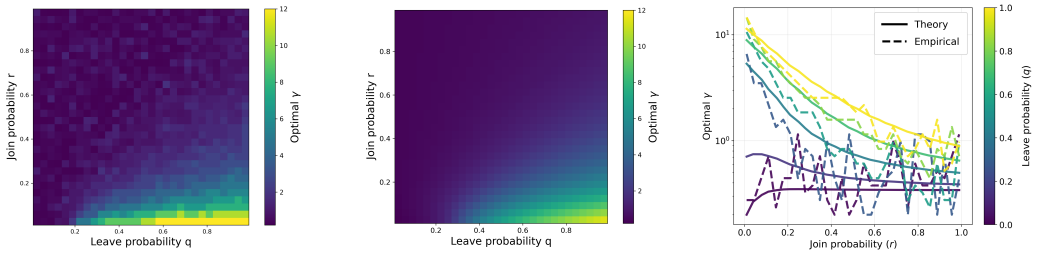
$$\widetilde{\theta}_t = \theta_t + \eta_t, \quad \theta_t = A_t^{-1} X_t^\top y_t, \quad A_t = X_t^\top X_t + \lambda I_d, \quad \eta_t \sim \mathcal{N}\Big(0, \frac{\gamma^2}{N_t} I_d\Big), \tag{14}$$

where X_t and y_t stack the features and labels, and λ is the regularization parameter. Leakage is detected as in Section 4, holding both the noise covariance and penalty λ fixed in the leave-one-out comparison. By the Sherman–Morrison formula, removing record i shifts the estimator by $\Delta\theta_i = \frac{e_i}{1-h_i} A_t^{-1} x_i$, with residual $e_i = y_i - x_i^\top \theta_t$ and leverage $h_i = x_i^\top A_t^{-1} x_i$. Hence $\Lambda(\widetilde{\theta}_t \mid x_i, y_i) \mid Z_t \sim \mathcal{N}(a_i, 2a_i)$ with $a_i = N_t \|\Delta\theta_i\|^2/(2\gamma^2)$: this is exactly equation 8, with $\Delta\theta_i$ in place of the mean shift $(z - \mu_{z,t})/(N_t - 1)$. The leakage probability, the growth factor $C_t(\gamma)$ and the population dynamics thus keep the same form (Appendix G), and bounding $\|\Delta\theta_i\|$ yields the critical noise threshold. Algorithm 2 presents the procedure underlying this setting.

Theorem 3. *For $q, r > 0$ and $\lambda > 0$, the condition $C_t(\gamma) \ge 1$ is guaranteed as soon as the noise scale of the ridge mechanism equation 14 satisfies $\gamma \ge \gamma_c$, where*

$$\gamma_c \triangleq \sqrt{N_t}\, \Delta_{\max} \left(\frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau} \right), \qquad \Delta_{\max} \triangleq \frac{RY}{\lambda} \left(1 + R\sqrt{\frac{N_t}{\lambda}} \right),$$

$\Phi_{q,r}^{-1} \triangleq \Phi^{-1}\left(\frac{q}{q+r}\right)$ as in Theorem 2, and $\Delta_{\max}$ bounds the shift $\|\Delta\theta_i\|_2$ induced by any single record.



(a) Empirical heatmap (b) Theoretical heatmap (c) γ^* line plot comparison

Figure 3: Comparison of the optimal noise scale γ^* obtained from theory and simulations in linear regression.

Remark 3 (Influence of a record compared to mean estimation)**.** *In mean estimation, the influence of a record on the release is the mean shift* $(z-\mu_{z,t})/(N_t-1)$*, so that outliers are the most exposed. In ridge regression, it becomes the leave-one-out shift* $\Delta\theta_i = \frac{e_i}{1-h_i} A_t^{-1} x_i$*, which combines the residual* e_i *and the leverage* h_i*: the most exposed records are those that are both atypical in feature space and poorly fitted by the model. Since the residual can grow with* $\|x_i\|$*, the worst-case bound* $\Delta_{\max}$ *scales with* R^2 *instead of* R*, but this worst case is conservative: for well-specified data, residuals are of order* σ *and* $A_t \approx N_t\Sigma + \lambda I$*, so typical shifts remain of order* $\sqrt{d}/N_t$*, as in mean estimation.*

Analogous to Section 4, under the mean-field approximation with nondecreasing growth factors, the cumulative expected excess risk is upper bounded by $\hat{\mathcal{E}}_{\text{total}}(\gamma)$ (Lemma 17). This bound has the same form as equation 13, with $\gamma^2\text{Tr}(\Sigma^{-1})+d$ replaced by $\gamma^2\text{Tr}(\Sigma)+\nu$, where $\nu = O(d)$ accounts for non-private estimation error, including regularization. We again select γ^* by minimizing this surrogate. Figure 3 compares the γ^* minimizing this objective in the growth regime ($\gamma > \gamma_c$) with the one found by simulation: their agreement supports applying the mean-estimation surrogate to the performative dynamics of parametric learning. As with mean estimation, providing privacy guarantees to users maximizes the long-term utility in repeated ridge regression, and the optimal level of protection increases with the probability q that a user leaves.

6 Discussion and Future Work

We showed that private estimation can outperform non-private estimation in the long term when participants drop out after their data is leaked. These results demonstrate that utility maximization can align with strong privacy protection when there is a sufficiently strong feedback loop. This paper introduces performative privacy as this study of the long-term interplay between privacy and utility under repeated model deployment, at the intersection of privacy-preserving machine learning and performative learning.

We focused on mean estimation and linear regression which enable precise theoretical characterization of the dynamics, but we plan to extend our results to more complex learning tasks. The feedback loop is also kept as simple as possible, but it could be made more realistic using epidemic models and more complex privacy preferences and composition over time, and by studying different definition of what is a privacy leakage, for instance by moving from membership inference to reconstruction. Interesting dynamics could also arise from having different platform competing for a finite based of user and using different privacy and utility guarantees. Additionally, the privacy budget can vary over time, creating more complex feedback loops as well.

AI use statement

In this work, we used generative AI tools to polish the writing, implement straightforward parts of the code (e.g., Matplotlib and seaborn code for generating figures), fill in few details of proofs for which we had already identified the key methods and arguments, and improve the completeness of the related work section. We did not use generative AI tools to develop the paper's core ideas, define the problem setting, or prove the main theoretical results. We take full responsibility for the

final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

Ethics statement

The authors confirm that this manuscript complies with all applicable ethical standards and does not violate any relevant code of ethics.

Reproducibility statement

All experiments were conducted on synthetically generated datasets that mimic the real world and have been sufficiently motivated in the paper as well as explained clearly in the code. We describe the simulated setup in sufficient detail to enable reproduction of the paper's figures. An anonymus Github link along with additional experimental details has been provided in Appendix D. All proofs are provided in Appendices E, F and G.

References

Michael Aerni, Jie Zhang, and Florian Tramèr. Evaluations of machine learning privacy defenses are misleading. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 1271–1284, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3690194.

W. C. Allee and Edith S. Bowen. Studies in animal aggregations: Mass protection against colloidal silver among goldfishes. *Journal of Experimental Zoology*, 61(2), February 1932. ISSN 1097-010X. doi: 10.1002/jez.1400610202.

Krishna B. Athreya and Peter E. Ney. *Branching Processes*. Springer Berlin Heidelberg, 1972. ISBN 9783642653711. doi: 10.1007/978-3-642-65371-1.

Achraf Azize and Debabrota Basu. Some targets are harder to identify than others: Quantifying the target-dependent membership leakage. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, 33:14475–14485, 2020.

Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, and Han Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1005–1014, 2021.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6045–6061, 2022.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pp. 635–658. Springer, 2016.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pp. 1–10, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327107. doi: 10.1145/2591796.2591877.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, A. Terzis, and Florian Tramèr. Membership inference attacks from first principles. *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2021.

Syomantak Chaudhuri, Konstantin Miagkov, and Thomas A. Courtade. Mean estimation under heterogeneous privacy demands. *IEEE Transactions on Information Theory*, 71(2):1362–1375, February 2025. ISSN 1557-9654. doi: 10.1109/tit.2024.3511498.

Franck Courchamp, Tim Clutton-Brock, and Bryan Grenfell. Inverse density dependence and the allee effect. *Trends in Ecology and Evolution*, 14(10), October 1999. ISSN 0169-5347. doi: 10.1016/s0169-5347(99)01683-3.

Rachel Cummings and David Durfee. Individual sensitivity preprocessing for data privacy. pp. 528–547, 2020.

Rachel Cummings, Katrina Ligett, Aaron Roth, Zhiwei Steven Wu, and Juba Ziani. Accuracy for sale: Aggregating data with a variance constraint. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 317–324, 2015.

Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. "I need a better description": An investigation into user expectations for differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 3037–3052, 2021.

Rachel Cummings, Hadi Elzayn, Vasilis Gkatzelis, Emmanouil Pountourakis, and Juba Ziani. Optimal data acquisition with privacy-aware agents. *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 210–224, 2022.

Edwige Cyffers. Setting epsilon is not the issue in differential privacy. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*, 2025.

Edwige Cyffers, Muni Sreenivas Pydi, Jamal Atif, and Olivier Cappé. Optimal classification under performative distribution shift. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2175.

Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5303–5311, 2021.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4), 2013.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, Berlin, Germany, 2006. Springer Berlin Heidelberg.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, volume 81 of *Proceedings of Machine Learning Research*, pp. 160–171, 2018.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Mathieu Even, Clément Berenfeld, Linus Bleistein, Tudor Cebere, Julie Josse, and Aurélien Bellet. Causal evaluation of membership inference attacks, 2026.

Alireza Fallah, Ali Makhdoumi, Azarakhsh Malekian, and Asuman E. Ozdaglar. Optimal and differentially private data acquisition: Central and local mechanisms. *Proceedings of the 23rd ACM Conference on Economics and Computation*, 2022.

Vitaly Feldman and Tijana Zrnic. Individual privacy accounting via a Rényi filter. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28080–28091, 2021.

Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pp. 199–208, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302616. doi: 10.1145/1993574.1993605.

Mahdi Haghifam, Adam Smith, and Jonathan Ullman. The sample complexity of membership inference and privacy auditing, 2025.

Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*, 2023.

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22969–22981, 2022.

Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12570–12586. PMLR, 23–29 Jul 2023.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938, 2018.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, August 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000167.

Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C. Pierce, and Aaron Roth. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium (CSF)*, pp. 398–410. IEEE, 2014.

IBM and Ponemon Institute. Cost of a Data Breach Report 2025. Technical report, IBM, 2025. Accessed 2026-04-30.

Andreas Ioannidis, Antonios Litke, and Nikolaos Papadakis. Randomized response and differential privacy. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 600–605, 2024.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *ICML*, 2022.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: how private is private sgd? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium*, pp. 1895–1912, 2019.

Zachary Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? personalized differential privacy. In *IEEE International Conference on Data Engineering (ICDE)*, pp. 1023–1034, 2015.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1376–1385, 2015.

Sai Praneeth Karimireddy, Wenshuo Guo, and Michael I. Jordan. Mechanisms that incentivize data sharing in federated learning. arXiv preprint arXiv:2207.04557, 2022.

H. Kesten and B. P. Stigum. A limit theorem for multidimensional galton-watson processes. *Ann. Math. Statist.*, 37(6):1211–1223, 1966.

Antti Koskela, Marlon Tobaben, and Antti Honkela. Individual privacy accounting with Gaussian differential privacy. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Jaewoo Lee and Chris Clifton. *How Much Is Enough? Choosing ϵ for Differential Privacy*, pp. 325–340. Springer Berlin Heidelberg, 2011. ISBN 9783642248610. doi: 10.1007/978-3-642-24861-0_22.

Tobias Leemann, Bardh Prenkaj, and Gjergji Kasneci. Is my data safe? predicting instance-level membership inference success for white-box and black-box attacks. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

Katrina Ligett and Aaron Roth. Take it or leave it: running a survey when privacy comes at a cost. In *Proceedings of the 8th International Conference on Internet and Network Economics*, WINE'12, pp. 378–391, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642353109. doi: 10.1007/978-3-642-35311-6_28.

Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Steven Z. Wu. Accuracy first: Selecting a differential privacy level for accuracy-constrained ERM. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4929–4939, 2020.

Thomas Michel, Debabrota Basu, and Emilie Kaufmann. Sequential membership inference attacks, 2026.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, August 2017. doi: 10.1109/csf.2017.11.

Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 866–882. IEEE, May 2021. doi: 10.1109/sp40001.2021.00069.

Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, 2004.

Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC)*, pp. 774–789, 2012.

Mallesh M. Pai and Aaron Roth. Privacy and mechanism design. *ACM SIGecom Exchanges*, 12(1): 8–29, 2013. ISSN 1551-9031. doi: 10.1145/2509013.2509016.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

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. ISSN 1076-9757. doi: 10.1613/jair.1.14649.

Aaron Roth and Grant Schoenebeck. Conducting truthful surveys, cheaply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pp. 826–843, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314152. doi: 10.1145/2229012.2229076.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. White-box vs black-box: Bayes optimal strategies for membership inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5558–5567. PMLR, 09–15 Jun 2019.

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)*, San Jose, CA, USA, May 2017.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Scott Skinner-Thompson. Performative privacy. *UC Davis Law Review*, 50(4), 2017. NYU School of Law, Public Law Research Paper No. 17-10; U of Colorado Law Legal Studies Research Paper No. 17-4.

Walter L. Smith and William E. Wilkinson. On branching processes in random environments. *Ann. Math. Statist.*, 40(6):814–827, 1969.

