BALL DIFFERENTIAL PRIVACY - HOW TO MITIGATE DATA RECONSTRUCTION WITH LESS NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vector embeddings of raw records, while not human-readable, do not preserve privacy of records: an adversary can reconstruct training records from a released model. Differential privacy (DP) is the principled defense, but its noise is calibrated to the worst-case indistinguishability; obscuring arbitrary single-record substitutions, including those far outside the set of plausible alternatives relevant to a reconstruction adversary. This results in noise levels far larger than what reconstruction robustness requires, degrading accuracy without a corresponding security benefit against reconstruction attacks. To mitigate this limitation of standard DP, we propose *Ball-DP*: (ε, δ) indistinguishability over single-record substitutions restricted to a ball of radius r according to a distance metric d in the embedding space. A deployment facing only local reconstruction threat can choose a small r, thereby reduce noise and recover wasted accuracy. The radius makes the scope of the privacy claim explicit against reconstruction attacks; standard DP is recovered when r covers the entire admissible record domain. We provide noise calibrations for regularized convex learning problems for Ball-DP, and derive the corresponding reconstruction-robustness certificates (named *Ball-ReRo*) – upper-bound on an attacker's reconstruction success. By deriving the Bayes-optimal finite-prior MAP reconstruction attack, we conduct empirical auditing of Ball-ReRo certificates for several benchmark learning tasks. We observe that calibrating noise to Ball-DP significantly improves utility, with marginal impact on reconstruction robustness.

1 INTRODUCTION

Modern machine learning systems rarely operate directly on raw data. Instead, representation-learning pipelines commonly map text, images, or other types of user records to vector embeddings that can be reused by downstream task-specific classifying heads (Bengio et al., 2013; Devlin et al., 2019; Reimers & Gurevych, 2019; Radford et al., 2021). This representation layer creates a false sense of privacy. Embeddings may not be human-readable, yet a model trained on them can reveal information about the training records. In the informed-adversary model, for example, an attacker knowing all but one training record can identify or reconstruct training records from a released model, even if the release is a simple convex model rather than a memorizing neural network (Balle et al., 2022).

Privacy-preserving AI requires guarantees that remain meaningful even against strong adversaries with side information. Differential privacy (DP) provides such a guarantee by requiring the released model to look nearly the same under arbitrary changes to a single training record (Cummings et al., 2024a). The parameters (ε, δ) control the strength of this indistinguishability requirement, with smaller values giving stronger protection. A common implementation is output perturbation: train the model, then add enough noise to mask the effect of any single record (Chaudhuri et al., 2011). This worst-case calibration is simple and broadly applicable, but it can require substantial noise and therefore reduce accuracy when strong privacy is desired.

DP is naturally aligned with membership inference attacks (MIA), which ask whether a particular record was used to train the model: by making the released model nearly indistinguishable under any single-record changes, DP limits what an attacker can infer about involvement of specific record. However, participation is not always the sensitive fact of interest (Hayes et al., 2023; Cummings et al., 2024b). In many settings, it may already be known that a person contributed data; the real privacy concern is whether the released model reveals the private content of their record.

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.

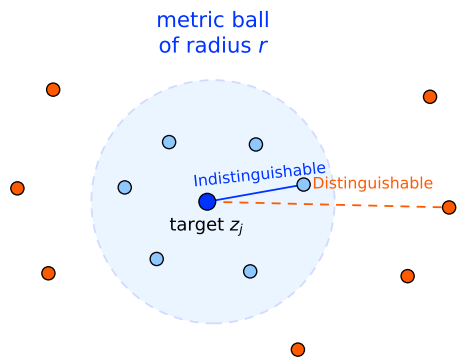

Figure 1: Illustration of reconstruction protection due to Ball-DP: an adversary can narrow down the target record to points highlighted in blue but cannot reliably tell any further. By limiting the scope of (ε, δ)-indistinguishability to replacements within a metric ball of radius r, Ball-DP mitigates accurate reconstruction while recovering higher model accuracy than standard DP.

This motivates training-record reconstruction attacks, where an adversary uses side information to recover the hidden record itself, often by choosing from a set of plausible candidates. The connection between DP guarantee and reconstruction risk is, however, quite loose (Hayes et al., 2023; Cummings et al., 2024b). DP mechanisms provide indistinguishability against worst-case single-record substitutions that can be arbitrarily far from each other in the embedding space. While a reconstruction adversary is interested in generating an estimate (i.e., choosing an element from the candidate set) that is in proximity to the target record. Thus, to prevent reconstruction, it suffices to induce indistinguishability over single-record substitutions that are close to each other in the underlying embedding space.

Ball-DP. We propose *Ball-DP*, which deploys the usual (ε, δ)-indistinguishability but applies it to only single-record substitutions within the ball of declared radius r. The radius becomes an explicit privacy parameter: smaller r efficiently mitigates local reconstruction threats using less noise, while larger r expands the protected set at a corresponding utility cost. For the regularized convex prediction heads that we consider, we can ensure (r, ε, δ)-Ball-DP via Gaussian output perturbation whose noise scales with r in addition to the indistinguishability parameters (ε, δ). Figure 1 illustrates the protected adjacency scope. Reconstruction guarantees depend on the adversary's prior and feasible-output baseline, not on r alone. When every admissible replacement has distance at most r, Ball-DP recovers the corresponding global replacement guarantee.

Contributions. Our contributions can be summarized as follows:

1. We define (r, ε, δ)-Ball-DP formally, i.e., (ε, δ)-indistinguishability over r-adjacency, and establish its connection to standard DP.
2. We derive ℓ_2-sensitivity of strongly convex empirical-risk minimization (ERM) under r-adjacency, and present the Gaussian release mechanism calibrated to Ball-DP guarantee.
3. We formally connect Ball-DP to reconstruction robustness; bounding an attacker's reconstruction success rate as a function of the oblivious prior and the privacy parameters (r, ε, δ).
4. We empirically validate our theoretical bounds, and compare the accuracy-reconstruction robustness trade-off of Ball-DP and standard DP across seven learning benchmarks.

Our empirical study suggests that Ball-DP provides comparable reconstruction robustness as standard DP while requiring less noise, alleviating the accuracy–reconstruction robustness trade-off. At $\varepsilon = 2$ and $\delta = 10^{-6}$, Ball-DP improves mean accuracy over the standard global-radius DP comparator across all seven embedding benchmarks at the reported local policy radii. For example, at the $q50$ label-preserving policy radii, Ball-DP improves mean accuracy by 0.57–36.47 percentage points. Even at the broader $q95$ local radii, Ball-DP remains more accurate than the global-radius comparator on every benchmark. Finite-prior maximum a posteriori (MAP) audits show comparable empirical exact-identification rates for candidate sets of size $m = 10$.

Comparisons with metric-DP. Unlike metric-DP or d_X-privacy, which impose distance-dependent privacy-loss bounds (Chatzikokolakis et al., 2013; Imola et al., 2022), Ball-DP uses the metric to define the binary relation $D \sim_r D'$ and enforces uniform (ε, δ) indistinguishability on that relation. Our definition specifies neither stronger protection for closer pairs nor a direct guarantee for substitutions beyond r. This separates adjacency policy from calibration: sensitivity-based mechanisms can be adapted using r-adjacency sensitivity, as we show for Gaussian output perturbation.

2 PRELIMINARIES: ERM, RECONSTRUCTION, AND GAUSSIAN RELEASES

Embedding ERM setup. Let raw examples live in an application-dependent space $\mathcal{X}_0$, e.g., photos, text and tabular records, and let labels lie in finite space $\mathcal{Y}$. A public encoder $\phi : \mathcal{X}_0 \to \mathbb{R}^d$ maps each raw example to an embedding. From this point onwards, a training record is $z_i = (x_i, y_i) = (\phi(a_i), y_i) \in \mathcal{Z} := \mathbb{R}^d \times \mathcal{Y}$, where $a_i \in \mathcal{X}_0$ is the corresponding raw example. The label space may either be binary, $\mathcal{Y} = \{0, 1\}$, or multiclass, $\mathcal{Y} = \{1, \ldots, K\}$. A dataset is an ordered tuple $D = (z_1, \ldots, z_n) \in \mathcal{Z}^n$. We consider ℓ_2-regularized empirical risk minimization (ERM) on the training records. For a parameter $\theta \in \mathbb{R}^p$, a loss function $\ell(\theta; z)$, and regularization parameter $\lambda > 0$, we define

$$F_D(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; z_i) + \frac{\lambda}{2} \|\theta\|_2^2, \quad \widehat{\theta}(D) := \arg\min_{\theta \in \mathbb{R}^p} F_D(\theta). \tag{1}$$

When $\ell(\cdot; z)$ is convex for every record z, the objective in (1) is λ-strongly convex, hence the minimizer is unique. This is a standard convex-learning formulation behind regularized linear and generalized-linear heads (Hastie et al., 2009; Shalev-Shwartz & Ben-David, 2014).

Reconstruction threat model. We consider an *informed reconstruction adversary*, following the setting in Balle et al. (2022); Hayes et al. (2023), that observes the released model and has substantial side information about the training set. In particular, all records except one target record z are treated as known; when needed below, we denote this ordered tuple of known records by $D^- \in \mathcal{Z}^{n-1}$. The adversary also knows the training algorithm, hyperparameters, release rule, and any public information used to form beliefs about the hidden record.

The relevant reconstruction task is therefore not an unconstrained search over the full record space. Side information typically narrows the target to a finite shortlist, or more generally a prior distribution, over plausible alternatives (Balle et al., 2022). A privacy breach occurs when the released model lets the adversary identify the hidden record, or a sufficiently close substitute, better than what could be achieved otherwise from its prior information alone. For example, if the adversary has narrowed the target to m plausible candidates and has no further information, an oblivious exact-identification guess succeeds with probability $1/m$; a meaningful reconstruction audit should be interpreted relative to this baseline.

Differential privacy. The gold standard for privacy-preservation of training records is differential privacy or DP (Dwork & Roth, 2014). Two datasets $D, D' \in \mathcal{Z}^n$ are adjacent, written $D \sim D'$, if they comprise identical records except one. The two differing records must lie in the admissible record domain. A randomized mechanism $M : \mathcal{Z}^n \to \Theta$ satisfies (ε, δ)-differential privacy if, for every adjacent pair $D \sim D'$ and measurable set $S \subseteq \Theta$, $\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S] + \delta$. This indistinguishability bounds how much the release can improve an adversary's inference about any protected record; it does not imply zero leakage.