Daniel J. Solove. *Understanding Privacy*. Harvard University Press, Cambridge, MA, 2008. GWU Legal Studies Research Paper No. 420, GWU Law School Public Law Research Paper No. 420.

Seamus Somerstep, Ya'acov Ritov, and Yuekai Sun. Algorithmic fairness in performative policy learning: Escaping the impossibility of group fairness. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, FAccT '24, pp. 616–630. ACM, 2024. doi: 10.1145/3630106.3658929.

Rohan Taori and Tatsunori Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33883–33920, 2023.

Anvith Thudi, Ilia Shumailov, Franziska Boenisch, and Nicolas Papernot. Bounding membership inference. arXiv preprint arXiv:2202.12232, 2022.

Yu-Xiang Wang. Per-instance differential privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019.

Yue Wang, Xintao Wu, and Donghui Hu. Using randomized response for differential privacy preserving data collection. In *EDBT/ICDT Workshops*, volume 1558, pp. 0090–6778, 2016.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. doi: 10.1080/01621459.1965.10480775.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. Towards effective differential privacy communication for users' data sharing decision and comprehension. In *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 392–410, 2020. doi: 10.1109/SP40000.2020.00088.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 3093–3106, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2017.

Da Yu, Gautam Kamath, Janardhan Kulkarni, Tie-Yan Liu, Jian Yin, and Huishuai Zhang. Individual privacy accounting for differentially private stochastic gradient descent. *Transactions on Machine Learning Research*, 2023.

Xueru Zhang, Mohammad Mahdi Khalili, Cem Tekin, and Mingyan Liu. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Haiqing Zhu, Tijana Zrnic, and Celestine Mendler-Dünner. Look-ahead reasoning on learning platforms. In *Advances in Neural Information Processing Systems*, 2025.

Appendix

Table of Contents

A Numerical Experiments For Section 4 **16**

B Detailed Plots from Experiments in Section 3 **17**

C The Agnostic Model **17**

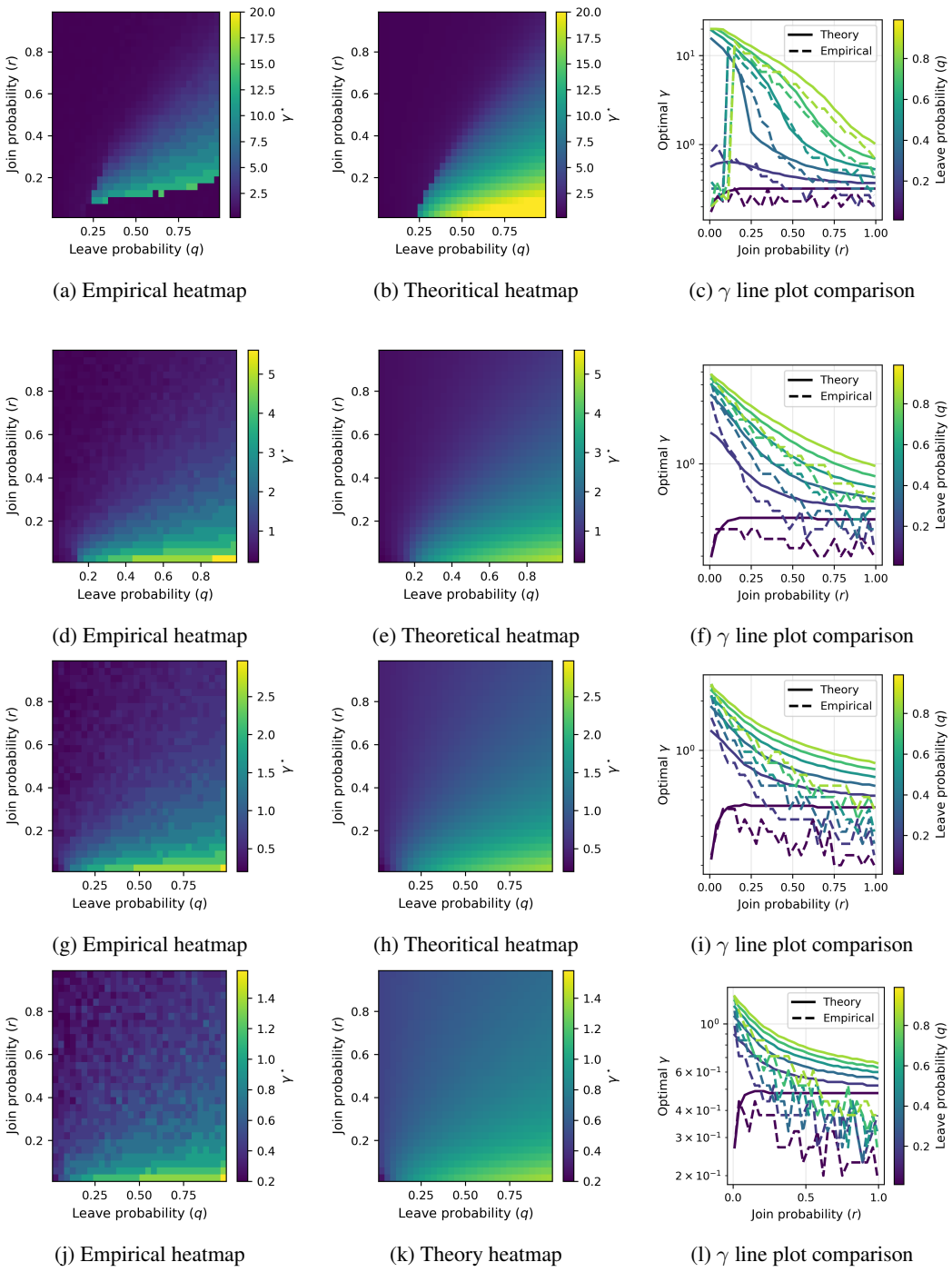
D Details About Numerical Experiments **18**

E Proofs from Section 3 **19**
E.1 Properties of the Estimator 19
E.2 Population Evolution and Critical Noise Threshold 20
E.3 Galton–Watson Analysis of the Population 21

F Proofs from Section 4 **22**
F.1 Generic Results 22
F.2 Results for Gaussian Mechanism 23
F.3 Comparison of the Generic Bound with the Exact Gaussian Leakage 25
F.4 Population Evolution and Critical Noise Threshold 26
F.5 Bounds and Asymptotics of the Error 28
F.6 High-Probability Population Lower Bound 30

G Theoretical Results from Section 5 **31**

A NUMERICAL EXPERIMENTS FOR SECTION 4



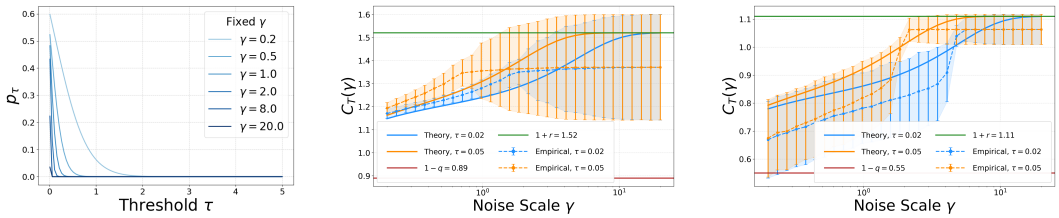
(a) Empirical heatmap (b) Theoritical heatmap (c) γ line plot comparison

(d) Empirical heatmap (e) Theoretical heatmap (f) γ line plot comparison

(g) Empirical heatmap (h) Theoritical heatmap (i) γ line plot comparison

(j) Empirical heatmap (k) Theory heatmap (l) γ line plot comparison

Figure 4: Comparison of the optimal threshold γ obtained from theory and simulations. The first row corresponds to $\tau = 0.01$, the second row corresponds to $\tau = 0.05$, the third row corresponds to $\tau = 0.1$ and the last row corresponds to $\tau = 0.2$. For each value of τ, the left panel shows the empirical heatmap, the middle panel shows the theoretical heatmap, and the right panel compares the theoretical and empirical optimal γ as a function of the join probability.

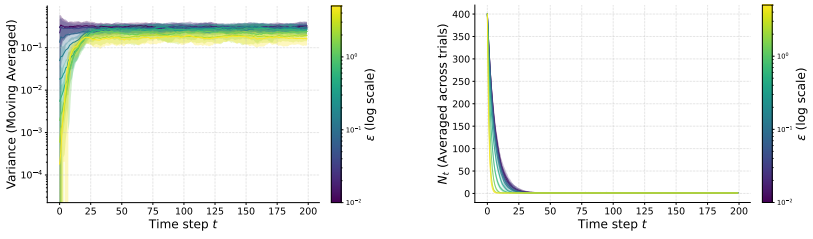


(a) p_τ vs τ (b) $C(\gamma)$ vs γ for $r > q$ (c) $C(\gamma)$ vs γ for $q > r$

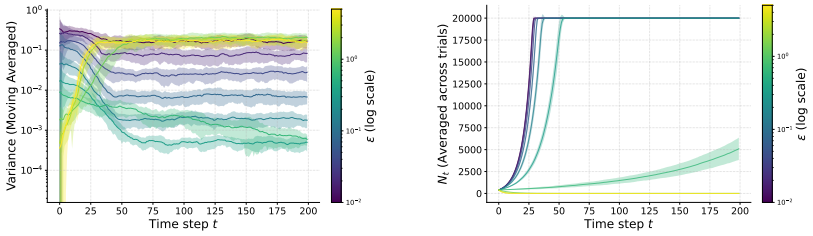
Figure 5: Behavior of p_τ with τ and $C(\gamma)$ with γ, comparing theoretical predictions with empirical simulations. (a) $N = 100$; (b) $q = 0.11, r = 0.52, T = 10$; (c) $q = 0.45, r = 0.11, T = 10$, with the population clipped at 20,000.

Figures 2 and 4 reveal an interesting observation: as τ increases, the empirical noise increasingly underestimates the theoretical prediction because leakage becomes more sensitive to rare, extreme outliers. While the theoretical model assumes an unbounded population, the experimental setup imposes a population cap of 20,000 so the discrepancy between theory and experiment becomes more pronounced at larger τ.

B Detailed Plots from Experiments in Section 3



(a) Moving-averaged variance ($q = 0.6$, $r = 0.3$) (b) Population evolution ($q = 0.6$, $r = 0.3$)

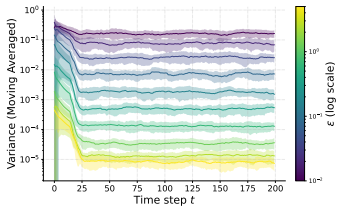


(c) Moving-averaged variance ($q = 0.3$, $r = 0.6$) (d) Population evolution ($q = 0.3$, $r = 0.6$)

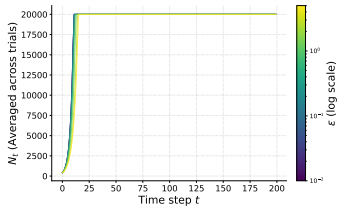
Figure 6: Comparison of moving-averaged variance and population evolution under two parameter settings: the top row corresponds to $(q, r) = (0.6, 0.3)$, while the bottom row corresponds to $(q, r) = (0.3, 0.6)$.

C The Agnostic Model

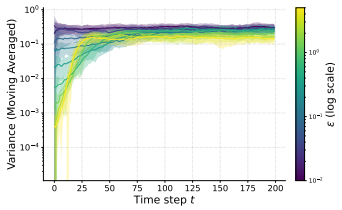
This section explores the agnostic model mentioned in Section 3. Under this framework, we assume population growth is directly proportional to the number of participants in the previous iteration, $t - 1$, regardless of wether or not their data were leaked. Applying a derivation analogous to the



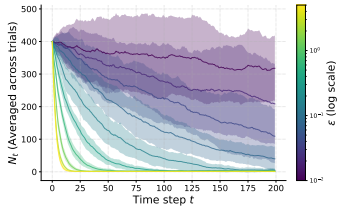
(a) Moving averaged variance for $(r, q) = (0.6, 0.3)$



(b) Population evolution for $(r, q) = (0.6, 0.3)$



(c) Moving averaged variance for $(r, q) = (0.3, 0.6)$



(d) Population evolution for $(r, q) = (0.3, 0.6)$

Figure 7: In a growing population ($r = 0.6, q = 0.3$), the system quickly reaches capacity across all privacy budgets, causing variance to stratify based on noise levels and favoring low privacy due to the high growth rate. In contrast, a shrinking population ($r = 0.3, q = 0.6$) triggers performative collapse; while weak privacy (high ϵ) initially yields higher utility, it accelerates population decline. This causes variance to spike and eventually converge with stricter privacy regimes as the disappearing user base eliminates any initial advantage.

proof of Lemma 6, we characterize the evolution of the participant pool as follows:

$$\mathbb{E}[N_t \mid N_{t-1}] = N_{t-1} - N_{t-1}q\left(\frac{1}{2} + \beta\right) + rN_{t-1} = N_{t-1}\left(1 - q\left(\frac{1}{2} + \beta\right) + r\right) \tag{15}$$

Under this model, we once again simulate the behavior of the population and variance under the same two regimes of $(r, q) = (0.3, 0.6)$ and $(r, q) = (0.6, 0.3)$ respectively as demonstrated in Figure 7. In the regime ($r > q$), rapid growth outweighs attrition, driving the population to maximum capacity regardless of the privacy budget. Unlike the previous model — where low privacy could stifle growth or lead to suboptimal stability, this model provides no incentive for privacy, as the highest ϵ consistently yields the lowest variance. Conversely, when $q > r$, the variance mirrors the original model's behavior, yet population decay is notably more gradual, with certain ϵ values inducing high-variance stochastic collapse across different runs.