Gaussian output perturbation. In convex learning, we can ensure DP by output perturbation. Specifically, instead of releasing the ERM output $\widehat{\theta}(D)$, we release $\widetilde{\theta}(D) = \widehat{\theta}(D) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I_p)$. Here, $\mathcal{N}(0, \sigma^2 I_p)$ denotes a Gaussian distribution with mean 0 and covariance matrix $\sigma^2 I_p$. The variance σ^2 scales with the ℓ_2 sensitivity of the original unperturbed model (Dwork & Roth, 2014; Chaudhuri et al., 2011; Balle & Wang, 2018): $\Delta_2 := \sup_{D \sim D'} \|\widehat{\theta}(D) - \widehat{\theta}(D')\|_2$. Specifically, when $\sigma \geq \frac{\Delta_2}{\varepsilon} \sqrt{2 \log(1.25/\delta)}$, the Gaussian mechanism satisfies (ε, δ)-DP.

Finite-prior exact-identification audit.

We audit reconstruction by identifying the hidden record among plausible candidates inside the declared neighborhood. The auditor computes the trained model for each candidate and selects the candidate with highest posterior probability given the observed Gaussian release. Under a uniform prior, this reduces to the nearest-model rule. Section 4.1 formalizes the attack and the audit algorithm.

3 BALL-DP: INDISTINGUISHABILITY OVER RESTRICTED ADJACENCY

Ball-DP introduces an explicit adjacency policy for mitigating reconstruction threats. Specifically, instead of asking the privacy mechanism to hide every admissible one-record replacement, it hides

one-record replacements inside a declared metric ball. The radius of this ball becomes part of the overall privacy budget along with the indistinguishability parameters.

3.1 Adjacency policy definition

Fix a public metric d on the record domain $\mathcal{Z}$. For embeddings, d is typically the Euclidean distance on the embedding coordinate, possibly with label constraints as in (2).

Definition 3.1 (r-adjacency)**.** For radius $r > 0$, datasets $D = (z_1, \ldots, z_n)$ and $D' = (z'_1, \ldots, z'_n)$ are *r-adjacent*, written $D \sim_r D'$, if $\exists$ an index $j \in [n]$ such that $z_i = z'_i$ for all $i \neq j$ and $d(z_j, z'_j) \leq r$.

We now define Ball-DP as (ε, δ)-indistinguishability over r-adjacency.

Definition 3.2 (Ball-DP)**.** A randomized mechanism $M : \mathcal{Z}^n \to \Theta$ satisfies (r, ε, δ)*-Ball-DP* if, for every measurable $S \subseteq \Theta$ and every $D \sim_r D'$, $\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S] + \delta$.

The radius r changes the scope of the adjacency relation being protected, not the semantics of (ε, δ) indistinguishability on that relation. For $0 \leq r' \leq r$, every mechanism satisfying (r, ε, δ)-Ball-DP also satisfies $(r', \varepsilon, \delta)$-Ball-DP. Increasing r enables protection against a larger number of substitutions, at the cost of additional noise; a trade-off made precise later in the following.

Proposition 3.3 (DP as a special case)**.** *Suppose that $\forall z \in \mathcal{Z}$, we have $\|z\|_2 \leq B$ and $d(z, z') := \|z - z'\|_2$. Then, (r, ε, δ)-Ball-DP implies (ε, δ)-DP when $r \geq 2B$.*

The metric d and radius r specify the substitutions protected by Ball-DP. Throughout, we use the following label-preserving metric; extending the analysis to other metrics is left to future work.

Label-preserving distance metric. Throughout this paper, records are pairs (x, y) and we use the label-preserving Euclidean metric:

$$d\big((x, y), (x', y')\big) = \begin{cases} \|x - x'\|_2, & y = y', \\ \infty, & y \neq y'. \end{cases} \tag{2}$$

Finite-radius adjacency therefore protects embedding changes conditional on the label. It does not protect label changes; that would require a metric assigning finite distance to cross-label substitutions. This choice also makes class counts invariant over adjacent datasets, which is why count-aware calibration does not violate indistinguishability.

The metric, radius, embedding map, norm bound, and any count metadata used for calibration are treated as public or fixed by the adjacency policy. This is consistent with settings where a private learner is trained on fixed embeddings supplied by an external representation model. If any of these calibration quantities are estimated from sensitive records, that estimation must be privatized or justified as public side information before the Ball-DP claim applies.

3.2 Ball-DP for convex learning

We next connect radius-restricted adjacency to Gaussian output perturbation for the ERM problem (1). The key quantity is how much a single record replacement can change the gradient contribution of the loss. We capture this through the following uniform regularity assumption.

Assumption 3.4 (Lipschitz smoothness in sample)**.** For the parameter set $\Theta \subseteq \mathbb{R}^p$ containing all ERM minimizers of interest, assume that there exists $L_z \geq 0$ such that, for all $\theta \in \Theta$ and $z, z' \in \mathcal{Z}$, $\|\nabla_\theta \ell(\theta; z) - \nabla_\theta \ell(\theta; z')\|_2 \leq L_z d(z, z')$.

Section A gives the constant L_z for the ridge-prototype model used in the main experiments and states the corresponding constants for the additional convex heads. The derivations for the secondary heads are deferred to Appendix H.

Theorem 3.5 (r-adjacency ℓ_2-sensitivity of strongly convex ERM)**.** *Assume that F_D in (1) is differentiable and λ-strongly convex for every dataset D, and that Assumption 3.4 holds on parameter set Θ. Then, for all $D \sim_r D'$, we have $\left\|\widehat{\theta}(D) - \widehat{\theta}(D')\right\|_2 \leq \frac{L_z}{\lambda n} r$.*

The proof of Theorem 3.5 can be found in Appendix E.2. The theorem shows that the distance between the corresponding ERM solutions is bounded linearly in the radius r. Therefore, the Gaussian output perturbation can be calibrated to the protected radius r rather than to any arbitrary replacement.

In what follows, Φ denotes the standard normal cumulative distribution function (CDF). For sharper Gaussian calibration, we use the analytic Gaussian mechanism scale of (Balle & Wang, 2018). Given an ℓ_2 sensitivity bound Δ, define

$$\mathsf{AGM}(\varepsilon, \delta, \Delta) := \inf\left\{\sigma > 0 : \Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) \le \delta\right\},$$

where $\mathsf{AGM}(\varepsilon, \delta, 0) = 0$ by convention.

Theorem 3.6 (Ball-Gaussian output perturbation)**.** *Fix $r, \varepsilon > 0$ and $\delta \in (0, 1)$. Let $f(D) = \widehat{\theta}(D)$ and suppose r-adjacency ℓ_2-sensitivity is bounded by Δ_r, i.e., $\sup_{D \sim_r D'} \|f(D) - f(D')\|_2 \le \Delta_r$. If $M(D) = f(D) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I_p)$ and $\sigma \ge \mathsf{AGM}(\varepsilon, \delta, \Delta_r)$, then M satisfies (r, ε, δ)-Ball-DP.*

The proof is in Appendix E.3.

Alternatively, if $\sigma \ge \frac{\Delta_r}{\varepsilon}\sqrt{2\log(1.25/\delta)}$, then the same conclusion follows from the classic Gaussian mechanism bound (Dwork & Roth, 2014). Moreover, under Assumption 3.4, due to Theorem 3.5 it suffices to use $\Delta_r = \frac{L_z}{\lambda n} r$. Note that the Gaussian mechanism calibrated to (r, ε, δ)-Ball-DP also satisfies $(t, \varepsilon_t, \delta_t)$-Ball-DP for $t > r$ as long as the indistinguishability parameters ε_t, δ_t scales with the ratio t/r. The proof of Proposition 3.7 is deferred to Appendix E.4.

Proposition 3.7 (Distance-dependent guarantee for Gaussian release)**.** *Suppose Assumption 3.4 holds true and the Gaussian mechanism M with noise parameter σ satisfies (r, ε, δ)-Ball-DP. Then, for any $t \ge r$, M satisfies $(t, \varepsilon_t, \delta_t)$-Ball-DP as long as*

$$\delta_t \ge \Phi\left(\frac{\Delta_t}{2\sigma} - \frac{\varepsilon_t\sigma}{\Delta_t}\right) - e^{\varepsilon_t}\Phi\left(-\frac{\Delta_t}{2\sigma} - \frac{\varepsilon_t\sigma}{\Delta_t}\right),$$

where $\Delta_t := \frac{L_z}{\lambda n} t$. Moreover, under the classical sufficient calibration $\sigma \ge (\Delta_r/\varepsilon)\sqrt{2\log(1.25/\delta)}$, with $\Delta_r := L_z r/(\lambda n)$, the mechanism satisfies $(t, \varepsilon_t, \delta)$-Ball-DP for $\varepsilon_t := \varepsilon t/r$.

Thus, a release calibrated at radius r admits weaker privacy at larger radii $t > r$. If $\|z\|_2 \le B$ for all records, setting $t = 2B$ recovers a standard-DP guarantee via Propositions 3.3 and 3.7.

For completeness, Appendix F.2 provides the Ball-Rényi DP analysis.

4 Ball-ReRo Certificates and Audits

We now translate radius-restricted indistinguishability into reconstruction-robustness statements. First, we state the exact finite-prior MAP audit used in the experiments. Second, we define the Ball-ReRo baseline and derive a certificate from Ball-DP. Third, we give the sharper direct Gaussian certificate used empirically.

Throughout this section, the adversary knows the public remainder dataset D^-, the training algorithm, the hyperparameters, the Ball-DP policy radius, and the candidate-construction protocol. The hidden record is Z, the completed dataset is $D^- \cup \{Z\}$, and the release is $\widetilde{\theta} = \widehat{\theta}(D^- \cup \{Z\}) + \xi$, with $\xi \sim \mathcal{N}(0, \sigma^2 I)$. The audit asks whether $\widetilde{\theta}$ identifies the hidden record better than the prior and allowed candidate set alone.

4.1 Finite-prior MAP audit for Gaussian releases

We consider a finite candidate set $S = \{z_1, \ldots, z_m\} \subseteq \mathcal{B}_d(u, r)$ containing the hidden target, with prior weights $\pi_1, \ldots, \pi_m > 0$ satisfying $\sum_{i=1}^m \pi_i = 1$. This is an adversary-favorable audit: the side information has already reduced the reconstruction problem to choosing among plausible local alternatives. The remaining question is whether the released model identifies which candidate was used as the hidden training record.