We also report the utility-maximizing privacy budget ϵ for each parameter pair in Figure 8. For growing populations ($r > q$), the models diverge sharply: the original model shows a broad transition favoring intermediate privacy, whereas the agnostic model overwhelmingly favors minimal privacy (highest ϵ), as rapid growth drives the population to capacity regardless of data use. Maximum privacy is beneficial only within a narrow band just below $r = q$; outside this regime, strict privacy provides no utility gain.

D Details About Numerical Experiments

To empirically validate our theoretical framework, we simulate both the DP mean estimation and DP-OLS regression mechanisms over a finite horizon of $T = 50$ rounds with an initial population of $N_0 = 1000$ agents and feature dimension $d = 12$. For the mean estimation task, individual data records are drawn from a standard Gaussian $\mathcal{N}(0, I_d)$ for simplicity; while our theoretical analysis explicitly accommodates a shifting true mean μ_t, the stationary setting sufficiently demonstrates the population dynamics. To prove that our results hold robustly under continuous distributional drift,

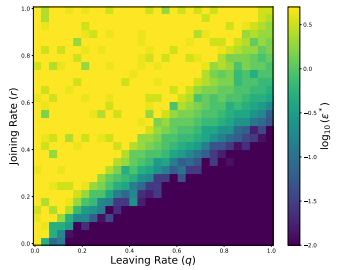


Figure 8: Global mapping of the optimal privacy budget ϵ^* across the (r, q) parameter space for the agnostic model

we implement a dynamic target in the DP-OLS regression experiments: at each round t, the true regression coefficient vector θ_t^* is freshly sampled from the surface of the d-dimensional unit ball, inducing persistent concept drift in the target variables y_i. For evaluating the theoretical cumulative error (from equation 27) in the regression setting, the term ν, which bounds the non-private estimation error, is explicitly configured as $\nu = d\sigma_\epsilon^2$. Model utility is measured via Mahalanobis distance for the mean estimator and via mean squared prediction error on a fresh, held-out test set ($N_{\text{test}} = 2000$) for the regression model. To ensure our evaluation accurately reflects the risk of demographic collapse, any trajectory where the population drops below the necessary OLS rank requirement ($N_t \leq d$) is assigned a fixed maximum penalty loss for all remaining rounds. Complete details regarding data clipping radii, ridge regularization constants for matrix stability, grid search intervals for the noise scale γ, and the full implementation code can be found in our anonymous GitHub repository: (`https://anonymous.4open.science/r/performative-privacy-70F1/`).

E Proofs from Section 3

E.1 Properties of the Estimator

Lemma 4. *The estimator $\hat{\mu}_t$ for population Bernoulli mean μ_t as described in equation 2 is unbiased.*

Proof. To prove that the estimator $\hat{\mu}_t$ is unbiased, we first find the expected value of a single privatized response $Y_{i,t}$. As $X_{i,t}$ is a Bernoulli variable with mean μ_t, we have:

$$\begin{aligned}
\mathbb{E}[Y_{i,t}] &= \mathbb{P}(X_{i,t} = 1)\mathbb{E}[Y_{i,t} \mid X_{i,t} = 1] + \mathbb{P}(X_{i,t} = 0)\mathbb{E}[Y_{i,t} \mid X_{i,t} = 0] \\
&= \mu_t \left(\frac{1}{2} + \beta\right) + (1 - \mu_t)\left(\frac{1}{2} - \beta\right) \\
&= \mu_t \left(\frac{1}{2} + \beta\right) + \frac{1}{2} - \beta - \mu_t \left(\frac{1}{2} - \beta\right) \\
&= \mu_t \left(\frac{1}{2} + \beta - \frac{1}{2} + \beta\right) + \frac{1}{2} - \beta \\
&= 2\beta\mu_t + \frac{1}{2} - \beta
\end{aligned}$$

Now we evaluate the expected value of the sample mean $\hat{\nu}_t$. By linearity of expectation:

$$\begin{aligned}
\mathbb{E}[\hat{\nu}_t] &= \mathbb{E}\left[\frac{1}{N_t}\sum_{i=1}^{N_t} Y_{i,t}\right] = \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbb{E}[Y_{i,t}] \\
&= \frac{1}{N_t} \cdot N_t \left(2\beta\mu_t + \frac{1}{2} - \beta\right) = 2\beta\mu_t + \frac{1}{2} - \beta
\end{aligned}$$

Substituting $\mathbb{E}[\hat{\nu}_t]$ into the expected value of our estimator $\hat{\mu}_t$:

$$\begin{aligned}\mathbb{E}[\hat{\mu}_t] &= \mathbb{E}\left[\frac{\hat{\nu}_t - \frac{1}{2} + \beta}{2\beta}\right] = \frac{\mathbb{E}[\hat{\nu}_t] - \frac{1}{2} + \beta}{2\beta} \\ &= \frac{\left(2\beta\mu_t + \frac{1}{2} - \beta\right) - \frac{1}{2} + \beta}{2\beta} \\ &= \frac{2\beta\mu_t}{2\beta} = \mu_t\end{aligned}$$

This concludes the proof. □

Lemma 5. *For any given privacy budget $\epsilon > 0$, setting the perturbation parameter to*

$$\beta = \frac{1}{2}\tanh\left(\frac{\epsilon}{2}\right)$$

guarantees that the randomized response mechanism satisfies ϵ-local differential privacy (ϵ-LDP).

Proof. To satisfy ϵ-LDP, the maximum ratio of output probabilities (true reporting vs. false reporting) must equal e^{ϵ}:

$$\frac{\frac{1}{2} + \beta}{\frac{1}{2} - \beta} = e^{\epsilon}$$

Solving directly for β gives:

$$\beta = \frac{1}{2}\left(\frac{e^{\epsilon} - 1}{e^{\epsilon} + 1}\right)$$

Using the identity $\tanh(z) = \frac{e^{2z}-1}{e^{2z}+1}$ with $z = \frac{\epsilon}{2}$, this simplifies to:

$$\beta = \frac{1}{2}\tanh\left(\frac{\epsilon}{2}\right)$$

□

E.2 Population Evolution and Critical Noise Threshold

Lemma 6. *The evolution of the number of participants is given by:*

$$\begin{aligned}\mathbb{E}[N_t|N_{t-1}] &= N_{t-1} - N_{t-1}q\left(\frac{1}{2} + \beta\right) + r\left(\frac{1}{2} - \beta\right)N_{t-1} \\ &= N_{t-1}\left(1 - q\left(\frac{1}{2} + \beta\right) + r\left(\frac{1}{2} - \beta\right)\right)\end{aligned}$$

Proof. Let D_t and A_t be the number of departures and arrivals at t. The population at time t follows the basic balance equation:

$$N_t = N_{t-1} - D_t + A_t$$

By construction of the model, each of the N_{t-1} participants independently leaves with probability $q(\frac{1}{2} + \beta)$, recruits a new participant with probability $r(\frac{1}{2} - \beta)$, or does neither, so that, conditionally on N_{t-1}, A_t and D_t follow Binomial$(N_{t-1}, r(1/2 - \beta))$ and Binomial$(N_{t-1}, q(\beta + 1/2))$ respectively (jointly, $(D_t, A_t, N_{t-1} - D_t - A_t)$ is multinomial).

Thus, $\mathbb{E}[A_t|N_{t-1}] = r\left(\frac{1}{2} - \beta\right)N_{t-1}$ and $\mathbb{E}[D_t|N_{t-1}] = N_{t-1}q\left(\frac{1}{2} + \beta\right)$. Taking the conditional expectation of the balance equation yields, exactly:

$$\begin{aligned}\mathbb{E}[N_t|N_{t-1}] &= N_{t-1} - N_{t-1}q\left(\frac{1}{2} + \beta\right) + r\left(\frac{1}{2} - \beta\right)N_{t-1} \\ &= N_{t-1}\left(1 - q\left(\frac{1}{2} + \beta\right) + r\left(\frac{1}{2} - \beta\right)\right)\end{aligned}$$

□

Theorem 1. *To maintain $C(\beta) \geq 1$, the parameter β must satisfy the explicit upper bound $\beta \leq \beta_c$, where the critical threshold β_c is defined as:*

$$\beta_c \triangleq \max\left\{0, \frac{r-q}{2(q+r)}\right\}$$

For $r > q$, this is equivalent to $\epsilon \leq \epsilon_c \triangleq \ln(r/q)$.

Proof. Substituting the simplified linear expression for the performative growth factor yields:

$$1 + \frac{r-q}{2} - \beta(q+r) \geq 1$$
$$\implies \frac{r-q}{2} - \beta(q+r) \geq 0$$

Rearranging the inequality to isolate the β term yields:

$$\beta(q+r) \leq \frac{r-q}{2}$$

Dividing both sides by $(q+r)$,

$$\beta \leq \frac{r-q}{2(q+r)}$$

However, when $r \leq q$, the critical point get clipped to zero. Thus,

$$\beta \leq \max\left\{0, \frac{r-q}{2(r+q)}\right\}$$

Finally, for $r > q$, since $2\beta = \tanh(\epsilon/2) = \frac{e^\epsilon - 1}{e^\epsilon + 1}$ is increasing in ϵ, $\beta \leq \beta_c$ is equivalent to $(e^\epsilon - 1)(r+q) \leq (r-q)(e^\epsilon + 1)$, i.e., $qe^\epsilon \leq r$, i.e., $\epsilon \leq \ln(r/q)$. This completes the proof. □

E.3 Galton–Watson Analysis of the Population

Lemma 6 and Theorem 1 only describe the expected population. Since leakage, departures, and recruitments are independent across participants and rounds, (N_t) is in fact exactly a Galton–Watson branching process (Athreya & Ney, 1972), which yields an exact description of the random population.

Theorem 4 (Extinction and Growth)**.** *Let $q, r > 0$, $\epsilon_c = \ln(r/q)$, and $\sigma^2(\beta) \triangleq q(\frac{1}{2} + \beta) + r(\frac{1}{2} - \beta) - (C(\beta) - 1)^2$ the offspring variance. Starting from N_0 participants, the population goes extinct with probability*

$$\Pr(\exists t,\ N_t = 0) = \min\left(1, \frac{q}{r}e^\epsilon\right)^{N_0} = e^{-N_0(\epsilon_c - \epsilon)_+}.$$

If $\epsilon < \epsilon_c$, $N_t/C(\beta)^t$ converges almost surely to a random variable W, positive exactly on survival, with $\mathbb{E}[W] = N_0$ and $\mathrm{Var}(W) = N_0\sigma^2(\beta)/\big(C(\beta)(C(\beta)-1)\big)$.

In particular, if $q \geq r$, then $\epsilon_c \leq 0$ and the population goes extinct almost surely for every ϵ. If $r > q$, extinction is only likely within $\mathcal{O}(1/N_0)$ of ϵ_c, and the deterministic trajectory $N_0 C(\beta)^t$ is accurate up to the random factor W/N_0, whose relative standard deviation is small unless $C(\beta) - 1 = \mathcal{O}(1/N_0)$.

Proof. **Branching structure.** Let $a \triangleq q(\frac{1}{2} + \beta)$ and $c \triangleq r(\frac{1}{2} - \beta)$. Conditionally on the past, each of the N_t participants of round t independently produces $\xi \in \{0, 1, 2\}$ participants for round $t+1$: $\xi = 0$ if their answer is leaked and they leave (probability a), $\xi = 2$ if it is not leaked and they recruit (probability c), and $\xi = 1$ otherwise. Hence (N_t) is a Galton–Watson process (Athreya & Ney, 1972) with offspring generating function $f(s) = a + (1 - a - c)s + cs^2$, offspring mean $f'(1) = 1 - a + c = C(\beta)$, and offspring variance $f''(1) + f'(1) - f'(1)^2 = 2c + (1 - a + c) - (1 - a + c)^2 = a + c - (c - a)^2 = \sigma^2(\beta)$. By Lemma 5, $\frac{1/2+\beta}{1/2-\beta} = e^\epsilon$, so that $\frac{a}{c} = \frac{q}{r}e^\epsilon = e^{\epsilon - \epsilon_c}$.

Extinction. Since $a > 0$, the offspring distribution is not concentrated on 1, so the extinction probability of the process started from a single participant is the smallest root in $[0, 1]$ of $f(s) = s$

(Athreya & Ney, 1972). Here, $f(s) - s = cs^2 - (a+c)s + a = (s-1)(cs-a)$, whose roots are 1 and a/c, so this probability is $\min(1, a/c) = \min(1, e^{\epsilon - \epsilon_c})$. The N_0 initial participants generate independent Galton–Watson processes whose sum is (N_t), and the population goes extinct if and only if all of them do, which happens with probability $\min(1, e^{\epsilon-\epsilon_c})^{N_0} = e^{-N_0(\epsilon_c - \epsilon)_+}$.