Proposition 4.1 (Exact finite-prior MAP attack for Gaussian release)**.** *For the Gaussian release above, a Bayes-optimal exact-identification attack under the finite prior $\pi_S = \sum_{i=1}^m \pi_i \delta_{z_i}$ is*

$$\widehat{z}_{\mathrm{MAP}} \in \arg\max_{z_i \in S}\left\{\log \pi_i - \frac{1}{2\sigma^2}\left\|\widetilde{\theta} - \widehat{\theta}(D^- \cup \{z_i\})\right\|_2^2\right\}.$$

Algorithm 1 Finite-prior MAP audit for Gaussian release

Require: observed release $\widetilde{\theta}$, public remainder D^-, noise scale σ
Require: candidate set $S = \{z_1, \dots, z_m\} \subseteq \mathcal{B}_d(u, r)$ and prior weights $\pi_1, \dots, \pi_m$
1: **for** $i = 1, \dots, m$ **do**
2: $\quad \theta_i \leftarrow \widehat{\theta}(D^- \cup \{z_i\})$ ▷ candidate model under target z_i
3: $\quad s_i \leftarrow \log \pi_i - \dfrac{1}{2\sigma^2}\|\widetilde{\theta} - \theta_i\|_2^2$ ▷ Gaussian log-posterior up to constants
4: **end for**
5: **return** z_i for any $\hat{\imath} \in \arg\max_i s_i$

For a uniform prior on S, the $\log \pi_i$ term is constant and the attack reduces to constrained least squares over the candidate set. The proof is deferred to Appendix G.

4.2 Ball-supported priors and feasible outputs

A reconstruction certificate must be interpreted relative to the adversary's side information. In particular, if the audit protocol restricts the reconstructor to a feasible output set, then the oblivious baseline should obey the same restriction. We therefore define the prior support, the allowed-output baseline, and the Ball-ReRo property using the same feasible-output family.

Definition 4.2 (Ball-supported prior)**.** A family of priors $\{\pi_{u,r}\}_{u \in \mathcal{Z}}$ is *Ball-supported* if each $\pi_{u,r}$ is supported on the policy ball $\mathcal{B}_d(u, r) \coloneqq \{z \in \mathcal{Z} : d(z, u) \le r\}$. The record u is a public anchor or reference record and need not itself be sampled from the prior.

Definition 4.3 (Allowed-output reconstruction baseline)**.** Fix a prior π, reconstruction loss $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$, tolerance $\eta \ge 0$, and feasible output set $\mathcal{A} \subseteq \mathcal{Z}$. Define $\kappa_{\pi,\rho}(\eta; \mathcal{A}) := \sup_{a \in \mathcal{A}} \Pr_{Z \sim \pi}[\rho(Z, a) \le \eta]$. This is the best success probability of an oblivious adversary that ignores the mechanism output while obeying the same feasible-output restriction as the reconstructor. When the prior, loss, tolerance, and feasible set are clear from context, we write this baseline simply as κ.

Definition 4.4 (Ball-ReRo)**.** Fix a radius $r > 0$, a reconstruction loss ρ, a tolerance $\eta \ge 0$, a Ball-supported prior family $\{\pi_{u,r}\}_{u \in \mathcal{Z}}$, and feasible output sets $\{\mathcal{A}_{u,r}\}_{u \in \mathcal{Z}}$ with $\mathcal{A}_{u,r} \subseteq \mathcal{Z}$. A mechanism $M : \mathcal{Z}^n \to \Theta$ satisfies (η, γ, r)-Ball-ReRo with respect to $(\rho, \{\pi_{u,r}\}, \{\mathcal{A}_{u,r}\})$ if, for every public remainder dataset $D^- \in \mathcal{Z}^{n-1}$, every anchor $u \in \mathcal{Z}$, and every reconstructor $R : \Theta \times \mathcal{Z}^{n-1} \to \mathcal{A}_{u,r}$,

$$\Pr_{Z \sim \pi_{u,r},\, \theta \sim M(D^- \cup \{Z\})} \big[\rho(Z, R(\theta, D^-)) \le \eta\big] \le \gamma.$$

For finite exact-identification audits, the reconstructor is constrained to output one of the candidates in S. With a uniform prior on m candidates and exact-identification loss, the baseline in Definition 4.3 is therefore $\kappa = 1/m$. A smaller policy radius may make it harder to construct a large, well-separated feasible set near a given anchor; this is part of the operational trade-off and must be reported explicitly with the audit.

Theorem 4.5 (Ball-DP implies Ball-ReRo)**.** *Fix $D^- \in \mathcal{Z}^{n-1}$, an anchor $u \in \mathcal{Z}$, a Ball-supported prior $\pi_{u,r}$, a loss ρ, a tolerance η, and a feasible output set $\mathcal{A}_{u,r} \subseteq \mathcal{Z}$. Let $\kappa \coloneqq \kappa_{\pi_{u,r},\rho}(\eta; \mathcal{A}_{u,r})$. If M satisfies (r, ε, δ)-Ball-DP, then every reconstructor $R : \Theta \times \mathcal{Z}^{n-1} \to \mathcal{A}_{u,r}$ satisfies $\Pr_{Z \sim \pi_{u,r},\, \theta \sim M(D^- \cup \{Z\})}[\rho(Z, R(\theta, D^-)) \le \eta] \le e^{\varepsilon}\kappa + \delta$. Equivalently, M satisfies $(\eta, e^{\varepsilon}\kappa + \delta, r)$-Ball-ReRo for this prior and feasible-output set.*

The proof is in Appendix F.1. This radius-restricted DP-to-reconstruction bound (Balle et al., 2022) is interpreted relative to κ, the best oblivious success probability under the same prior and feasible-output restriction. Audits therefore compare reconstruction success with κ, rather than treating a high raw success rate as evidence of release-induced leakage.

4.3 Gaussian Ball-ReRo bounds

The generic Ball-DP certificate above can be loose for Gaussian releases. We therefore use the testing, or blow-up-function, viewpoint (Balle et al., 2022; Kaissis et al., 2023; Hayes et al., 2023). For release

distributions P, Q and $\kappa \in [0,1]$, write $\mathcal{B}_\kappa(P,Q) := \sup\{P(E) : E \text{ measurable and } Q(E) \le \kappa\}$. This quantity asks how large an attack-success event can be under the target-dependent release law if the same event has baseline probability at most κ under a reference release law. In our Gaussian output-perturbation setting, the relevant output laws have equal covariance, so this testing problem has a closed form; Appendix F.5 proves the Gaussian identity used below. The next results are the certificates plotted in the experiments and should be read as release-level upper bounds under the stated feasible-set prior.

Theorem 4.6 (Direct Gaussian Ball-ReRo)**.** *Let $M(D) = f(D) + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 I)$, and let $\Delta_2(r)$ denote the r-adjacency sensitivity of f. Fix $D^- \in \mathcal{Z}^{n-1}$, an anchor $u \in \mathcal{Z}$, a Ball-supported prior $\pi_{u,r}$, a reconstruction loss ρ, a tolerance η, and a feasible output set $\mathcal{A}_{u,r} \subseteq \mathcal{Z}$. Define $\kappa := \kappa_{\pi_{u,r},\rho}(\eta; \mathcal{A}_{u,r})$ and $\bar{\Delta}(D^-) := \sup_{z \in \mathrm{supp}(\pi_{u,r})} \|f(D^- \cup \{z\}) - f(D^- \cup \{u\})\|_2$. Then every reconstructor R with outputs in $\mathcal{A}_{u,r}$ satisfies*

$$\Pr[\rho(Z, R(\theta, D^-)) \le \eta] \le \Phi\left(\Phi^{-1}(\kappa) + \frac{\bar{\Delta}(D^-)}{\sigma}\right) \le \Phi\left(\Phi^{-1}(\kappa) + \frac{\Delta_2(r)}{\sigma}\right),$$

Here the displayed bound is extended continuously at the endpoint baselines: if $\kappa = 0$ the bound is 0, and if $\kappa = 1$ the bound is 1. The proof is in Appendix F.6.

Corollary 4.7 (Uniform finite prior under exact identification)**.** *Suppose $\pi_S = m^{-1} \sum_{i=1}^m \delta_{z_i}$ is uniform on m candidates in $S \subseteq \mathcal{B}_d(u, r)$, and the decoder is constrained to output a candidate in S. Consider the exact-identification loss $\rho_{\mathrm{ID}}(z, z') := \mathbf{1}\{z \ne z'\}$ with tolerance $\eta = 0$. Then, every reconstructor R satisfies*

$$\Pr[R(\theta, D^-) = Z] \le \Phi\left(\Phi^{-1}(1/m) + \frac{\Delta_2(r)}{\sigma}\right),$$

where the probability is over $Z \sim \pi_S$ and $\theta \sim M(D^- \cup \{Z\})$. The feasible-output oblivious baseline is exactly $\kappa = 1/m$.

The proof is in Appendix F.7. This is the certificate used in the experiments. At fixed (ε, δ, m), homogeneous calibration keeps $\Delta_2(r)/\sigma$ constant: smaller radii reduce noise and restrict protection, rather than tightening the certificate. MAP audits measure identification within the feasible candidate sets. Appendix F.8 gives a sharper support-specific bound, not used in the reported experiments.

5 Experimental Analysis

Experimental settings. All ERM experiments use $\delta = 10^{-6}$, $m = 10$, and Gaussian output perturbation. Ball-DP uses same-label distance quantiles $q50$, $q80$, and $q95$ as policy radii. The standard comparator uses the global bounded-replacement radius $2B$ for utility experiments and the empirical same-label diameter for reconstruction audits.

5.1 Experimental model

The seven-dataset utility and MAP audits use a ridge-prototype classifier on fixed embeddings because its optimizer, sensitivity, and informed-adversary inverse problem are explicit. Ball-DP is not ridge-specific: Section A.2 gives the required constants for logistic and squared-hinge heads, and Section D evaluates additional convex heads on IMDb.

For record (z, y) with $z \in \mathbb{R}^d$ and $y \in \{1, \dots, C\}$, the ridge-prototype model learns one prototype $\mu_c \in \mathbb{R}^d$ per class by minimizing

$$F_D(\mu) = \frac{1}{n} \sum_{i=1}^n \|z_i - \mu_{y_i}\|_2^2 + \frac{\lambda}{2} \sum_{c=1}^C \|\mu_c\|_2^2. \tag{3}$$

Prediction uses the nearest prototype and all reported runs use $\lambda = 0.01$; model-specific statements and proofs are collected in Section A.

For this model, Corollary A.2 gives the count-worst-case local sensitivity $\Delta_{\mathrm{proto}}(r) = 2r/(2 + \lambda n)$. When class counts are public, the count-aware variant is stated separately in Proposition A.3. Gaussian output perturbation is then calibrated by substituting the relevant sensitivity into Theorem 3.6.

5.2 Utility at local policy radii

At $\varepsilon = 2$, all three Ball-DP policies outperform the global-radius comparator on all seven benchmarks (Appendix B, Table 1). Accuracy gains are 0.57–36.47, 0.51–29.86, and 0.47–25.07 percentage points at $q50$, $q80$, and $q95$, respectively; $q95$ provides the broadest local protection.

Text embeddings use `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019; Wang et al., 2020; Sentence Transformers, 2026), and image embeddings use ImageNet-pretrained ResNet-18 features (He et al., 2016; Paszke et al., 2019; PyTorch contributors, 2026). These encoders are treated as public preprocessing. The simple ridge-prototype head is intended to isolate privacy calibration rather than provide benchmark-leading accuracy.

5.3 Accuracy at a fixed reconstruction certificate

Figure 2 plots private-ERM accuracy against the certified exact-identification bound γ from Corollary 4.7 on four benchmarks; the seven-benchmark frontier is in Appendix B. Higher accuracy is better, and a smaller γ (farther left) corresponds to stronger certified reconstruction robustness.

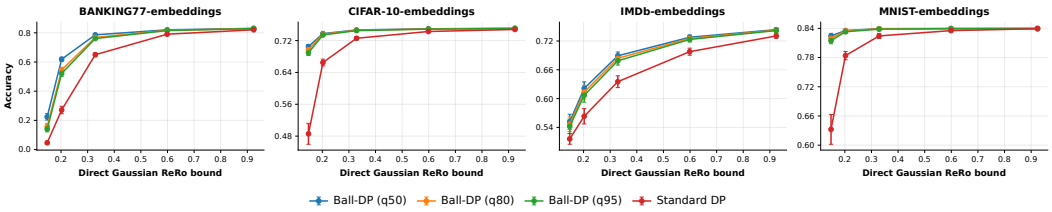

Figure 2: Private-ERM accuracy versus the direct exact-ID certificate γ on four benchmarks. Solid curves are Ball-DP at $q50$, $q80$, and $q95$; the dashed curve is the global-radius comparator. Points are seed-first means over ten releases. Full seven-benchmark results are in Appendix B.

At fixed ε, matched Gaussian calibration gives the same numerical certificate across radii; r instead controls which substitutions are protected and how much noise is required.

5.4 Finite-prior reconstruction diagnostics

We use a uniform prior over $m = 10$ candidates, giving the oblivious exact-ID baseline $1/m = 0.1$. Proposition 4.1 also covers non-uniform priors, but all aggregate results below use the uniform case.

For each $\varepsilon \in \{2, 4, 8\}$ and Ball-DP policy $q50$, $q80$, or $q95$, each setup seed constructs four same-label feasible supports of size $m = 10$ inside the policy ball using farthest-point selection. Every candidate serves once as the hidden target, yielding 40 release-and-attack trials per seed and 400 per dataset and mechanism over ten seeds. We average trials within each seed before computing 95% confidence intervals across seeds. Standard DP uses the empirical same-label diameter for calibration and is evaluated on the $q80$ candidate-support protocol.

Figure 3 compares local Ball-DP with the global-radius Standard-DP comparator; complete numerical results appear in Appendix C, Table 2. The mechanisms protect different adjacency scopes, so empirical exact-ID rates need not be ordered. For $\varepsilon = 2, 4, 8$, all reported mean rates are below the respective Gaussian certificates 0.2024, 0.3286, and 0.5987; chance is $1/m = 0.1$.

6 Related Work

Adjacency design beyond the global worst case. The choice of adjacency is central to the meaning of DP Dwork & Roth (2014). Metric formulations such as $d_{\mathcal{X}}$-privacy and later work on metric DP replace the uniform adjacency graph by guarantees that depend on a record metric, especially in location and embedding domains Chatzikokolakis et al. (2013); Imola et al. (2022). Ball-DP is simpler and more policy-oriented: it keeps the standard (ε, δ) semantics unchanged and restricts the protected substitutions to a declared radius-r neighborhood.

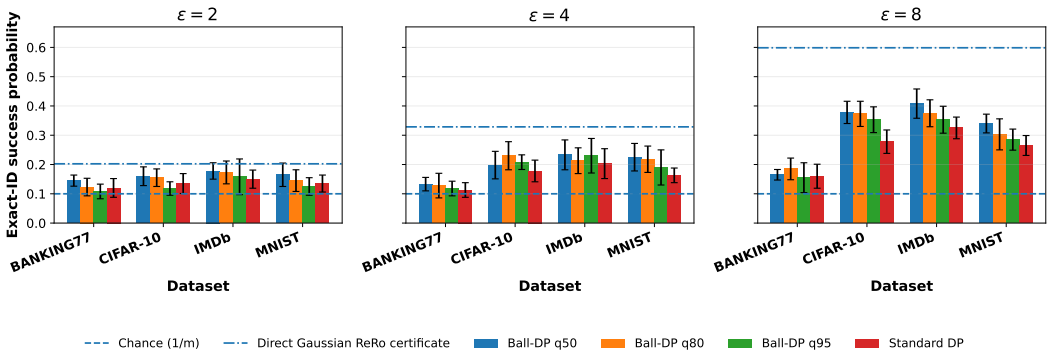

Figure 3: Exact-identification MAP audit on the four main-text benchmarks. Each panel fixes $\varepsilon \in \{2, 4, 8\}$. Bars show Ball-DP calibrated at $q50$, $q80$, and $q95$, and the global-radius Standard-DP comparator evaluated on the $q80$ candidate-support protocol. Error bars are seed-first 95% confidence intervals over ten setup seeds. The dashed line is chance, $1/m = 0.1$, and the dash-dotted line is the direct Gaussian Ball-ReRo certificate. Complete seven-benchmark results are reported in Appendix C.

Private convex ERM and output perturbation. Private ERM under convexity has a long history, including output- and objective-perturbation methods (Chaudhuri et al., 2011). In the Gaussian output-perturbation regime, analytic calibration improves over the classical tail-bound formula (Balle & Wang, 2018). Our generic ERM sensitivity theorem is the local counterpart of the usual output-perturbation analysis. Ridge prototypes give an exact equality rather than only an upper bound, while the binary-logistic, softmax-logistic, and squared-hinge SVM constants show how the same local-calibration argument extends to common convex linear heads.

Reconstruction attacks and reconstruction robustness. Balle et al. (2022) formalized the informed-adversary reconstruction setting and showed that convex models can admit exact or near-exact inversion from noiseless releases. Hayes et al. (2023) derived reconstruction bounds and matching attacks for DP-SGD. The hypothesis-testing view of privacy, developed in the f-DP line of work, provides a natural route from privacy guarantees to attack bounds (Dong et al., 2022). Kaissis et al. (2023) used this perspective to derive direct reconstruction bounds for Gaussian and subsampled mechanisms. Our Gaussian Ball-ReRo result is the local, radius-thresholded analogue of that direct Gaussian testing argument.

Prototype-based private learning. Prototype representations are common in classical nearest-centroid classifiers and modern embedding-based learning (Snell et al., 2017). Recent private learning work has also explored prototype releases in transfer-learning settings (Wahdany et al., 2025). Our use of prototypes is different: the model is chosen as the closed-form audit case because its convex structure yields exact local sensitivity and an exact informed-adversary audit. The broader convex-head contribution is captured by the generic ERM theorem and the explicit logistic/SVM constants.

7 Summary and Future Directions

This paper introduced Ball-DP, which restricts the protected adjacency relation to a declared radius while retaining standard (ε, δ) indistinguishability within that neighborhood. For strongly convex ERM, r-adjacency sensitivity scales linearly with r, enabling radius-aware Gaussian output perturbation. We instantiate the analysis for ridge prototypes and give constants for logistic and squared-hinge heads. For ridge prototypes, sensitivity and finite-prior MAP reconstruction are explicit, enabling theorem-aligned reconstruction audits.

Across all the seven tested embedding benchmarks, local calibration improves utility relative to global replacement while empirical exact-ID rates remain below the Ball-ReRo certificates. Additional IMDb experiments show the same utility-certificate pattern for other convex heads. The present analysis is limited to convex one-shot Gaussian releases; extending Ball-DP to DP-SGD transcripts and nonconvex representation learning is left for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing for concision and readability, LaTeX restructuring, and identifying minor notation and consistency issues. We did not use generative AI tools to conceive the project idea or develop the theoretical results, formulate mathematical claims, construct proofs, process data, or interpret the results. The authors reviewed all AI-assisted suggestions and take responsibility for the final content of the manuscript.

ETHICS STATEMENT

This paper studies privacy guarantees and reconstruction attacks for machine-learning models trained on potentially sensitive data. The experiments use fixed embedding representations and theorem-aligned attack protocols to audit privacy leakage under controlled assumptions, not to facilitate unauthorized extraction of personal data from deployed systems. Empirical attack evaluations should be run only on datasets for which the researchers have authorization, and released artifacts should omit raw private records. No human-subject intervention is involved in the analysis reported here. The experiments evaluate candidate identification in embedding space rather than recovered raw images or text.

REFERENCES

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 2018. URL https://proceedings.mlr.press/v80/balle18a.html.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *Proceedings of the 43rd IEEE Symposium on Security and Privacy (S&P)*, pp. 1138–1156. IEEE, 2022. doi: 10.1109/SP46214.2022.9833677.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás E. Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies: 13th International Symposium, PETS 2013*, volume 7981 of *Lecture Notes in Computer Science*, pp. 82–102. Springer, 2013. doi: 10.1007/978-3-642-39077-7_5.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011. doi: 10.5555/1953048.2021036. URL https://jmlr.org/papers/v12/chaudhuri11a.html.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii, Yu-Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, and Wanrong Zhang. Advancing differential privacy: Where we are now and future directions for real-world deployment. *Harvard Data Science Review*, 6(1), 2024a. doi: 10.1162/99608f92.d3197524.

Rachel Cummings, Shlomi Hod, Jayshree Sarathy, and Marika Swanberg. ATTAXONOMY: Unpacking differential privacy guarantees against practical adversaries. *CoRR*, abs/2405.01716, 2024b. doi: 10.48550/arXiv.2405.01716. URL https://arxiv.org/abs/2405.01716.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

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. doi: 10.1111/rssb.12454.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 2 edition, 1999. ISBN 9780471317166.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7.

Jamie Hayes, Borja Balle, and Saeed Mahloujifar. Bounding training data reconstruction in DP-SGD. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f8928b073ccbec15d35f2a9d39430bfd-Abstract-Conference.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

Jacob Imola, Shiva Prasad Kasiviswanathan, Stephen White, Abhinav Aggarwal, and Nathanael Teissier. Balancing utility and scalability in metric differential privacy. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 885–894. PMLR, 2022. URL https://proceedings.mlr.press/v180/imola22a.html.