Growth. Assume $\epsilon < \epsilon_c$, i.e., $a < c$ and $C \triangleq C(\beta) > 1$, and let $M_t \triangleq N_t / C^t$. Since $\mathbb{E}[N_{t+1} \mid N_t] = CN_t$, (M_t) is a nonnegative martingale with $\mathbb{E}[M_t] = N_0$. By the law of total variance, $\mathrm{Var}(N_{t+1}) = \sigma^2 \mathbb{E}[N_t] + C^2 \mathrm{Var}(N_t) = \sigma^2 N_0 C^t + C^2 \mathrm{Var}(N_t)$, so that $\mathrm{Var}(M_{t+1}) = \mathrm{Var}(M_t) + \sigma^2 N_0 C^{-t-2}$ and, as $\mathrm{Var}(M_0) = 0$,

$$\mathrm{Var}(M_t) = \sigma^2 N_0 \sum_{k=0}^{t-1} C^{-k-2} = \frac{N_0 \sigma^2 (1 - C^{-t})}{C(C-1)}.$$

The martingale (M_t) is thus bounded in L^2, so it converges almost surely and in L^2 to a random variable W with $\mathbb{E}[W] = N_0$ and $\mathrm{Var}(W) = N_0 \sigma^2 / (C(C-1))$. Finally, since the offspring number is bounded, $\mathbb{E}[\xi \log \xi] < \infty$, and the Kesten-Stigum theorem (Kesten & Stigum, 1966) gives $\Pr(W = 0) = \Pr(\text{extinction})$. □

Remark 4 (Beyond randomized response)**.** *In Sections 4 and 5, conditionally on the release noise η_t, each round is still approximately a Galton–Watson generation, with $\frac{1}{2} + \beta$ replaced by a leakage probability depending on η_t and N_t, but two features break the Galton–Watson structure. First, η_t is shared by all users, so the fraction of leaked users fluctuates across rounds without averaging out over users: (N_t) is a branching process in a random environment, whose almost sure extinction is characterized, for i.i.d. environments, by the expected logarithm of the random growth factor rather than by its mean (Smith & Wilkinson, 1969). Second, the leakage probability decreases with N_t (proof of Lemma 11), so the growth factor increases with the population size, a positive dependence known in population ecology as an Allee effect (Allee & Bowen, 1932; Courchamp et al., 1999). This is why we rely on the mean-field approximation in these sections.*

F Proofs from Section 4

F.1 Generic Results

Lemma 7. *Let Z_t be a dataset drawn from a distribution with true mean μ and covariance matrix Σ, and neglect the clipping bias (i.e., assume $\bar{z}_i = z_i$). Assuming the aggregated noise vector η_d has covariance $C_{\boldsymbol{\gamma}} = \frac{\gamma^2}{N_t} I$, the expected squared Mahalanobis error is:*

$$\mathbb{E}\left[\|o_t - \mu_t\|^2_{\Sigma^{-1}} \;\middle|\; N_t\right] = \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_t} \tag{16}$$

Proof. Since the mechanism's injected noise η_d is sampled independently from the underlying data distribution $\mathcal{D}$, the expected squared error decouples into two components: the variance introduced by the privacy mechanism and the intrinsic geometric variance of the data sample itself.

By applying the linearity of expectation and utilizing the trace trick for quadratic forms ($\mathbb{E}[x^T A x] = \mathrm{Tr}(A\Sigma_x) + \mu_x^T A \mu_x$), we expand the conditional error:

$$\begin{aligned}
\mathbb{E}\left[\|o_t - \mu_t\|^2_{\Sigma^{-1}} \;\middle|\; N_t\right] &= \mathbb{E}_{\eta_d}\left[\|\eta_d\|^2_{\Sigma^{-1}}\right] + \frac{1}{N_t^2}\mathbb{E}_{\mathcal{D}}\left[\Big\| \sum_{z_i \in Z_t} (z_i - \mu)\Big\|^2_{\Sigma^{-1}}\right] \\
&= \mathrm{Tr}(\Sigma^{-1} C_{\boldsymbol{\gamma}}) + \frac{1}{N_t^2} \sum_{z_i \in Z_t} \mathbb{E}_{\mathcal{D}}\left[\|z_i - \mu_t\|^2_{\Sigma^{-1}}\right] \\
&= \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1})}{N_t} + \frac{1}{N_t^2} \sum_{z_i \in Z_t} \mathrm{Tr}(\Sigma^{-1}\Sigma) \\
&= \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_t}
\end{aligned}$$

where we substitute $C_{\boldsymbol{\gamma}} = \frac{\gamma^2}{N_t} I$ and recognize that $\mathrm{Tr}(\Sigma^{-1}\Sigma) = \mathrm{Tr}(I_d) = d$. This bounds the conditional expected error, completing the proof. □

Lemma 1. *Let Z_{t-1} be the dataset at round t and $z \in Z_t$ be a target record. Let $o_t \sim \mathcal{M}(Z_t)$ be the candidate output. Define the factual and counterfactual output distributions as $P_{in}^{\mathcal{M}}(\cdot) = \Pr[\mathcal{M}(Z_t) = \cdot \mid z \in Z_t]$ and $P_{out}^{\mathcal{M}}(\cdot) = \Pr[\mathcal{M}(Z_t \setminus \{z\}) = \cdot]$.*

For the log-likelihood ratio $\Lambda(o_t|z) = \ln\left(\frac{P_{in}^{\mathcal{M}}(o_t)}{P_{out}^{\mathcal{M}}(o_t)}\right)$ evaluated at output o_t, and for any threshold $\tau > 0$ and order $\alpha > 1$, the tail probability of structural compromisation is bounded by:

$$\Pr_{o_t \sim P_{in}^{\mathcal{M}}}(\Lambda(o_t|z) > \tau) \leq \exp\Big(- (\alpha - 1)\big(\tau - D_\alpha(P_{in}^{\mathcal{M}} \parallel P_{out}^{\mathcal{M}})\big)\Big)$$

where $D_\alpha(P_{in}^{\mathcal{M}} \parallel P_{out}^{\mathcal{M}})$ is the Rényi divergence of order α.

Proof. We use the Cramer-Chernoff technique to bound the leakage probability of z. By the monotonicity of the exponential function, for any $\alpha > 1$, the tail event $\Lambda(o_t|z) > \tau$ is strictly equivalent to:

$$e^{(\alpha-1)\Lambda(o_t|z)} > e^{(\alpha-1)\tau}$$

Applying Markov's inequality to the non-negative random variable $e^{(\alpha-1)\Lambda(o_t|z)}$ over the factual distribution $P_{\text{in}}^{\mathcal{M}}$ yields:

$$\Pr_{o_t \sim P_{\text{in}}^{\mathcal{M}}}(\Lambda(o_t|z) > \tau) \leq \frac{\mathbb{E}_{o_t \sim P_{\text{in}}^{\mathcal{M}}}\left[e^{(\alpha-1)\Lambda(o_t|z)}\right]}{e^{(\alpha-1)\tau}} \tag{17}$$

Expanding the definition of $\Lambda(o_t|z)$ within the expectation, we have:

$$\mathbb{E}_{o_t \sim P_{\text{in}}^{\mathcal{M}}}\left[e^{(\alpha-1)\ln\left(\frac{P_{\text{in}}^{\mathcal{M}}(o_t)}{P_{\text{out}}^{\mathcal{M}}(o_t)}\right)}\right] = \mathbb{E}_{o_t \sim P_{\text{in}}^{\mathcal{M}}}\left[\left(\frac{P_{\text{in}}^{\mathcal{M}}(o_t)}{P_{\text{out}}^{\mathcal{M}}(o_t)}\right)^{\alpha-1}\right]$$

By the definition of Rényi divergence, $D_\alpha(P_{\text{in}}^{\mathcal{M}} \parallel P_{\text{out}}^{\mathcal{M}}) = \frac{1}{\alpha-1}\ln \mathbb{E}_{P_{\text{in}}^{\mathcal{M}}}\left[\left(\frac{P_{\text{in}}^{\mathcal{M}}(o_t)}{P_{\text{out}}^{\mathcal{M}}(o_t)}\right)^{\alpha-1}\right]$. Exponentiating both sides to isolate the expectation yields:

$$\mathbb{E}_{o_t \sim P_{\text{in}}^{\mathcal{M}}}\left[\left(\frac{P_{\text{in}}^{\mathcal{M}}(o_t)}{P_{\text{out}}^{\mathcal{M}}(o_t)}\right)^{\alpha-1}\right] = \exp\Big((\alpha-1)D_\alpha(P_{\text{in}}^{\mathcal{M}} \parallel P_{\text{out}}^{\mathcal{M}})\Big)$$

Substituting this equivalence back into the numerator of equation 17 gives:

$$\Pr_{o_t \sim P_{\text{in}}^{\mathcal{M}}}(\Lambda(o_t|z) > \tau) \leq \frac{\exp\Big((\alpha-1)D_\alpha(P_{\text{in}}^{\mathcal{M}} \parallel P_{\text{out}}^{\mathcal{M}})\Big)}{\exp\Big((\alpha-1)\tau\Big)}$$

Combining the exponents completes the proof:

$$\Pr_{o_t \sim P_{\text{in}}^{\mathcal{M}}}(\Lambda(o_t|z) > \tau) \leq \exp\Big(- (\alpha - 1)\big(\tau - D_\alpha(P_{\text{in}}^{\mathcal{M}} \parallel P_{\text{out}}^{\mathcal{M}})\big)\Big)$$

□

F.2 Results for Gaussian Mechanism

Lemma 8. *Let $\mathcal{M}$ be a Gaussian mechanism for mean estimation over a dataset Z_t of size N_t, which adds noise scaled to the population such that the variance is $\frac{\gamma^2}{N_t}I$. Let $P_{in}^{\mathcal{M}}(\cdot \mid Z_t)$ and $P_{out}^{\mathcal{M}}(\cdot \mid Z_t)$ denote the conditional distributions of the mechanism's output on adjacent datasets Z_t and $Z_t \setminus \{z\}$, respectively. For any order $\alpha > 1$, the Rényi divergence between these distributions is given by:*

$$D_\alpha(P_{in}^{\mathcal{M}}(\cdot \mid Z_t)||P_{out}^{\mathcal{M}}(\cdot \mid Z_t)) = \frac{\alpha N_t}{2\gamma^2(N_t-1)^2}\|z - \mu_{z,t}\|_2^2$$

Proof. Let the output distributions of the mechanism on the adjacent datasets be defined as $P_{\text{in}}^{\mathcal{M}}(\cdot \mid Z_t) = \mathcal{N}(\mu_{z,t}, C_\gamma)$ and $P_{\text{out}}^{\mathcal{M}}(\cdot \mid Z_t) = \mathcal{N}(\mu_{-z}, C_\gamma)$, with shared covariance matrix $C_\gamma = \frac{\gamma^2}{N_t} I_{d\times d}$. Assuming the mechanism computes the mean using the respective dataset sizes, the true, un-noised mean vectors and their difference are:

$$\mu_{z,t} = \frac{1}{N_t}\sum_{x\in Z_t} x, \quad \mu_{-z} = \frac{1}{N_t - 1}\sum_{x\in Z_t\setminus\{z\}} x, \quad \mu_{z,t} - \mu_{-z} = \frac{z - \mu_{z,t}}{N_t - 1}$$

For two multivariate Gaussian distributions sharing the same covariance matrix C_γ, the Rényi divergence of order α simplifies to the scaled Mahalanobis distance between their means. Substituting the known difference and the inverse covariance matrix $C_\gamma^{-1} = \frac{N_t}{\gamma^2} I$, we obtain:

$$\begin{aligned} D_\alpha(P_{\text{in}}^{\mathcal{M}}(\cdot \mid Z_t) || P_{\text{out}}^{\mathcal{M}}(\cdot \mid Z_t)) &= \frac{\alpha}{2}(\mu_{z,t} - \mu_{-z})^\top C_\gamma^{-1}(\mu_{z,t} - \mu_{-z}) \\ &= \frac{\alpha}{2}\left(\frac{z - \mu_{z,t}}{N_t - 1}\right)^\top \left(\frac{N_t}{\gamma^2} I\right)\left(\frac{z - \mu_{z,t}}{N_t - 1}\right) \\ &= \frac{\alpha}{2}\cdot\frac{N_t}{\gamma^2}\cdot\frac{1}{(N_t-1)^2}(z - \mu_{z,t})^T I (z - \mu_{z,t}) \\ &= \frac{\alpha N_t}{2\gamma^2(N_t-1)^2}\|z - \mu_{z,t}\|_2^2 \end{aligned}$$

□

Lemma 9. *Let a mechanism output o_t be perturbed by additive isotropic Gaussian noise $\eta_t \sim \mathcal{N}(0, \frac{\gamma^2}{N_t} I_{d\times d})$, representing the noise. The log-likelihood ratio $\Lambda(o_t|z)$ conditioned on the dataset Z_t is given by:*

$$\Lambda(o_t|z) \mid Z_t = \frac{N_t\|z - \mu_{z,t}\|^2}{2(N_t-1)^2\gamma^2} + \frac{N_t\langle z - \mu_{z,t}, \eta_t\rangle}{(N_t-1)\gamma^2}$$

And it is distributed as:

$$\Lambda(o_t|z) \mid Z_t \sim \mathcal{N}\left(\frac{N_t\|z - \mu_{z,t}\|^2}{2(N_t-1)^2\gamma^2}, \ \frac{N_t\|z - \mu_{z,t}\|^2}{(N_t-1)^2\gamma^2}\right)$$