Georgios Kaissis, Jamie Hayes, Alexander Ziller, and Daniel Rueckert. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy. *CoRR*, abs/2307.03928, 2023. doi: 10.48550/arXiv.2307.03928. URL https://arxiv.org/abs/2307.03928.

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

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. doi: 10.1098/rsta.1933.0009.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019. URL https://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.

PyTorch contributors. TorchVision resnet-18 model documentation. PyTorch documentation, 2026. URL https://docs.pytorch.org/vision/stable/models/generated/torchvision.models.resnet18.html. Accessed 2026-06-11.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

Sentence Transformers. sentence-transformers/all-MiniLM-L6-v2. Hugging Face model card, 2026. URL https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed 2026-06-11.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4077–4087, 2017. URL https://papers.nips.cc/paper/6996-prototypical-networks-for-few-shot-learning.

Dariush Wahdany, Matthew Jagielski, Adam Dziedzic, and Franziska Boenisch. Differentially private prototypes for imbalanced transfer learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):20991–20999, 2025. doi: 10.1609/aaai.v39i20.35395.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

A Technical Statements for the Experiments

This section records the model-specific facts used by the experimental section. The statements serve two purposes: they specify the Gaussian sensitivity calibration for the ridge-prototype head, and they show how the generic Ball-DP theorem extends to the additional convex heads used in the IMDb comparison. Proofs are deferred to the appendices.

A.1 Ridge-prototype head

The ridge-prototype head is the closed-form audit case. Its optimizer, sensitivity, and informed-adversary inverse problem are explicit, so it supports exact MAP reconstruction diagnostics without numerical approximation error. The statements below are therefore the ridge-specific instantiation of the general convex-ERM theory, not the full scope of the method.

Theorem A.1 (Exact gradient Lipschitz constant for ridge prototypes)**.** *Under the label-preserving distance metric defined in Section 3.1, Assumption 3.4 holds for $\ell(\mu;(z,y)) = \|z-\mu_y\|_2^2$ with the exact constant $L_z = 2$.*

Corollary A.2 (Exact r-adjacency sensitivity)**.** *Let $\widehat{\mu}(D)$ be the unique minimizer of* (3)*. The count-worst-case sensitivity at radius r is*

$$\Delta_{\mathrm{proto}}(r) = \sup_{D\sim_r D'} \|\widehat{\mu}(D) - \widehat{\mu}(D')\|_2 = \frac{2r}{2+\lambda n}.$$

This is sharper than the generic upper bound from Theorem 3.5.

Corollary A.2 is count-worst-case: it allows the changed record to be the only record in its class. If class counts are public, or are fixed component metadata of the label-preserving adjacency graph, the calibration can instead use the smallest nonzero public class count.

Proposition A.3 (Count-aware ridge-prototype sensitivity)**.** *Let n_c be the public number of records in class c, and let*

$$n_{\min} := \min\{n_c : n_c > 0\}.$$

Conditional on these counts, the exact r-adjacency sensitivity of the ridge-prototype optimizer is

$$\Delta_{\mathrm{proto}}(r; n_1, \ldots, n_C) = \max_{c:n_c>0} \frac{2r}{2n_c+\lambda n} = \frac{2r}{2n_{\min}+\lambda n}.$$

The maximum is attained by a class with count $n_{\min}$, since one changed record moves rare-class prototypes the most. Because $n_{\min} \geq 1$, the count-aware calibration is never looser than the count-worst-case bound $2r/(2+\lambda n)$, and it is strictly tighter whenever $n_{\min} > 1$.

Theorem A.4 (Exact reconstruction from a noiseless release)**.** *Let $D = D^- \cup \{(z, y)\}$ and suppose an informed adversary knows D^-, n, λ, and the noiseless optimum $\widehat{\mu}(D)$. If the missing label y is known, then*

$$z = \frac{2(n_y^- + 1) + \lambda n}{2}\, \widehat{\mu}_y(D) - \sum_{(z',y') \in D^- :\, y' = y} z',$$

where n_y^- is the number of public label-y examples.

If the missing label is unknown, define

$$\mu_c^0(D^-) := \frac{2 \sum_{(z',y') \in D^- :\, y' = c} z'}{2n_c^- + \lambda n}$$

for every class c with $n_c^- > 0$. If there is a unique class c such that $\widehat{\mu}_c(D) \neq \mu_c^0(D^-)$, then $y = c$ and the same formula reconstructs z.

The only label-identification obstruction is the degenerate event in which the missing example equals the D^--implied prototype of its own class. In that case, adding the missing record does not move the corresponding released prototype.

Corollary A.5 (Known-label noisy inversion)**.** *Fix D^- and suppose the missing label y is known. Let*

$$A_y := 2(n_y^- + 1) + \lambda n, \qquad S_y^- := \sum_{(z',y') \in D^- :\, y' = y} z'.$$

For the Gaussian release, define

$$W_y = (A_y/2)\widetilde{\mu}_y - S_y^-.$$

If the hidden record is Z, then

$$W_y = Z + \eta_y, \qquad \eta_y \sim \mathcal{N}(0, \tau_y^2 I), \qquad \tau_y = (A_y/2)\sigma.$$

Thus, for a uniform finite prior with common known label y, the exact MAP attack is nearest-neighbor decoding of W_y over the candidate support.

A.2 Logistic and SVM heads

The same ERM sensitivity theorem applies to the additional convex heads used in the experiments once Assumption 3.4 is verified for each model. For the main text, we record only the resulting constants needed for the Ball-DP calibration; the proof of Proposition A.6 is deferred to Appendix H.

Throughout this subsection we use the label-preserving metric, assume $\|z\|_2 \le B$, and write

$$\widetilde{B} := \sqrt{B^2 + 1}$$

for the augmented feature norm.

Proposition A.6 (Explicit L_z constants for common convex heads)**.** *For same-label substitutions, Assumption 3.4 holds for the following convex heads with the corresponding dataset-independent constants:*

- Squared hinge loss:
$$L_z \le 2 + 4B\sqrt{2/\lambda}.$$
- Binary logistic loss:
$$L_z \le 1 + \frac{\widetilde{B}^2}{4\lambda}.$$
- Softmax logistic loss:
$$L_z \le \sqrt{2}\left(1 + \frac{\widetilde{B}^2}{2\lambda}\right).$$

Consequently, for any of these heads, Theorem 3.5 gives

$$\|\widehat{\theta}(D) - \widehat{\theta}(D')\| \le \frac{L_z}{\lambda n} r \qquad \text{for all } D \sim_r D',$$

with L_z chosen from the corresponding case above.

B Full Seven-Benchmark Utility Results

Table 1 gives the complete $\varepsilon = 2$ utility comparison across all seven embedding benchmarks, and Figure 4 gives the corresponding full private-ERM accuracy–certificate frontier. The experimental settings and seed-first aggregation are the same as in Section 5; Figure 2 shows the compact four-benchmark main-text view.

Table 1: Mean accuracies over 10 independent Gaussian release seeds for the ridge-prototype private ERM experiment at $\varepsilon = 2$ and $m = 10$ across 7 embedding benchmarks. Ball-DP columns use local adjacency radii $q50$, $q80$, and $q95$; Std is the global-radius comparator (standard DP) calibrated with the same Gaussian rule.

Data embeddings	Acc $q50$	Acc $q80$	Acc $q95$	Acc (Std)
AG News	0.8616	0.8610	0.8606	0.8559
BANKING77	0.6358	0.5697	0.5218	0.2711
CIFAR-10	0.7374	0.7360	0.7340	0.6649
Emotion	0.3010	0.2932	0.2887	0.2458
IMDb	0.6212	0.6135	0.6071	0.5632
MNIST	0.8356	0.8345	0.8330	0.7840
TREC-6	0.2496	0.2460	0.2412	0.2036

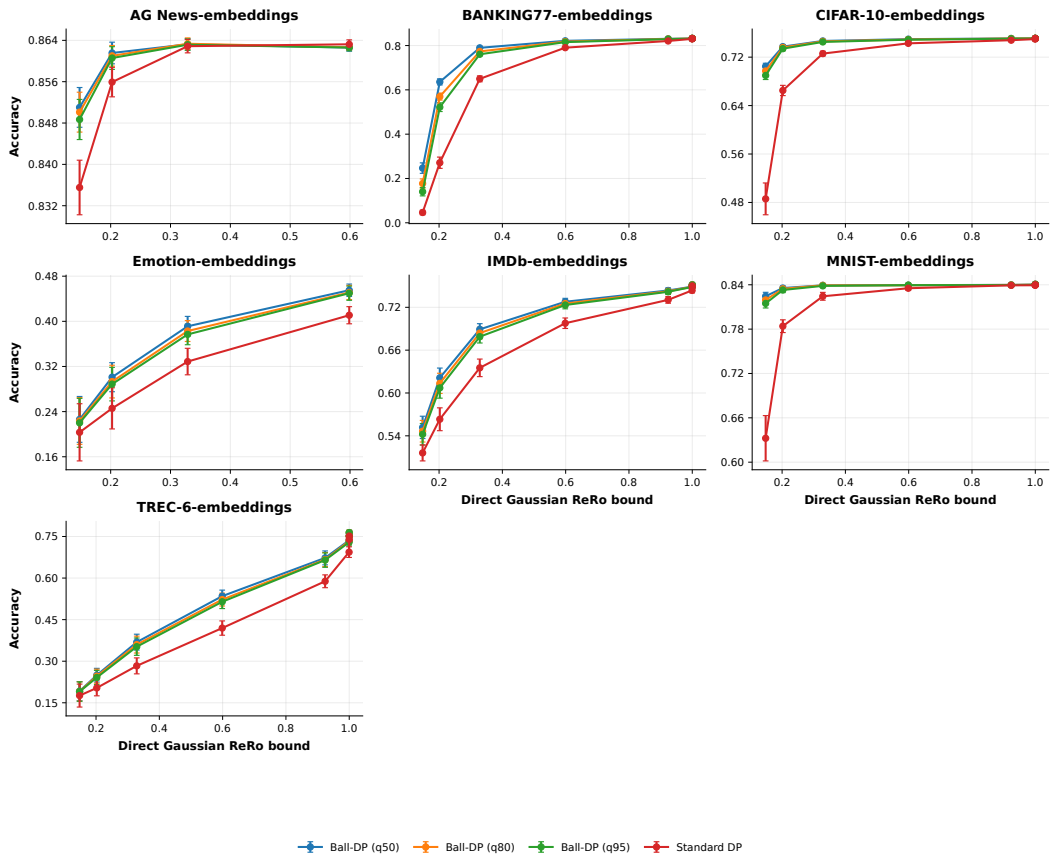

Figure 4: Complete seven-benchmark private-ERM accuracy versus the direct exact-identification upper bound γ for the ridge-prototype head. Solid curves are Ball-DP mechanisms calibrated at local radii $q50$, $q80$, and $q95$; the dashed curve is the global-radius comparator. Points are seed-first means over ten independent Gaussian release seeds. This is the full version of the compact four-benchmark view in Figure 2.