Proof. By definition, the log-likelihood ratio compares the probability density of observing the noisy output o_t when z is present (with mean $\mu_{z,t}$) versus when z is absent (with mean μ_{-z}). Because the aggregated additive noise η_t is Gaussian with distribution $\mathcal{N}(0, \frac{\gamma^2}{N_t} I)$, the log-likelihood conditioned on Z_t evaluates to:

$$\begin{aligned} \Lambda(o_t|z) \mid Z_t &= \ln\left(\frac{\exp\left(-\frac{N_t}{2\gamma^2}\|o_t - \mu_{z,t}\|^2\right)}{\exp\left(-\frac{N_t}{2\gamma^2}\|o_t - \mu_{-z}\|^2\right)}\right) \\ &= \frac{N_t}{2\gamma^2}\Big(\|o_t - \mu_{-z}\|^2 - \|o_t - \mu_{z,t}\|^2\Big) \end{aligned}$$

Substituting the noisy observation $o_t = \mu_{z,t} + \eta_t$ and the shift $\mu_{-z} = \mu_{z,t} - \frac{z - \mu_{z,t}}{N_t - 1}$:

$$\begin{aligned} \Lambda(o_t|z) \mid Z_t &= \frac{N_t}{2\gamma^2}\left(\left\|\mu_{z,t} + \eta_t - \left(\mu_{z,t} - \frac{z - \mu_{z,t}}{N_t - 1}\right)\right\|^2 - \|\mu_{z,t} + \eta_t - \mu_{z,t}\|^2\right) \\ &= \frac{N_t}{2\gamma^2}\left(\left\|\frac{z - \mu_{z,t}}{N_t - 1} + \eta_t\right\|^2 - \|\eta_t\|^2\right) \\ &= \frac{N_t}{2\gamma^2}\left(\frac{\|z - \mu_{z,t}\|^2}{(N_t-1)^2} + \frac{2}{N_t - 1}\langle z - \mu_{z,t}, \eta_t\rangle + \|\eta_t\|^2 - \|\eta_t\|^2\right) \\ &= \frac{N_t\|z - \mu_{z,t}\|^2}{2(N_t-1)^2\gamma^2} + \frac{N_t\langle z - \mu_{z,t}, \eta_t\rangle}{(N_t-1)\gamma^2} \end{aligned}$$

Conditioned on Z_t, the term $\frac{N_t\|z-\mu_{z,t}\|^2}{2(N_t-1)^2\gamma^2}$ is a deterministic scalar constant. The sole source of randomness is the inner product with the aggregated noise vector $\eta_t \sim \mathcal{N}(0, \frac{\gamma^2}{N_t}I)$. This linear combination $\langle z-\mu_{z,t}, \eta_t\rangle$ forms a univariate Gaussian with mean 0 and variance

$$\frac{\gamma^2}{N_t}\|z-\mu_{z,t}\|^2.$$

Scaling this by $\frac{N_t}{(N_t-1)\gamma^2}$, the conditional variance reduces to:

$$\mathbb{V}\left(\frac{N_t\langle z-\mu_{z,t},\eta_t\rangle}{(N_t-1)\gamma^2}\middle| Z_t\right) = \frac{N_t^2}{(N_t-1)^2\gamma^4}\left(\frac{\gamma^2}{N_t}\|z-\mu_{z,t}\|^2\right) = \frac{N_t\|z-\mu_{z,t}\|^2}{(N_t-1)^2\gamma^2}$$

Thus,

$$\Lambda(o_t|z) \mid Z_t \sim \mathcal{N}\left(\frac{N_t\|z-\mu_{z,t}\|^2}{2(N_t-1)^2\gamma^2}, \frac{N_t\|z-\mu_{z,t}\|^2}{(N_t-1)^2\gamma^2}\right).$$

This completes the proof. □

Lemma 2. *Given the aggregated empirical mean mechanism, the exact probability that a specific individual data point z is exposed (i.e., its log-likelihood ratio breaches the threshold τ) conditioned on Z_t is given in closed form by:*

$$\Pr(\Lambda(o_t|z) > \tau \mid Z_t) = 1 - \Phi\left(\frac{\tau(N_t-1)\gamma}{\sqrt{N_t}\|z-\mu_{z,t}\|} - \frac{\sqrt{N_t}\|z-\mu_{z,t}\|}{2(N_t-1)\gamma}\right)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

Proof. From lemma 9, the conditional log-likelihood ratio for a specific data point z and $\mu_{z,t}$ strictly follows a Gaussian distribution:

$$\Lambda(o_t|z) \mid Z_t \sim \mathcal{N}\left(\mu'_{z,t}, \sigma_z^2\right) \quad \text{where} \quad \mu'_{z,t} = \frac{N_t\|z-\mu_{z,t}\|^2}{2(N_t-1)^2\gamma^2}, \;\; \sigma_z = \frac{\sqrt{N_t}\|z-\mu_{z,t}\|}{(N_t-1)\gamma} \tag{18}$$

The probability that $\Lambda(o_t|z)$ exceeds τ is precisely the tail probability of this Gaussian distribution. We evaluate this using the standard normal cumulative distribution function Φ:

$$\begin{aligned}\Pr(\Lambda(o_t|z) > \tau \mid Z_t) &= \Pr\left(\frac{\Lambda(o_t|z)-\mu'_{z,t}}{\sigma_z} > \frac{\tau-\mu'_{z,t}}{\sigma_z}\middle| Z_t\right)\\ &= 1-\Phi\left(\frac{\tau-\mu'_{z,t}}{\sigma_z}\right)\end{aligned}$$

Substituting the explicit formulae for $\mu'_{z,t}$ and σ_z from equation 18:

$$\begin{aligned}\Pr(\Lambda(o_t|z) > \tau \mid Z_t) &= 1-\Phi\left(\frac{\tau-\frac{N_t\|z-\mu_{z,t}\|^2}{2(N_t-1)^2\gamma^2}}{\frac{\sqrt{N_t}\|z-\mu_{z,t}\|}{(N_t-1)\gamma}}\right)\\ &= 1-\Phi\left(\frac{\tau(N_t-1)\gamma}{\sqrt{N_t}\|z-\mu_{z,t}\|} - \frac{N_t\|z-\mu_{z,t}\|^2(N_t-1)\gamma}{2(N_t-1)^2\gamma^2\sqrt{N_t}\|z-\mu_{z,t}\|}\right)\\ &= 1-\Phi\left(\frac{\tau(N_t-1)\gamma}{\sqrt{N_t}\|z-\mu_{z,t}\|} - \frac{\sqrt{N_t}\|z-\mu_{z,t}\|}{2(N_t-1)\gamma}\right)\end{aligned}$$

This completes the proof. □

F.3 Comparison of the Generic Bound with the Exact Gaussian Leakage

The bound of Lemma 1 applies to the output of any DP mechanism, as it only depends on the Rényi divergence. It decays exponentially with the threshold τ, and its tightest version is obtained by optimizing over the order α, which must remain strictly greater than 1. We illustrate it on the Gaussian mechanism, conditioning the leakage probability on Z_t for analytical tractability, and compare it with the exact expression of Lemma 2.

Lemma 10 (Upper Bound for the Gaussian Mechanism)**.** *Substituting into Lemma 1 the Rényi divergence of the Gaussian mechanism equation 5 between the outputs with and without the record z conditioned on Z_t given by Lemma 8, the tail bound on the leakage probability of a record z specializes to:*

$$\Pr_{o_t \sim P_{in}^{\mathcal{M}}}(\Lambda(o_t|z) > \tau \mid Z_t) \leq \exp\left(-(\alpha-1)\left(\tau - \frac{\alpha N_t \|z-\mu_{z,t}\|^2}{2\gamma^2(N_t-1)^2}\right)\right). \tag{19}$$

This bound makes explicit, first, the dependence on $\|z-\mu_{z,t}\|^2$: outliers face an exponentially increasing probability of leakage as their deviation from the empirical mean grows. Second, when the population size N_t increases, the divergence decreases, capturing the privacy amplification afforded by aggregation and leading to a virtuous circle: if the system is in a regime private enough to gain users, it becomes more likely to keep growing as the leakage risk decreases.

Minimizing equation 19 with respect to α yields the optimal order

$$\alpha^* = \frac{1}{2} + \frac{\tau\gamma^2(N_t-1)^2}{N_t\|z-\mu_{z,t}\|^2} \qquad \text{subject to } \alpha^* > 1.$$

Thus, the optimized bound is valid and non-vacuous only when

$$\tau > a \triangleq \frac{N_t\|z-\mu_{z,t}\|^2}{2\gamma^2(N_t-1)^2},$$

i.e., only when the individual's privacy tolerance strictly exceeds their deviation from the empirical mean rescaled by the current parameters. This condition fails precisely for extreme outliers and in low-noise regimes ($\gamma \to 0$), where $\alpha^* \leq 1$ either invalidates the divergence or trivializes the bound (i.e., it becomes larger than 1). For $\tau > a$, substituting α^* back into equation 19 yields the optimized bound

$$\Pr_{o_t \sim P_{\text{in}}^{\mathcal{M}}}(\Lambda(o_t|z) > \tau \mid Z_t) \leq \exp\left(-\frac{(\tau-a)^2}{4a}\right). \tag{20}$$

Since Lemma 9 shows that $\Lambda(o_t|z) \mid Z_t \sim \mathcal{N}(a, 2a)$ exactly, the standard Gaussian tail bound $1-\Phi(u) \leq e^{-u^2/2}$ applied at $u = (\tau-a)/\sqrt{2a}$ gives the same exponent: the optimized Rényi bound is tight up to the Φ-versus-exp gap. Conversely, unlike the generic bound, which is conservative and subject to the constraint $\tau > a$, Lemma 2 characterizes the leakage probability exactly, including for outliers.

F.4 Population Evolution and Critical Noise Threshold

Assumption 1. *The records are drawn independently and identically at each round t from a d-dimensional multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, with probability density function*

$$f_{\mathcal{N}}(z) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right).$$

Lemma 3. *Given assumption 1 and the performative update rule in mechanism 1, the expected population size at round $t+1$, conditioned on the current population N_t, is given by:*

$$\mathbb{E}[N_{t+1} \mid N_t] = N_t\Big(1 + r - (q+r)p_\tau^{(t)}\Big)$$

Proof. Let $\mathbb{1}(\Lambda(o_t|z_i) > \tau)$ be the indicator variable denoting whether the i-th data point experiences a privacy leakage exceeding the threshold τ at round t. Averaging the independent departure and recruitment coins, conditional on the dataset and released output, gives:

$$\mathbb{E}[N_{t+1} \mid Z_t, o_t] = N_t - q\sum_{i=1}^{N_t}\mathbb{1}\big(\Lambda(o_t|z_i) > \tau\big) + r\left(N_t - \sum_{i=1}^{N_t}\mathbb{1}\big(\Lambda(o_t|z_i) > \tau\big)\right)$$

Where $z_i \in Z_t$. Grouping the terms yields:

$$\mathbb{E}[N_{t+1} \mid Z_t, o_t] = N_t(1+r) - (q+r)\sum_{i=1}^{N_t}\mathbb{1}\big(\Lambda(o_t|z_i) > \tau\big)$$

Taking the conditional expectation given N_t, we evaluate the joint expectation over the entire dataset $Z_t = \{z_1, \ldots, z_{N_t}\}$ and the mechanism's noise η_t:

$$\mathbb{E}[N_{t+1} \mid N_t] = N_t(1+r) - (q+r)\mathbb{E}_{Z_t,\eta_t}\left[\sum_{i=1}^{N_t} \mathbb{1}\big(\Lambda(o_t|z_i) > \tau\big)\right]$$

By Fubini's theorem, we distribute the expectation inside the sum. As established, $\Lambda(o_t|z_i)$ for a specific point depends on z_i, $\mu_{z,t}$, and the noise η_t. We evaluate the expectation by utilizing the tower property of expectation:

$$\mathbb{E}_{Z_t,\eta_t}\left[\sum_{i=1}^{N_t} \mathbb{1}\big(\Lambda(o_t|z_i) > \tau\big)\right] = \sum_{i=1}^{N_t} \mathbb{E}_{Z_t}\left[\Pr\big(\Lambda(o_t|z_i) > \tau \mid Z_t\big)\right]$$

From lemma 2, the conditional probability of leakage for any specific data point z is given as:

$$\Pr(\Lambda(o_t|z) > \tau \mid Z_t) = 1 - \Phi\left(\frac{\tau(N_t-1)\gamma}{\sqrt{N_t}\|z-\mu_{z,t}\|} - \frac{\sqrt{N_t}\|z-\mu_{z,t}\|}{2(N_t-1)\gamma}\right)$$

Since all N_t data points are identically distributed, we proceed as follows:

$$\begin{aligned}\mathbb{E}[N_{t+1} \mid N_t] &= N_t(1+r) \\ &\quad - (q+r)N_t\mathbb{E}_{Z_t}\left[1 - \Phi\left(\frac{\tau(N_t-1)\gamma}{\sqrt{N_t}\|z-\mu_{z,t}\|} - \frac{\sqrt{N_t}\|z-\mu_{z,t}\|}{2(N_t-1)\gamma}\right)\right]\end{aligned}$$

Since mechanism $\mathcal{M}$ in algorithm 1 clips the samples to $\bar{z}$, we denote the empirical mean on the clipped samples as $\mu_{\bar{z},t}$. Thus, the marginal leakage probability for z is

$$p_\tau^{(t)} = \mathbb{E}_{Z_t}\left[1 - \Phi\left(\frac{\tau(N_t-1)\gamma}{\sqrt{N_t}\|\bar{z}-\mu_{\bar{z},t}\|} - \frac{\sqrt{N_t}\|\bar{z}-\mu_{\bar{z},t}\|}{2(N_t-1)\gamma}\right)\right].$$

Which expands as shown in equation 9. Therefore, our final expression is:

$$\begin{aligned}\mathbb{E}[N_{t+1} \mid N_t] &= N_t(1+r) - (q+r)N_t p_\tau^{(t)} \\ &= N_t\Big(1 + r - (q+r)p_\tau^{(t)}\Big)\end{aligned}$$

This completes the proof. □

Theorem 2. *The condition $C_t(\gamma) \geq 1$ is guaranteed as soon as the mechanism's noise scale satisfies the explicit lower bound $\gamma \geq \gamma_c$. For $q, r > 0$, $N_t > 1$ and clipping bound R, a sufficient threshold γ_c is defined as:*

$$\gamma_c \triangleq \frac{2R\sqrt{N_t}}{N_t - 1}\left(\frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau}\right)$$

where $\Phi_{q,r}^{-1} \triangleq \Phi^{-1}\left(\frac{q}{q+r}\right)$ is the probit function (the inverse of the standard normal cumulative distribution function) evaluated at the performative survival threshold.

Proof. Nonnegative expected growth requires $C(\gamma) \geq 1$. Given the definition of the performative growth factor, this restricts the maximum permissible expected marginal leakage probability p_τ:

$$C(\gamma) \geq 1 \implies 1 + r - (q+r)p_\tau \geq 1 \implies p_\tau \leq \frac{r}{q+r}$$

Clipping gives the uniform bound

$$\|\bar{z} - \mu_{\bar{z},t}\| \leq 2R.$$

It is therefore sufficient to require:

$$1 - \Phi\left(\frac{\tau(N_t-1)\gamma}{2R\sqrt{N_t}} - \frac{R\sqrt{N_t}}{(N_t-1)\gamma}\right) \leq \frac{r}{q+r}.$$

Rearranging to isolate the CDF and applying the Probit function (Φ^{-1}) yields:

$$\frac{\tau(N_t-1)\gamma}{2R\sqrt{N_t}} - \frac{R\sqrt{N_t}}{(N_t-1)\gamma} \geq \Phi^{-1}\left(1-\frac{r}{q+r}\right) = \Phi^{-1}\left(\frac{q}{q+r}\right) \triangleq \Phi_{q,r}^{-1}. \tag{21}$$

Let

$$u = \frac{(N_t-1)\gamma}{2R\sqrt{N_t}}.$$

Substituting u transforms equation 21 into a standard quadratic equation:

$$\tau u - \frac{1}{2u} \geq \Phi_{q,r}^{-1} \implies 2\tau u^2 - 2\Phi_{q,r}^{-1}u - 1 \geq 0.$$

Taking the valid positive root of the quadratic provides the sufficient lower bound for u:

$$u \geq \frac{2\Phi_{q,r}^{-1} + \sqrt{4(\Phi_{q,r}^{-1})^2 + 8\tau}}{4\tau} = \frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau}. \tag{22}$$

Substituting $u = \frac{(N_t-1)\gamma}{2R\sqrt{N_t}}$ back into equation 22 and using $R = \Theta(\sqrt{d})$, we get:

$$\begin{aligned}\gamma &\geq \frac{2R\sqrt{N_t}}{N_t-1}\left(\frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau}\right) \triangleq \gamma_c \\ &= \Theta\left(\frac{\sqrt{dN_t}}{N_t-1}\left(\frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau}\right)\right)\end{aligned}$$

This establishes a sufficient noise level for nonnegative expected growth, completing the proof. □

F.5 Bounds and Asymptotics of the Error

Theorem 5. *Under our assumptions, conditioned on $N_{t-1} \geq 2$, the expected squared Mahalanobis error of the private empirical mean o_t is bounded from below and above by:*

$$\frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{C_{t-1}(\gamma)N_{t-1} + 1} \leq \mathbb{E}\left[\|o_t - \mu_t\|_{\Sigma^{-1}}^2 \;\middle|\; N_{t-1}\right] \leq \kappa_{q,r}\,\frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{C_{t-1}(\gamma)N_{t-1}} \tag{23}$$

$$\textit{with} \quad \kappa_{q,r} \triangleq \frac{2(1+r)}{1-q} = \mathcal{O}\left(\frac{1}{1-q}\right), \quad \textit{since } 0 \leq r \leq 1,\ 0 \leq q < 1$$

where the asymptotic notation in the upper bound is particularly useful in the regime $q \to 1$. This floored variant is used for the error bound; Algorithm 1 instead stops when the population becomes too small.

Proof. Recall that the exact squared Mahalanobis error, conditioned strictly on a known population size N_t, isolates the population in the denominator:

$$\mathbb{E}\left[\|o_t - \mu_t\|_{\Sigma^{-1}}^2 \;\middle|\; N_t\right] = \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_t}.$$

Taking the conditional expectation given N_{t-1} gives

$$\mathbb{E}\left[\|o_t - \mu_t\|_{\Sigma^{-1}}^2 \;\middle|\; N_{t-1}\right] = \Big(\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d\Big)\mathbb{E}\left[\frac{1}{N_t} \;\middle|\; N_{t-1}\right].$$

Lower Bound. The floor increases the raw population by at most one, so $\mathbb{E}[N_t \mid N_{t-1}] \leq C_{t-1}(\gamma)N_{t-1} + 1$. We apply Jensen's inequality to the convex function $x \mapsto 1/x$ on $(0, \infty)$:

$$\mathbb{E}\left[\frac{1}{N_t} \;\middle|\; N_{t-1}\right] \geq \frac{1}{\mathbb{E}[N_t \mid N_{t-1}]} \geq \frac{1}{C_{t-1}(\gamma)N_{t-1} + 1}.$$

Upper Bound. Pre-draw an independent departure coin for each participant. The number S whose coin prevents departure is distributed as $\mathrm{Bin}(N_{t-1}, 1-q)$. These participants remain regardless of leakage, so $N_t \geq \max\{1, S\}$. Since $1/\max\{1, s\} \leq 2/(s+1)$ for integers $s \geq 0$,

$$\mathbb{E}\left[\frac{1}{N_t}\middle| N_{t-1}\right] \leq 2\mathbb{E}\left[\frac{1}{S+1}\middle| N_{t-1}\right] = \frac{2(1-q^{N_{t-1}+1})}{(N_{t-1}+1)(1-q)}.$$

Using $C_{t-1}(\gamma) \leq 1+r$ directly yields:

$$\mathbb{E}\left[\frac{1}{N_t}\middle| N_{t-1}\right] \leq \frac{2(1+r)}{1-q}\frac{1}{C_{t-1}(\gamma)N_{t-1}} = \frac{\kappa_{q,r}}{C_{t-1}(\gamma)N_{t-1}}.$$

Multiplying the lower and upper bounds on the expected inverse population by $\left(\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d\right)$ establishes equation 23. □

Corollary 1 (Zero-Noise Collapse)**.** *For any purge rate $0 < q < 1$, as the algorithmic noise vanishes ($\gamma \to 0^+$), the expected estimation error increases geometrically while the population remains large, under the mean field approximation of N_t.*

Proof. As $\gamma \to 0^+$, the log-likelihood ratio diverges for any z with $\|z - \mu_{z,t}\| > 0$, driving the marginal leakage probability $p_\tau^{(t)} \to 1$ for all t. The demographic growth factor reduces to $C_t(\gamma) \to 1-q$ for all t, resulting in a mean-field population of $N_0(1-q)^{t-1}$. Using the single-round error formula, the expected estimation error is of order $\frac{d}{N_0(1-q)^{t-1}}$. For $q > 0$, the denominator decreases geometrically, causing the error to increase while the population remains large. This approximation does not describe behavior after the population reaches a floor or stopping rule. □

Lemma 11. *For nondegenerate Gaussian data, for fixed $\gamma > 0$, under mean-field approximation, if $C_t(\gamma) > 1$ for all t (i.e., the expected population strictly grows), then $C_t(\gamma)$ is a strictly increasing sequence.*

Proof. Under the mean-field approximation of N_t, the expected population size at step t is given by $N_{t+1} = C_t(\gamma)N_t$. Since we are given that $C_t(\gamma) > 1$ for all t, the population strictly grows, meaning $N_{t+1} > N_t$ (and consequently its inverse shrinks: $1/N_{t+1} < 1/N_t$).

To justify the decrease in leakage, note that for a Gaussian sample of size N, $z_i - \mu_z \sim \mathcal{N}(0, (1 - 1/N)\Sigma)$. Writing $G \sim \mathcal{N}(0, \Sigma)$, the standardized shift in Lemma 2 has distribution

$$\frac{\sqrt{N}\|z_i - \mu_z\|}{\gamma(N-1)} \overset{d}{=} \frac{\|G\|}{\gamma\sqrt{N-1}}.$$

For each nonzero G, this decreases with N, and so does its leakage probability $1 - \Phi(\tau\gamma\sqrt{N-1}/\|G\| - \|G\|/(2\gamma\sqrt{N-1}))$. Taking expectations preserves the decrease. Thus, $p_\tau^{(t)} > p_\tau^{(t+1)}$.

To determine the behavior of $C_t(\gamma)$, we evaluate the difference between consecutive terms:

$$C_{t+1}(\gamma) - C_t(\gamma) = \left[1 + r - (q+r)p_\tau^{(t+1)}\right] - \left[1 + r - (q+r)p_\tau^{(t)}\right] = (q+r)\left(p_\tau^{(t)} - p_\tau^{(t+1)}\right)$$

Since $q + r > 0$ and $p_\tau^{(t)} - p_\tau^{(t+1)} > 0$, the product is positive $C_{t+1}(\gamma) - C_t(\gamma) > 0$, concluding the proof. □

Corollary 2 (Infinite Horizon Asymptotics)**.** *For any positive recruitment rate $r > 0$ and any fixed finite noise scale $\gamma > 0$ such that $C_t(\gamma) > 1$ for all t, the expected estimation error at round t converges to zero as $t \to \infty$, at the geometric rate $\mathcal{O}(C_1(\gamma)^{-(t-1)})$ under the mean field approximation of N_t.*

Proof. For a fixed γ with $C_t(\gamma) > 1$ for all t, using the single-round error formula under the mean-field approximation of N_t, $\mathcal{E}_t(\gamma)$ is at most $\frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_0 \prod_{k=1}^{t-1} C_k(\gamma)}$. Due to lemma 11, this quantity is upper

bounded by $\frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1})+d}{N_0 C_1(\gamma)^{t-1}}$ and decays to zero as $t \to \infty$. As $\gamma \to \infty$, the marginal leakage probability vanishes ($p_\tau \to 0$), maximizing the growth factor to $C(\gamma) \to 1 + r$ and the mean-field population toward $N_0(1+r)^{t-1}$: for any fixed finite noise level in the growth regime, exponential population expansion eventually dominates the fixed-in-t mechanism variance term γ^2.

This conclusion continues to hold when q is large (even as $q \to 1$) and $r > 0$ is small but nonzero, provided $C(\gamma) > 1$. The numerator $\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d$ remains fixed in t, while $C_1(\gamma)^{t-1}$ grows exponentially, so the expected estimation error still converges to zero. □

Lemma 12. *For a finite horizon T, work under the deterministic mean-field approximation. Assume $C_1(\gamma) > 1$. The cumulative expected Mahalanobis error aggregated across all T rounds, defined as $\mathcal{E}_{total}(\gamma) \triangleq \sum_{t=1}^{T} \mathcal{E}_t(\gamma)$ is upper bounded by $\mathcal{E}_{total}(\gamma) \leq \hat{\mathcal{E}}_{total}(\gamma)$, where:*

$$\hat{\mathcal{E}}_{total}(\gamma) \triangleq \frac{C_1(\gamma)(\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d)}{N_0(C_1(\gamma) - 1)} \left(1 - \frac{1}{[C_1(\gamma)]^T}\right) \tag{24}$$

Proof. To evaluate the true cost of privacy over a finite horizon T, we must aggregate the expected error across all rounds. The cumulative expected Mahalanobis error is the sum of the single-round errors:

$$\mathcal{E}_{total}(\gamma) = \sum_{t=1}^{T} \mathbb{E}\left[\|o_t - \mu_t\|^2_{\Sigma^{-1}}\right] = \sum_{t=1}^{T} \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_0 \prod_{k=1}^{t-1} C_k(\gamma)} \overset{(a)}{\leq} \sum_{t=1}^{T} \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_0 [C_1(\gamma)]^{t-1}}$$

Where (a) holds due to lemma 11. We apply the formula for sum of a finite geometric series:

$$\begin{aligned}
\hat{\mathcal{E}}_{total}(\gamma) &\triangleq \sum_{t=1}^{T} \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_0 [C_1(\gamma)]^{t-1}} \\
&= \frac{\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d}{N_0} \left(\frac{1 - \frac{1}{[C_1(\gamma)]^T}}{1 - \frac{1}{C_1(\gamma)}}\right) \\
&= \frac{C_1(\gamma)(\gamma^2 \mathrm{Tr}(\Sigma^{-1}) + d)}{N_0(C_1(\gamma) - 1)} \left(1 - \frac{1}{[C_1(\gamma)]^T}\right)
\end{aligned}$$

This completes the proof. □

Remark 5 (Optimization in the Growth Regime)**.** *In the growth regime, where $C_1(\gamma) > 1$, under the mean-field approximation, directly optimizing the upper bound $\hat{\mathcal{E}}_{total}(\gamma)$ rather than the exact cumulative error $\mathcal{E}_{total}(\gamma)$ is a practical surrogate objective. It captures the noise–growth trade-off, without requiring its minimizer to coincide with that of the exact cumulative error. Computationally, this substitution is advantageous as it depends solely on the initial expected growth factor $C_1(\gamma)$, bypassing the need to recursively track the sequence $C_t(\gamma)$ across all T rounds. Furthermore, empirical evaluations confirm that optimizing this upper bound yields an effective configuration in practice.*

F.6 High-Probability Population Lower Bound

Lemma 13. *Let Z_t be the dataset realized at round t. Let $\kappa_{t+1} \triangleq \mathbb{E}[N_{t+1} \mid N_t]$ denote the expected population at round $t+1$, integrating over both the mechanism's privacy noise and the individuals' stochastic performative decisions. For any relative deviation $\epsilon \in (0, 1)$, the probabilities that the realized population N_{t+1} deviates below or above this conditional expectation decay exponentially:*

(a) $\Pr\left(N_{t+1} \leq (1-\epsilon)\kappa_{t+1} \mid Z_t\right) \leq \exp\left(-\frac{\epsilon^2 C_t(\gamma)\kappa_{t+1}}{8}\right)$

(b) $\Pr\left(N_{t+1} \geq (1+\epsilon)\kappa_{t+1} \mid Z_t\right) \leq \exp\left(-\frac{\epsilon^2 C_t(\gamma)\kappa_{t+1}}{8}\right)$

where $C_t(\gamma) \triangleq \frac{\kappa_{t+1}}{N_t} = 1 + r - (q+r)p_\tau^{(t)}(\gamma)$ is the expected performative growth factor evaluated at round t.