C Full Seven-Benchmark Exact-Identification Results

Table 2 reports exact-ID MAP audit results for all seven benchmarks, including the four shown in Figure 3.

Table 2: Complete seven-benchmark exact-ID MAP audit. Entries are seed-first means $\pm$ 95% confidence intervals over ten setup seeds. Chance is $1/m = 0.1$; the final column is the direct Gaussian certificate.

Dataset	ε	Ball-DP $q50$	Ball-DP $q80$	Ball-DP $q95$	Std. DP ($q80$ audit)	γ_{rel}
AG News	2	0.153 ± 0.035	0.155 ± 0.029	0.170 ± 0.043	0.138 ± 0.032	0.2024
AG News	4	0.223 ± 0.034	0.205 ± 0.034	0.195 ± 0.041	0.195 ± 0.041	0.3286
AG News	8	0.370 ± 0.052	0.373 ± 0.074	0.335 ± 0.043	0.340 ± 0.041	0.5987
BANKING77	2	0.145 ± 0.019	0.123 ± 0.030	0.108 ± 0.025	0.120 ± 0.032	0.2024
BANKING77	4	0.133 ± 0.023	0.128 ± 0.042	0.118 ± 0.025	0.113 ± 0.025	0.3286
BANKING77	8	0.165 ± 0.018	0.185 ± 0.037	0.155 ± 0.051	0.160 ± 0.041	0.5987
CIFAR-10	2	0.160 ± 0.032	0.155 ± 0.030	0.118 ± 0.023	0.135 ± 0.034	0.2024
CIFAR-10	4	0.198 ± 0.047	0.230 ± 0.048	0.208 ± 0.025	0.178 ± 0.037	0.3286
CIFAR-10	8	0.378 ± 0.038	0.373 ± 0.043	0.353 ± 0.044	0.278 ± 0.040	0.5987
Emotion	2	0.093 ± 0.033	0.120 ± 0.022	0.080 ± 0.023	0.130 ± 0.035	0.2024
Emotion	4	0.088 ± 0.034	0.095 ± 0.014	0.090 ± 0.018	0.105 ± 0.028	0.3286
Emotion	8	0.153 ± 0.032	0.120 ± 0.027	0.118 ± 0.030	0.148 ± 0.023	0.5987
IMDb	2	0.178 ± 0.028	0.173 ± 0.039	0.158 ± 0.061	0.150 ± 0.031	0.2024
IMDb	4	0.233 ± 0.051	0.213 ± 0.044	0.230 ± 0.059	0.203 ± 0.051	0.3286
IMDb	8	0.408 ± 0.050	0.375 ± 0.046	0.353 ± 0.046	0.325 ± 0.037	0.5987
MNIST	2	0.165 ± 0.040	0.145 ± 0.037	0.125 ± 0.030	0.135 ± 0.029	0.2024
MNIST	4	0.225 ± 0.047	0.218 ± 0.045	0.190 ± 0.060	0.163 ± 0.025	0.3286
MNIST	8	0.340 ± 0.032	0.303 ± 0.053	0.285 ± 0.036	0.265 ± 0.034	0.5987
TREC-6	2	0.095 ± 0.019	0.123 ± 0.034	0.108 ± 0.023	0.085 ± 0.033	0.2024
TREC-6	4	0.105 ± 0.031	0.100 ± 0.032	0.128 ± 0.039	0.105 ± 0.036	0.3286
TREC-6	8	0.123 ± 0.035	0.118 ± 0.036	0.098 ± 0.019	0.108 ± 0.028	0.5987

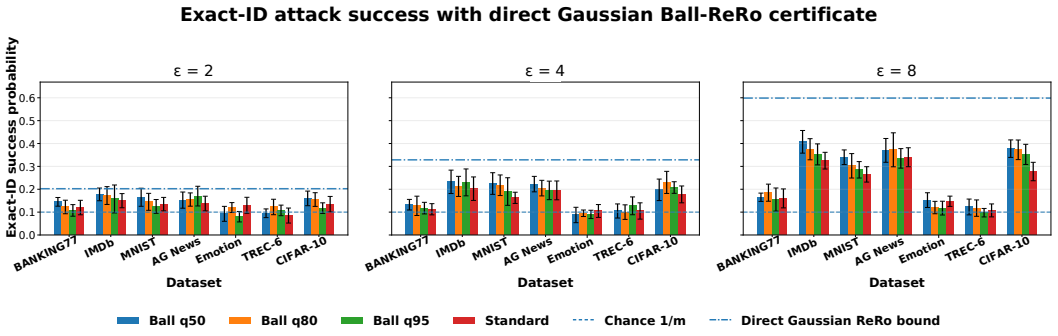

Figure 5: Complete seven-benchmark version of the exact-identification MAP audit in Figure 3. Each panel fixes $\varepsilon \in \{2, 4, 8\}$ and compares Ball-DP at $q50$, $q80$, and $q95$ with the global-radius Standard-DP comparator evaluated on the $q80$ candidate-support protocol. Error bars are seed-first 95% confidence intervals over ten setup seeds; horizontal reference lines show chance and the direct Gaussian Ball-ReRo certificate.

D Convex-head comparison on IMDb

The preceding seven-dataset frontier uses ridge prototypes to keep the exact finite-prior inverse problem explicit. The calibration theorem, however, applies to every convex head for which Assumption 3.4 is verified. Proposition A.6 gives those constants for binary logistic regression, softmax logistic regression, and squared-hinge SVMs. To make this scope visible empirically, Figure 6 repeats

the private-ERM utility–certificate sweep on IMDb embeddings at the local radius policy $q50$ for binary logistic regression, ridge prototypes, and the squared-hinge SVM. This experiment uses five Gaussian release seeds per mechanism and nominal privacy levels $\varepsilon \in \{8, 16, 32\}$.

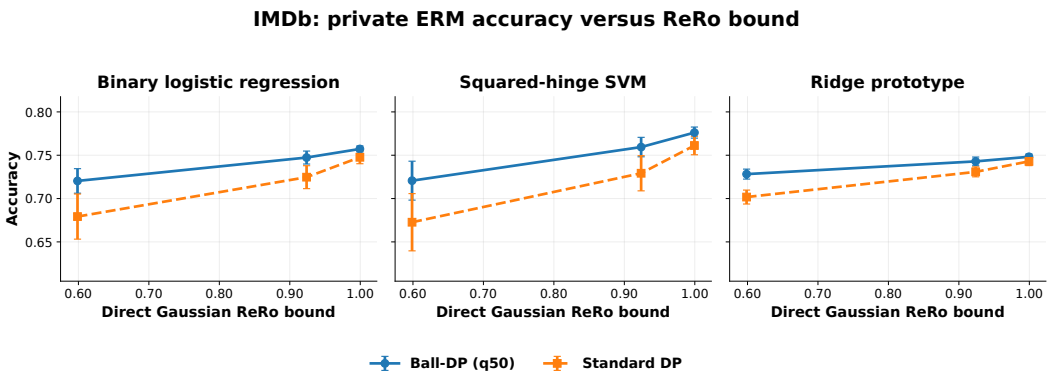

Figure 6: IMDb convex-head comparison at local radius policy $q50$. Each panel shows private-ERM accuracy as a function of the direct finite-prior ReRo certificate γ. Solid curves are Ball-DP releases calibrated at the local $q50$ policy radius; dashed curves are the global-radius Standard-DP comparator evaluated on the corresponding $q50$ row set. Points correspond to $\varepsilon \in \{8, 16, 32\}$, with $\gamma = 0.5987, 0.9238, 0.9995$, respectively. Error bars are seed-first 95% normal intervals over five Gaussian release seeds.

Table 3: IMDb convex-head comparison at local radius policy $q50$. Each cell reports mean accuracy $\pm$ seed-first 95% confidence interval over five Gaussian release seeds. The ReRo certificates are $\gamma = 0.5987, 0.9238, 0.9995$ for $\varepsilon = 8, 16, 32$, respectively, and are the horizontal-axis values used in Figure 6.

Head	Mechanism	$\varepsilon = 8$	$\varepsilon = 16$	$\varepsilon = 32$
Binary logistic	Ball-DP ($q50$)	0.720 ± 0.014	0.747 ± 0.008	0.757 ± 0.004
Binary logistic	Standard DP	0.679 ± 0.026	0.725 ± 0.013	0.748 ± 0.007
Ridge prototype	Ball-DP ($q50$)	0.728 ± 0.006	0.743 ± 0.005	0.748 ± 0.003
Ridge prototype	Standard DP	0.702 ± 0.008	0.731 ± 0.006	0.743 ± 0.005
Squared-hinge SVM	Ball-DP ($q50$)	0.721 ± 0.023	0.759 ± 0.011	0.776 ± 0.006
Squared-hinge SVM	Standard DP	0.673 ± 0.033	0.729 ± 0.020	0.761 ± 0.011

At every displayed privacy level and for all three convex heads, Ball-DP has higher mean accuracy than the global-radius comparator at the same reported ReRo certificate. The gains are largest at the strongest privacy level, $\varepsilon = 8$: 4.12 accuracy points for binary logistic regression, 2.65 points for ridge prototypes, and 4.80 points for the squared-hinge SVM. As privacy weakens, the gap shrinks: at $\varepsilon = 32$, the corresponding gains are 0.96, 0.53, and 1.49 points. This pattern is consistent with the calibration story: local radius-aware noise matters most when the Gaussian perturbation is large.

The head comparison also shows a utility trade-off that is separate from the privacy certificate. When one can afford high ε, the secondary convex heads can outperform the ridge prototype: at $\varepsilon = 32$, the squared-hinge SVM reaches the highest mean accuracy in this sweep, 0.776 under Ball-DP. However, these higher-utility heads are also more sensitive to stronger privacy in the reported experiment. From $\varepsilon = 32$ to $\varepsilon = 8$, Ball-DP accuracy drops by 3.68 points for binary logistic regression and 5.54 points for the squared-hinge SVM, compared with 2.00 points for the ridge prototype. Thus the secondary heads can be attractive in the high-utility, weak-privacy regime, while the ridge prototype remains a stable closed-form audit head for the main seven-dataset reconstruction study.