Proof. Given the dataset Z_t, the total population at round $t+1$ is the sum of the individual demographic updates: $N_{t+1} = \sum_{i=1}^{N_t} Y_{i,t}$, where each $Y_{i,t} \in [0, 2]$ represents the net presence of individual i (e.g., 0 for leaving, 1 for staying, 2 for bringing a joiner).

While the shared privacy noise η_t induces a weak dependency among the final decisions $\{Y_{i,t}\}$, we can rigorously bound the aggregate deviation without conditioning on the noise by constructing a Doob martingale. Let us order the individuals arbitrarily and define the sequence of filtrations $\mathcal{F}_k = \sigma(Z_t, Y_{1,t}, \ldots, Y_{k,t})$ for $k \in \{0, \ldots, N_t\}$. The corresponding martingale is:

$$M_k = \mathbb{E}[N_{t+1} \mid \mathcal{F}_k]$$

with the boundary values $M_0 = \kappa_{t+1}$ and $M_{N_t} = N_{t+1}$.

The martingale difference $\Delta_k = M_k - M_{k-1}$ represents the change in the expected total population upon revealing the k-th individual's decision. Because an individual's decision is strictly bounded in the interval $[0, 2]$, and the posterior shift on the shared noise distribution induced by revealing a single Bernoulli observation in a large population is negligible, the martingale differences are bounded almost surely by $|M_k - M_{k-1}| \leq 2$.

Applying the Azuma-Hoeffding inequality to this martingale difference sequence yields the upper tail bound:

$$\Pr\left(N_{t+1} - \kappa_{t+1} \geq \epsilon\kappa_{t+1} \mid Z_t\right) = \Pr(M_{N_t} - M_0 \geq \epsilon\kappa_{t+1}) \leq \exp\left(-\frac{(\epsilon\kappa_{t+1})^2}{2\sum_{k=1}^{N_t} 2^2}\right)$$

Simplifying the denominator:

$$2\sum_{k=1}^{N_t} 4 = 8N_t$$

Substituting the deterministic relationship $N_t = \kappa_{t+1}/C_t(\gamma)$ yields:

$$\Pr\left(N_{t+1} \geq (1+\epsilon)\kappa_{t+1} \mid Z_t\right) \leq \exp\left(-\frac{\epsilon^2\kappa_{t+1}^2}{8\kappa_{t+1}/C_t(\gamma)}\right) = \exp\left(-\frac{\epsilon^2 C_t(\gamma)\kappa_{t+1}}{8}\right)$$

By the symmetry of the Azuma-Hoeffding bound, an identical exponential decay holds for the lower tail:

$$\Pr\left(N_{t+1} \leq (1-\epsilon)\kappa_{t+1} \mid Z_t\right) \leq \exp\left(-\frac{\epsilon^2 C_t(\gamma)\kappa_{t+1}}{8}\right)$$

This establishes the two-sided exponential decay for the population trajectory given the current round's dataset. □

G Theoretical Results from Section 5

Lemma 14. *Let $Z_t = (X, y)$ be a dataset containing N_t records. Let $A = X^\top X + \lambda I$ and $\theta = A^{-1}X^\top y$ denote the exact ridge regression estimator on Z_t. Let the differentially private mechanism output $\widetilde{\theta} = \theta + \eta$, where $\eta \sim \mathcal{N}(0, \frac{\gamma^2}{N_t} I_{d\times d})$. For a target record $(x_i, y_i) \in Z_t$, let $e_i = y_i - x_i^\top\theta$ denote its residual and $h_i = x_i^\top A^{-1} x_i$ its leverage score. The log-likelihood ratio $\Lambda(\widetilde{\theta} \mid x_i, y_i)$ conditioned on Z_t is given by:*

$$\Lambda(\widetilde{\theta} \mid x_i, y_i) \mid Z_t = \frac{N_t e_i^2 x_i^\top A^{-2} x_i}{2\gamma^2(1-h_i)^2} + \frac{N_t e_i \langle A^{-1}x_i, \eta\rangle}{\gamma^2(1-h_i)}$$

And it is distributed as:

$$\Lambda(\widetilde{\theta} \mid x_i, y_i) \mid Z_t \sim \mathcal{N}\left(\frac{N_t e_i^2 x_i^\top A^{-2} x_i}{2\gamma^2(1-h_i)^2},\ \frac{N_t e_i^2 x_i^\top A^{-2} x_i}{\gamma^2(1-h_i)^2}\right)$$

Proof. By definition, the log-likelihood ratio compares the probability density of observing the noisy output $\widetilde{\theta}$ when (x_i, y_i) is present in the training set versus when it is absent. Let θ_{-i} denote the ridge

estimator computed on $Z_t \setminus \{(x_i, y_i)\}$ with penalty λ and noise covariance held fixed. Because the additive noise η is distributed as $\mathcal{N}(0, \frac{\gamma^2}{N_t} I)$, the log-likelihood ratio conditioned on Z_t evaluates to:

$$\begin{aligned}\Lambda(\widetilde{\theta} \mid x_i, y_i) \mid Z_t &= \ln \left(\frac{\exp\left(-\frac{N_t}{2\gamma^2}\|\widetilde{\theta} - \theta\|^2\right)}{\exp\left(-\frac{N_t}{2\gamma^2}\|\widetilde{\theta} - \theta_{-i}\|^2\right)} \right) \\ &= \frac{N_t}{2\gamma^2}\left(\|\widetilde{\theta} - \theta_{-i}\|^2 - \|\widetilde{\theta} - \theta\|^2\right)\end{aligned}$$

By the Sherman-Morrison formula, the shift in the estimator due to the removal of (x_i, y_i) is given by $\Delta\theta = \theta - \theta_{-i} = \frac{e_i}{1-h_i} A^{-1} x_i$, which implies $\theta_{-i} = \theta - \Delta\theta$. Substituting the noisy observation $\widetilde{\theta} = \theta + \eta$ and the expression for θ_{-i}:

$$\begin{aligned}\Lambda(\widetilde{\theta} \mid x_i, y_i) \mid Z_t &= \frac{N_t}{2\gamma^2}\left(\|\theta + \eta - (\theta - \Delta\theta)\|^2 - \|\theta + \eta - \theta\|^2\right) \\ &= \frac{N_t}{2\gamma^2}\left(\|\Delta\theta + \eta\|^2 - \|\eta\|^2\right) \\ &= \frac{N_t}{2\gamma^2}\left(\|\Delta\theta\|^2 + 2\langle\Delta\theta, \eta\rangle + \|\eta\|^2 - \|\eta\|^2\right) \\ &= \frac{N_t\|\Delta\theta\|^2}{2\gamma^2} + \frac{N_t\langle\Delta\theta, \eta\rangle}{\gamma^2}\end{aligned}$$

Substituting $\Delta\theta = \frac{e_i}{1-h_i} A^{-1} x_i$, the squared Euclidean norm evaluates to $\|\Delta\theta\|^2 = \left(\frac{e_i}{1-h_i}\right)^2 x_i^\top A^{-2} x_i$. This yields:

$$\Lambda(\widetilde{\theta} \mid x_i, y_i) \mid Z_t = \frac{N_t e_i^2 x_i^\top A^{-2} x_i}{2\gamma^2(1-h_i)^2} + \frac{N_t e_i \langle A^{-1} x_i, \eta\rangle}{\gamma^2(1-h_i)}$$

Conditioned on Z_t, the term $\frac{N_t\|\Delta\theta\|^2}{2\gamma^2}$ is a deterministic scalar. The variance is derived strictly from the inner product $\langle\Delta\theta, \eta\rangle$, which is a linear combination of the elements of $\eta \sim \mathcal{N}(0, \frac{\gamma^2}{N_t} I)$. Consequently, $\langle\Delta\theta, \eta\rangle$ is a univariate Gaussian with mean 0 and variance $\frac{\gamma^2}{N_t}\|\Delta\theta\|^2$. Scaling this random variable by $\frac{N_t}{\gamma^2}$, the conditional variance reduces to:

$$\mathrm{Var}\left(\frac{N_t\langle\Delta\theta, \eta\rangle}{\gamma^2} \;\middle|\; Z_t\right) = \frac{N_t^2}{\gamma^4}\left(\frac{\gamma^2}{N_t}\|\Delta\theta\|^2\right) = \frac{N_t\|\Delta\theta\|^2}{\gamma^2} = \frac{N_t e_i^2 x_i^\top A^{-2} x_i}{\gamma^2(1-h_i)^2}$$

Thus, the log-likelihood ratio follows the distribution:

$$\Lambda(\widetilde{\theta} \mid x_i, y_i) \mid Z_t \sim \mathcal{N}\left(\frac{N_t e_i^2 x_i^\top A^{-2} x_i}{2\gamma^2(1-h_i)^2},\ \frac{N_t e_i^2 x_i^\top A^{-2} x_i}{\gamma^2(1-h_i)^2}\right)$$

This completes the proof. □

Lemma 15. *Given the differentially private ridge regression estimator, the exact probability that a specific record (x_i, y_i) is exposed (i.e., its log-likelihood ratio breaches the threshold τ) conditioned on Z_t is given in closed form by:*

$$\Pr(\Lambda(\widetilde{\theta} \mid x_i, y_i) > \tau \mid Z_t) = 1 - \Phi\left(\frac{\tau\gamma(1-h_i)}{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}} - \frac{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}}{2\gamma(1-h_i)}\right)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

Proof. Lemma 14 aids us in establishing the log-likelihood ratio for a specific record as a Gaussian distribution:

$$\Lambda(\widetilde{\theta} \mid x_i, y_i) \mid Z_t \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

where the distribution parameters are $\mu_i = \frac{N_t e_i^2 x_i^\top A^{-2} x_i}{2\gamma^2(1-h_i)^2}$ and $\sigma_i = \frac{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}}{\gamma(1-h_i)}$.

Evaluating the tail probability $\Pr(\Lambda > \tau \mid Z_t)$ via the standard normal cumulative distribution function $\Phi\left(\frac{\tau-\mu_i}{\sigma_i}\right)$ and executing the final algebraic transformation yields:

$$\Pr(\Lambda(\widetilde{\theta} \mid x_i, y_i) > \tau \mid Z_t) = 1 - \Phi\left(\frac{\tau\gamma(1-h_i)}{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}} - \frac{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}}{2\gamma(1-h_i)}\right)$$

This completes the proof. □

Lemma 16. *Given the performative update rule in the differentially private mechanism, the expected population size at round $t+1$, conditioned on the current population N_t, is given by:*

$$\mathbb{E}[N_{t+1} \mid N_t] = N_t\Big(1 + r - (q+r)p_\tau^{(t)}\Big)$$

Proof. Let $\mathbb{1}\big(\Lambda(\widetilde{\theta}_t \mid x_i, y_i) > \tau\big)$ be the indicator variable denoting whether the i-th data point experiences a privacy leakage exceeding the threshold τ at round t. Averaging the independent departure and recruitment coins, conditional on the dataset and released output, gives:

$$\mathbb{E}[N_{t+1} \mid Z_t, \widetilde{\theta}_t] = N_t - q\sum_{i=1}^{N_t} \mathbb{1}\big(\Lambda(\widetilde{\theta}_t \mid x_i, y_i) > \tau\big) + r\left(N_t - \sum_{i=1}^{N_t} \mathbb{1}\big(\Lambda(\widetilde{\theta}_t \mid x_i, y_i) > \tau\big)\right)$$

Where $(x_i, y_i) \in Z_t$. Grouping the terms yields:

$$\mathbb{E}[N_{t+1} \mid Z_t, \widetilde{\theta}_t] = N_t(1+r) - (q+r)\sum_{i=1}^{N_t} \mathbb{1}\big(\Lambda(\widetilde{\theta}_t \mid x_i, y_i) > \tau\big)$$

Taking the conditional expectation given N_t, we evaluate the joint expectation over the entire dataset $Z_t = \{(x_1, y_1), \ldots, (x_{N_t}, y_{N_t})\}$ and the mechanism's noise η_t:

$$\mathbb{E}[N_{t+1} \mid N_t] = N_t(1+r) - (q+r)\mathbb{E}_{Z_t,\eta_t}\left[\sum_{i=1}^{N_t} \mathbb{1}\big(\Lambda(\widetilde{\theta}_t \mid x_i, y_i) > \tau\big)\right]$$