E Proofs for Main Results

E.1 Proof of Proposition 3.3

Proof. If $\|z\|_2 \le B$ and $\|z'\|_2 \le B$, the triangle inequality gives

$$\|z - z'\|_2 \le \|z\|_2 + \|z'\|_2 \le 2B.$$

Thus every admissible single-record replacement is $2B$-adjacent, so the bounded-replacement adjacency graph is contained in the Ball adjacency graph at radius $2B$. □

E.2 Proof of Theorem 3.5

We first state a generic optimizer-perturbation bound that does not depend on the metric or on any Lipschitz-in-sample assumption.

Lemma E.1 (Single-record optimizer perturbation)**.** *Let $D, D' \in \mathcal{Z}^n$ differ only at index j, with records z_j and z'_j. Suppose F_D is differentiable and λ-strongly convex, and suppose the minimizers $\widehat{\theta}(D)$ and $\widehat{\theta}(D')$ lie in a parameter set Θ. Then*

$$\|\widehat{\theta}(D) - \widehat{\theta}(D')\|_2 \le \frac{1}{\lambda n} \sup_{\theta \in \Theta} \left\|\nabla_\theta \ell(\theta; z_j) - \nabla_\theta \ell(\theta; z'_j)\right\|_2 .$$

Proof of Lemma E.1. Let

$$u := \widehat{\theta}(D), \qquad v := \widehat{\theta}(D').$$

If $u = v$, the result is immediate, so assume $u \neq v$. Since F_D is differentiable and λ-strongly convex, its gradient is λ-strongly monotone:

$$\left\langle \nabla F_D(a) - \nabla F_D(b),\, a - b \right\rangle \ge \lambda \|a - b\|_2^2 \qquad \text{for all } a, b.$$

Apply this with $a = u$ and $b = v$. Since u minimizes F_D, $\nabla F_D(u) = 0$, and therefore

$$\lambda \|u - v\|_2^2 \le \left\langle -\nabla F_D(v),\, u - v \right\rangle \le \|\nabla F_D(v)\|_2 \, \|u - v\|_2,$$

where the second inequality follows from Cauchy–Schwarz inequality. Dividing by $\|u - v\|_2$ gives

$$\lambda \|u - v\|_2 \le \|\nabla F_D(v)\|_2.$$

Since v minimizes $F_{D'}$, $\nabla F_{D'}(v) = 0$, so

$$\|\nabla F_D(v)\|_2 = \|\nabla F_D(v) - \nabla F_{D'}(v)\|_2.$$

The datasets differ only at index j, hence all unchanged records and the regularizer cancel in the gradient difference:

$$\nabla F_D(v) - \nabla F_{D'}(v) = \frac{1}{n} \left(\nabla_\theta \ell(v; z_j) - \nabla_\theta \ell(v; z'_j)\right).$$

Because $v \in \Theta$, we have

$$\left\|\nabla_\theta \ell(v; z_j) - \nabla_\theta \ell(v; z'_j)\right\|_2 \le \sup_{\theta \in \Theta} \left\|\nabla_\theta \ell(\theta; z_j) - \nabla_\theta \ell(\theta; z'_j)\right\|_2 .$$

Combining the previous displays yields

$$\|\widehat{\theta}(D) - \widehat{\theta}(D')\|_2 \le \frac{1}{\lambda n} \sup_{\theta \in \Theta} \left\|\nabla_\theta \ell(\theta; z_j) - \nabla_\theta \ell(\theta; z'_j)\right\|_2 ,$$

as claimed. □

Proof of Theorem 3.5. Let $D \sim_r D'$ and let j be the differing coordinate. By Lemma E.1,

$$\|\widehat{\theta}(D) - \widehat{\theta}(D')\|_2 \le \frac{1}{\lambda n} \sup_{\theta \in \Theta} \left\|\nabla_\theta \ell(\theta; z_j) - \nabla_\theta \ell(\theta; z'_j)\right\|_2 .$$

Assumption 3.4 gives

$$\sup_{\theta \in \Theta} \left\|\nabla_\theta \ell(\theta; z_j) - \nabla_\theta \ell(\theta; z'_j)\right\|_2 \le L_z \, d(z_j, z'_j).$$

Since $D \sim_r D'$, we have $d(z_j, z'_j) \le r$. Therefore

$$\|\widehat{\theta}(D) - \widehat{\theta}(D')\|_2 \le \frac{L_z}{\lambda n} r.$$

□

E.3 Proof of Theorem 3.6

Proof. If $\Delta_r = 0$, then $f(D) = f(D')$ for every $D \sim_r D'$, so the corresponding release distributions coincide and Ball-DP holds.

Otherwise, for every $D \sim_r D'$, the releases are Gaussian with common covariance $\sigma^2 I_p$ and means separated by at most Δ_r. By the analytic Gaussian mechanism theorem (Balle & Wang, 2018, Theorem 8), $\sigma \geq \mathsf{AGM}(\varepsilon, \delta, \Delta_r)$ implies

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S] + \delta$$

for every measurable output event S. Hence M satisfies (r, ε, δ)-Ball-DP. □

E.4 Proof of Proposition 3.7

Proof. Fix $t > 0$ and consider any pair $D \sim_t D'$. By Theorem 3.5 with radius t,

$$\|\widehat{\theta}(D) - \widehat{\theta}(D')\|_2 \leq \Delta_t := \frac{L_z}{\lambda n} t.$$

Thus the two releases are Gaussian distributions with common covariance $\sigma^2 I_p$ and means separated by at most Δ_t. The Gaussian privacy profile therefore implies that the release is $(\varepsilon_t, \delta_t)$-DP on the t-adjacency relation whenever

$$\delta_t \geq \Phi\left(\frac{\Delta_t}{2\sigma} - \frac{\varepsilon_t \sigma}{\Delta_t}\right) - e^{\varepsilon} \Phi\left(-\frac{\Delta_t}{2\sigma} - \frac{\varepsilon_t \sigma}{\Delta_t}\right)$$

For the closed-form statement, assume

$$\sigma \geq \frac{\Delta_r}{\varepsilon} \sqrt{2 \log(1.25/\delta)}, \qquad \Delta_r := \frac{L_z}{\lambda n} r.$$

Set $\varepsilon_t = \varepsilon t / r$. Since

$$\frac{\Delta_t}{\varepsilon_t} = \frac{(L_z/(\lambda n))t}{\varepsilon t / r} = \frac{L_z r}{\lambda n \, \varepsilon} = \frac{\Delta_r}{\varepsilon},$$

the chosen σ also satisfies

$$\sigma \geq \frac{\Delta_t}{\varepsilon_t} \sqrt{2 \log(1.25/\delta)}.$$

Whenever ε_t lies in the usual parameter range for the classical Gaussian mechanism bound, that bound gives (ε_t, δ)-DP on the t-adjacency relation. □

E.5 Proofs for the Ridge-Prototype Model

E.5.1 Proof of Theorem A.1

Proof. We identify the parameter $\mu = (\mu_1, \ldots, \mu_C)$ with the concatenated vector in $\mathbb{R}^{Cd}$ obtained by stacking the class blocks. Under this identification, the per-example loss is

$$\ell(\mu; (z, y)) = \|z - \mu_y\|_2^2.$$

Fix a class index $c \in \{1, \ldots, C\}$. The gradient of ℓ with respect to block μ_c is

$$\nabla_{\mu_c} \ell(\mu; (z, y)) = 2(\mu_y - z) \, \mathbf{1}\{c = y\}.$$

Now take two records (z, y) and (z', y') and subtract their blockwise gradients:

$$\begin{aligned} &\nabla_{\mu_c} \ell(\mu; (z, y)) - \nabla_{\mu_c} \ell(\mu; (z', y')) \\ &\quad = 2(\mu_y - z) \, \mathbf{1}\{c = y\} - 2(\mu_{y'} - z') \, \mathbf{1}\{c = y'\}. \end{aligned}$$

Under the label-preserving metric of Equation (2), finite record distance occurs only when $y = y'$. Therefore the only nonvacuous case is $y = y'$. In that case the display above simplifies to

$$\nabla_{\mu_c} \ell(\mu; (z, y)) - \nabla_{\mu_c} \ell(\mu; (z', y)) = 2(z' - z) \, \mathbf{1}\{c = y\}.$$

This means that all blocks are zero except the block indexed by y. The full gradient difference therefore has exactly one nonzero block, equal to $2(z' - z)$. Its Euclidean norm is thus

$$\|\nabla_\mu \ell(\mu; (z, y)) - \nabla_\mu \ell(\mu; (z', y))\|_2 = 2\|z - z'\|_2.$$

By the definition of the label-preserving metric,

$$d\big((z, y), (z', y)\big) = \|z - z'\|_2.$$

Substituting this identity into the previous display yields

$$\|\nabla_\mu \ell(\mu; (z, y)) - \nabla_\mu \ell(\mu; (z', y))\|_2 = 2\, d\big((z, y), (z', y)\big).$$

Hence Assumption 3.4 holds with $L_z = 2$.