By Fubini's theorem, we distribute the expectation inside the sum. The log-likelihood ratio for a specific point depends on (x_i, y_i), the dataset Z_t, and the noise η_t. Utilizing the tower property of expectation:

$$\mathbb{E}_{Z_t,\eta_t}\left[\sum_{i=1}^{N_t} \mathbb{1}\big(\Lambda(\widetilde{\theta}_t \mid x_i, y_i) > \tau\big)\right] = \sum_{i=1}^{N_t} \mathbb{E}_{Z_t}\left[\Pr\big(\Lambda(\widetilde{\theta}_t \mid x_i, y_i) > \tau \mid Z_t\big)\right]$$

From lemma 15, the conditional probability of leakage for any specific data point is given as:

$$\Pr(\Lambda(\widetilde{\theta}_t \mid x_i, y_i) > \tau \mid Z_t) = 1 - \Phi\left(\frac{\tau\gamma(1-h_i)}{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}} - \frac{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}}{2\gamma(1-h_i)}\right)$$

Since all N_t data points are identically distributed, we proceed as follows:

$$\mathbb{E}[N_{t+1} \mid N_t] = N_t(1+r) - (q+r)N_t\mathbb{E}_{Z_t}\left[1 - \Phi\left(\frac{\tau\gamma(1-h_i)}{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}} - \frac{\sqrt{N_t}|e_i|\sqrt{x_i^\top A^{-2} x_i}}{2\gamma(1-h_i)}\right)\right]$$

The ridge regression mechanism enforces strict public bounds on the dataset before evaluating the sufficient statistics. Specifically, the inputs are clipped such that $\|\bar{x}_i\|_2 \le R$ and $|\bar{y}_i| \le Y$. Let $\bar{A} = \sum_{j=1}^{N_t} \bar{x}_j \bar{x}_j^\top + \lambda I$ denote the regularized design matrix on the clipped features, with clipped leverage scores $\bar{h}_i = \bar{x}_i^\top \bar{A}^{-1} \bar{x}_i$ and clipped residuals $\bar{e}_i = \bar{y}_i - \bar{x}_i^\top \bar{\theta}$. Thus, the marginal leakage probability $p_\tau^{(t)}$ over the clipped data distribution is:

$$p_\tau^{(t)} = \mathbb{E}_{Z_t}\left[1 - \Phi\left(\frac{\tau\gamma(1-\bar{h}_i)}{\sqrt{N_t}|\bar{e}_i|\sqrt{\bar{x}_i^\top \bar{A}^{-2} \bar{x}_i}} - \frac{\sqrt{N_t}|\bar{e}_i|\sqrt{\bar{x}_i^\top \bar{A}^{-2} \bar{x}_i}}{2\gamma(1-\bar{h}_i)}\right)\right]$$

Substituting $p_\tau^{(t)}$ into the population equation yields:

$$\begin{aligned}\mathbb{E}[N_{t+1} \mid N_t] &= N_t(1+r) - (q+r)N_t p_\tau^{(t)} \\ &= N_t\Big(1 + r - (q+r)p_\tau^{(t)}\Big)\end{aligned}$$

This completes the proof. □

Theorem 3. *For $q, r > 0$ and $\lambda > 0$, the condition $C_t(\gamma) \geq 1$ is guaranteed as soon as the noise scale of the ridge mechanism equation 14 satisfies $\gamma \geq \gamma_c$, where*

$$\gamma_c \triangleq \sqrt{N_t}\,\Delta_{\max}\left(\frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau}\right), \qquad \Delta_{\max} \triangleq \frac{RY}{\lambda}\left(1 + R\sqrt{\frac{N_t}{\lambda}}\right),$$

$\Phi_{q,r}^{-1} \triangleq \Phi^{-1}\left(\frac{q}{q+r}\right)$ as in Theorem 2, and $\Delta_{\max}$ bounds the shift $\|\Delta\theta_i\|_2$ induced by any single record.

Proof. Nonnegative expected growth requires $C_t(\gamma) \geq 1$. Given the definition of the performative growth factor, this restricts the maximum permissible expected marginal leakage probability p_τ:

$$C_t(\gamma) \geq 1 \implies 1 + r - (q+r)p_\tau \geq 1 \implies p_\tau \leq \frac{r}{q+r}.$$

Since the leave-one-out ridge objective at its minimizer is no larger than its value at zero, $\lambda\|\theta_{-i}\|^2 \leq N_t Y^2$. The normal equations give

$$\Delta\theta_i = A^{-1}x_i(y_i - x_i^\top\theta_{-i}).$$

Using $\|A^{-1}\| \leq 1/\lambda$, $\|x_i\| \leq R$ and $|y_i| \leq Y$, we obtain

$$\|\Delta\theta_i\| \leq \frac{RY}{\lambda}\left(1 + R\sqrt{\frac{N_t}{\lambda}}\right) = \Delta_{\max}.$$

It is therefore sufficient to require:

$$1 - \Phi\left(\frac{\tau\gamma}{\sqrt{N_t}\Delta_{\max}} - \frac{\sqrt{N_t}\Delta_{\max}}{2\gamma}\right) \leq \frac{r}{q+r}.$$

Rearranging to isolate the CDF and applying the Probit function (Φ^{-1}) yields:

$$\frac{\tau\gamma}{\sqrt{N_t}\Delta_{\max}} - \frac{\sqrt{N_t}\Delta_{\max}}{2\gamma} \geq \Phi^{-1}\left(1 - \frac{r}{q+r}\right) = \Phi^{-1}\left(\frac{q}{q+r}\right) \triangleq \Phi_{q,r}^{-1}. \tag{25}$$

Let

$$u = \frac{\gamma}{\sqrt{N_t}\Delta_{\max}}.$$

Substituting u transforms equation 25 into a standard quadratic equation:

$$\tau u - \frac{1}{2u} \geq \Phi_{q,r}^{-1} \implies 2\tau u^2 - 2\Phi_{q,r}^{-1}u - 1 \geq 0.$$

Taking the valid positive root of the quadratic provides the strict lower bound for u:

$$u \geq \frac{2\Phi_{q,r}^{-1} + \sqrt{4(\Phi_{q,r}^{-1})^2 + 8\tau}}{4\tau} = \frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau}. \tag{26}$$

Substituting $u = \frac{\gamma}{\sqrt{N_t}\Delta_{\max}}$ back into equation 26 yields:

$$\gamma \geq \sqrt{N_t}\Delta_{\max}\left(\frac{\Phi_{q,r}^{-1} + \sqrt{(\Phi_{q,r}^{-1})^2 + 2\tau}}{2\tau}\right) \triangleq \gamma_c.$$

This worst-case bound is conservative. For well-conditioned data and sufficiently large N_t, typical leave-one-out shifts instead decrease with the population size. □

Lemma 17. *For a finite horizon T, under the deterministic mean-field approximation with $N_1 = N_0$, let's assume $C_t(\gamma) \geq C_1(\gamma) > 1$. Assume that the eigenvalues of Σ are bounded above and away from zero, and that the fixed penalty $\lambda > 0$, $\|\theta_t^*\|$ and observation-noise variance are bounded uniformly over t. Let $\mathcal{L}(\theta) = \mathbb{E}_{(x,y)}[(y - x^\top\theta)^2]$ denote the expected test loss. The cumulative expected excess test loss aggregated across all T rounds, defined as $\mathcal{E}_{total}(\gamma) \triangleq \sum_{t=1}^{T} \mathcal{E}_t(\gamma)$ where $\mathcal{E}_t(\gamma) \triangleq \mathbb{E}_{Z_t,\eta_t}\left[\mathcal{L}(\widetilde{\theta}_t) - \mathcal{L}(\theta_t^*)\right]$, is upper bounded by $\mathcal{E}_{total}(\gamma) \leq \hat{\mathcal{E}}_{total}(\gamma)$, where:*

$$\hat{\mathcal{E}}_{total}(\gamma) \triangleq \frac{C_1(\gamma)(\gamma^2 \mathrm{Tr}(\Sigma) + \nu)}{N_0(C_1(\gamma) - 1)} \left(1 - \frac{1}{[C_1(\gamma)]^T}\right) \tag{27}$$

where $\nu = \mathcal{O}(d)$ accounts for non-private estimation error, including regularization.

Proof. To evaluate the true cost of privacy on model utility over a finite horizon T, we consider the expected excess test loss (excess risk) at round t. Let θ_t^* denote the actual underlying population weights at round t. For the squared loss objective in linear regression, the excess risk reduces exactly to the expected squared Σ distance under the feature covariance matrix Σ:

$$\mathcal{E}_t(\gamma) = \mathbb{E}_{Z_t,\eta_t}\left[\mathcal{L}(\widetilde{\theta}_t) - \mathcal{L}(\theta_t^*)\right] = \mathbb{E}_{Z_t,\eta_t}\left[\|\widetilde{\theta}_t - \theta_t^*\|_\Sigma^2\right]$$

Substituting the output-perturbed estimator $\widetilde{\theta}_t = \theta_t + \eta_t$, where θ_t denotes the exact non-private ridge estimator and $\eta_t \sim \mathcal{N}(0, \frac{\gamma^2}{N_t} I_{d\times d})$:

$$\begin{aligned}\mathcal{E}_t(\gamma) &= \mathbb{E}_{Z_t,\eta_t}\left[\|\theta_t - \theta_t^* + \eta_t\|_\Sigma^2\right] \\ &= \mathbb{E}_{Z_t}\left[\|\theta_t - \theta_t^*\|_\Sigma^2\right] + \mathbb{E}_{\eta_t}\left[\|\eta_t\|_\Sigma^2\right] + 2\mathbb{E}_{Z_t,\eta_t}\left[\langle \theta_t - \theta_t^*, \Sigma\eta_t\rangle\right]\end{aligned}$$

Because η_t has zero mean and is statistically independent of Z_t, the expectation of the inner product vanishes. The expected excess loss induced by the privacy noise evaluates via the trace operator:

$$\mathbb{E}_{\eta_t}\left[\|\eta_t\|_\Sigma^2\right] = \mathbb{E}_{\eta_t}\left[\eta_t^\top \Sigma \eta_t\right] = \mathrm{Tr}\left(\Sigma\, \mathbb{E}[\eta_t\eta_t^\top]\right) = \mathrm{Tr}\left(\Sigma\left(\frac{\gamma^2}{N_t} I_{d\times d}\right)\right) = \frac{\gamma^2 \mathrm{Tr}(\Sigma)}{N_t}$$

For Gaussian design and $N_t > d + 1$,

$$\mathbb{E}[(X_t^\top X_t)^{-1}] = \frac{\Sigma^{-1}}{N_t - d - 1}.$$

Conditional on X_t, the ridge bias is $-\lambda A_t^{-1}\theta_t^*$ and its variance contribution is $\sigma^2 \mathrm{Tr}(\Sigma A_t^{-1} X_t^\top X_t A_t^{-1})$. Using

$$A_t^{-1} X_t^\top X_t A_t^{-1} \preceq (X_t^\top X_t)^{-1}, \qquad \lambda^2 A_t^{-2} \preceq \frac{\lambda}{4}(X_t^\top X_t)^{-1},$$

we obtain

$$\mathbb{E}\|\theta_t - \theta_t^*\|_\Sigma^2 \leq \frac{\sigma^2 d + \frac{\lambda\|\Sigma\|}{4}(\theta_t^*)^\top \Sigma^{-1}\theta_t^*}{N_t - d - 1} \leq \frac{\nu}{N_t},$$

for a uniform $\nu = \mathcal{O}(d)$, since $N_t \geq N_0 \geq 2(d+1)$. Consequently,

$$\mathcal{E}_t(\gamma) \leq \frac{\gamma^2 \mathrm{Tr}(\Sigma) + \nu}{N_t}.$$

Aggregating across all T rounds and substituting the demographic evolution under mean field approximation $N_t = N_0 \prod_{k=1}^{t-1} C_k(\gamma)$:

$$\mathcal{E}_{\text{total}}(\gamma) \leq \sum_{t=1}^{T} \frac{\gamma^2 \mathrm{Tr}(\Sigma) + \nu}{N_0 \prod_{k=1}^{t-1} C_k(\gamma)} \overset{(a)}{\leq} \sum_{t=1}^{T} \frac{\gamma^2 \mathrm{Tr}(\Sigma) + \nu}{N_0 [C_1(\gamma)]^{t-1}}$$

where (a) uses the assumed nondecreasing growth factors; this is an additional assumption for regression. Summing the finite geometric series yields:

$$\begin{aligned}\hat{\mathcal{E}}_{\text{total}}(\gamma) &\triangleq \sum_{t=1}^{T} \frac{\gamma^2 \text{Tr}(\Sigma) + \nu}{N_0 [C_1(\gamma)]^{t-1}} \\ &= \frac{\gamma^2 \text{Tr}(\Sigma) + \nu}{N_0} \left(\frac{1 - \frac{1}{[C_1(\gamma)]^T}}{1 - \frac{1}{C_1(\gamma)}} \right) \\ &= \frac{C_1(\gamma)(\gamma^2 \text{Tr}(\Sigma) + \nu)}{N_0 (C_1(\gamma) - 1)} \left(1 - \frac{1}{[C_1(\gamma)]^T} \right)\end{aligned}$$

This completes the proof. □