To see that the constant is exact, choose any same-label pair with $z \neq z'$. The last display becomes an equality, so no smaller constant can work uniformly. □

E.5.2 Proof of Corollary A.2

Proof. Let

$$\widehat{\mu}(D) \in \arg\min_{\mu} F_D(\mu)$$

be the unique minimizer of the ridge-prototype objective. For each class c, define the class sum and class count

$$S_c(D) := \sum_{i: y_i = c} z_i, \qquad n_c(D) := \#\{i : y_i = c\}.$$

Because the objective separates over the class blocks, we can write

$$F_D(\mu) = \sum_{c=1}^{C} \left[\frac{1}{n} \sum_{i: y_i = c} \|z_i - \mu_c\|_2^2 + \frac{\lambda}{2} \|\mu_c\|_2^2 \right].$$

Fix a class c. Differentiating the cth summand with respect to μ_c gives

$$\nabla_{\mu_c} F_D(\mu) = \frac{2}{n} \sum_{i: y_i = c} (\mu_c - z_i) + \lambda \mu_c.$$

Expand the sum:

$$\nabla_{\mu_c} F_D(\mu) = \frac{2n_c(D)}{n} \mu_c - \frac{2}{n} S_c(D) + \lambda \mu_c.$$

Group the μ_c terms:

$$\nabla_{\mu_c} F_D(\mu) = \left(\frac{2n_c(D)}{n} + \lambda \right) \mu_c - \frac{2}{n} S_c(D).$$

At the optimum this gradient is zero, so

$$\left(\frac{2n_c(D)}{n} + \lambda \right) \widehat{\mu}_c(D) = \frac{2}{n} S_c(D).$$

Multiply both sides by n:

$$\big(2n_c(D) + \lambda n\big) \widehat{\mu}_c(D) = 2S_c(D).$$

Therefore the closed-form optimizer is

$$\widehat{\mu}_c(D) = \frac{2S_c(D)}{2n_c(D) + \lambda n}. \tag{4}$$

Now let D' be obtained from D by replacing one record (z, y) by (z', y) with $\|z - z'\|_2 \leq r$. Because the label is preserved, the class counts do not change:

$$n_c(D') = n_c(D) \qquad \text{for every class } c.$$

The class sums are also unchanged for all classes except the affected class y:

$$S_c(D') = S_c(D) \qquad \text{for } c \neq y,$$

while for the affected class,

$$S_y(D') = S_y(D) - z + z'.$$

Substituting these identities into (4) shows that

$$\widehat{\mu}_c(D') = \widehat{\mu}_c(D) \qquad \text{for every } c \neq y,$$

and

$$\widehat{\mu}_y(D') - \widehat{\mu}_y(D) = \frac{2\big(S_y(D') - S_y(D)\big)}{2n_y(D) + \lambda n} = \frac{2(z' - z)}{2n_y(D) + \lambda n}.$$

Thus the full parameter difference has exactly one nonzero block, namely the yth block. Its norm is therefore

$$\|\widehat{\mu}(D') - \widehat{\mu}(D)\|_2 = \|\widehat{\mu}_y(D') - \widehat{\mu}_y(D)\|_2 = \frac{2\|z' - z\|_2}{2n_y(D) + \lambda n}.$$

Since the changed record itself belongs to class y, we have $n_y(D) \geq 1$. Hence

$$\|\widehat{\mu}(D') - \widehat{\mu}(D)\|_2 \leq \frac{2r}{2n_y(D) + \lambda n} \leq \frac{2r}{2 + \lambda n}.$$

This proves the upper bound

$$\Delta_{\text{proto}}(r) \leq \frac{2r}{2 + \lambda n}.$$

To prove equality, choose a dataset in which the modified record is the *only* example from its class, so that $n_y(D) = 1$. Choose the replacement so that $\|z - z'\|_2 = r$. For this pair,

$$\|\widehat{\mu}(D') - \widehat{\mu}(D)\|_2 = \frac{2r}{2 + \lambda n}.$$

Therefore the upper bound is attained, and

$$\Delta_{\text{proto}}(r) = \frac{2r}{2 + \lambda n}.$$

□

E.5.3 Proof of Proposition A.3

Proof. Under the label-preserving metric, adjacent datasets differ by replacing (z, y) with (z', y), so the class labels and hence the count vector are unchanged along every adjacency edge. The proof of Corollary A.2 gives, for a replacement in class c,

$$\|\widehat{\mu}(D') - \widehat{\mu}(D)\|_2 = \frac{2\|z' - z\|_2}{2n_c + \lambda n}.$$

Conditional on the public count vector, the only remaining choices are the class c and the replacement distance. The radius policy imposes $\|z' - z\|_2 \leq r$, so the largest possible movement is

$$\max_{c : n_c > 0} \frac{2r}{2n_c + \lambda n}.$$

The denominator is smallest for a class with the smallest nonzero count $n_{\min}$, which gives $2r/(2n_{\min} + \lambda n)$. Equality is attained by choosing such a class and a replacement at distance exactly r. □

E.5.4 Proof of Theorem A.4

Proof. For each class $c \in \{1, \ldots, C\}$, define the class sum and class count in the known dataset D^- by

$$S_c^- := \sum_{\substack{(z', y') \in D^- \\ y' = c}} z',$$

$$n_c^- := \#\{(z', y') \in D^- : y' = c\}.$$

Since the full dataset is $D = D^- \cup \{(z, y)\}$, the full-data class statistics are

$$S_c(D) = S_c^- + z\,\mathbf{1}\{c = y\}\,, \qquad n_c(D) = n_c^- + \mathbf{1}\{c = y\}\,.$$

As shown in the proof of Corollary A.2, the released prototype satisfies

$$\widehat{\mu}_c(D) = \frac{2S_c(D)}{2n_c(D) + \lambda n}.$$

We now prove the two claims separately.

Part 1: exact record reconstruction when the label is known. Assume the adversary knows the missing label y. Then for the affected class,

$$S_y(D) = S_y^- + z, \qquad n_y(D) = n_y^- + 1.$$

Substituting these identities into the closed-form optimizer gives

$$\widehat{\mu}_y(D) = \frac{2(S_y^- + z)}{2(n_y^- + 1) + \lambda n}.$$

Multiply both sides by the denominator:

$$\big(2(n_y^- + 1) + \lambda n\big)\widehat{\mu}_y(D) = 2(S_y^- + z).$$

Expand the right-hand side:

$$\big(2(n_y^- + 1) + \lambda n\big)\widehat{\mu}_y(D) = 2S_y^- + 2z.$$

Subtract $2S_y^-$ from both sides:

$$\big(2(n_y^- + 1) + \lambda n\big)\widehat{\mu}_y(D) - 2S_y^- = 2z.$$

Finally divide by 2:

$$z = \frac{2(n_y^- + 1) + \lambda n}{2}\,\widehat{\mu}_y(D) - S_y^-.$$

Since $S_y^- = \sum_{(z',y') \in D^- :\, y' = y} z'$, this is exactly the displayed reconstruction formula.

Part 2: label identification from the changed prototype. For any class $c \neq y$, the missing record does not affect class c. Thus

$$S_c(D) = S_c^-, \qquad n_c(D) = n_c^-.$$

Substituting into the closed-form optimizer gives

$$\widehat{\mu}_c(D) = \frac{2S_c^-}{2n_c^- + \lambda n}.$$

Define the D^--implied baseline prototype

$$\mu_c^- := \frac{2S_c^-}{2n_c^- + \lambda n}.$$

Then we have shown that

$$\widehat{\mu}_c(D) = \mu_c^- \qquad \text{for every } c \neq y.$$

Therefore, if there is a unique class $c^\star$ such that

$$\widehat{\mu}_{c^\star}(D) \neq \mu_{c^\star}^-,$$

that class must be the true missing class: $c^\star = y$. Once the label is identified, Part 1 reconstructs the record exactly.

It remains to characterize the only case where the missing class fails to move. Suppose

$$\widehat{\mu}_y(D) = \mu_y^-.$$

Using the formulas above, this means

$$\frac{2(S_y^- + z)}{2(n_y^- + 1) + \lambda n} = \frac{2S_y^-}{2n_y^- + \lambda n}.$$

Cancel the common factor 2:

$$\frac{S_y^- + z}{2(n_y^- + 1) + \lambda n} = \frac{S_y^-}{2n_y^- + \lambda n}.$$

Cross-multiply:

$$(S_y^- + z)(2n_y^- + \lambda n) = S_y^-(2n_y^- + 2 + \lambda n).$$

Expand the left-hand side:

$$S_y^-(2n_y^- + \lambda n) + z(2n_y^- + \lambda n) = S_y^-(2n_y^- + \lambda n) + 2S_y^-.$$

Subtract the common term $S_y^-(2n_y^- + \lambda n)$ from both sides:

$$z(2n_y^- + \lambda n) = 2S_y^-.$$

Finally divide by $2n_y^- + \lambda n$ to get

$$z = \frac{2S_y^-}{2n_y^- + \lambda n} = \mu_y^-.$$

Thus,

$$\widehat{\mu}_y(D) = \mu_y^- \quad \Longrightarrow \quad z = \mu_y^-.$$

The converse is immediate: if $z = \mu_y^-$, then substituting this into the formula for $\widehat{\mu}_y(D)$ gives back μ_y^-. Hence

$$\widehat{\mu}_y(D) = \mu_y^- \quad \Longleftrightarrow \quad z = \mu_y^-.$$

So the only obstruction to label identification is the degenerate case in which the missing record equals the D^--implied prototype of its class. □

E.5.5 Proof of Corollary A.5

Proof. By the closed-form optimizer in (4), if the missing record is (Z, y) then

$$\widehat{\mu}_y(D^- \cup \{(Z, y)\}) = \frac{2(S_y^- + Z)}{A_y}, \qquad A_y = 2(n_y^- + 1) + \lambda n.$$

The Gaussian release gives

$$\widetilde{\mu}_y = \frac{2(S_y^- + Z)}{A_y} + \xi_y, \qquad \xi_y \sim \mathcal{N}(0, \sigma^2 I).$$

Multiplying by $A_y/2$ and subtracting S_y^- yields

$$W_y = \frac{A_y}{2}\widetilde{\mu}_y - S_y^- = Z + \frac{A_y}{2}\xi_y.$$

Thus $W_y = Z + \eta_y$ with $\eta_y \sim \mathcal{N}(0, (A_y\sigma/2)^2 I)$. For a uniform finite prior with common known label, the Gaussian likelihood is proportional to

$$\exp\left(-\frac{\|W_y - z_i\|_2^2}{2\tau_y^2}\right),$$

so MAP decoding is nearest-neighbor decoding over the candidate set. □

F PROOFS FOR BALL-RERO

F.1 PROOF OF THEOREM 4.5

Proof. For each record $z \in \mathcal{Z}$, define the completed dataset

$$D_z := D^- \cup \{z\},$$

and let

$$P_z := \mathrm{Law}\left(M(D_z)\right)$$

be the law of the mechanism output under D_z. Define the success indicator

$$A(z, \theta) := \mathbf{1}\{\rho(z, R(\theta, D^-)) \le \eta\}.$$

The reconstruction success probability is

$$p_{\mathrm{succ}} = \mathbb{E}_{Z\sim\pi_{u,r}}\mathbb{E}_{\theta\sim P_Z}[A(Z, \theta)].$$

Because $\pi_{u,r}$ is supported on $\mathcal{B}_d(u, r)$, every z in its support satisfies $d(z, u) \le r$. Hence D_z and D_u are r-adjacent. Since M is (ε, δ) r-adjacency DP, for every such z and every measurable event E,

$$P_z(E) \le e^{\varepsilon} P_u(E) + \delta.$$

Apply this inequality to

$$E_z := \{\theta : \rho(z, R(\theta, D^-)) \le \eta\}.$$

Then

$$\mathbb{E}_{\theta\sim P_z}[A(z, \theta)] = P_z(E_z) \le e^{\varepsilon} P_u(E_z) + \delta.$$

Averaging over $Z \sim \pi_{u,r}$ gives

$$p_{\mathrm{succ}} \le e^{\varepsilon}\mathbb{E}_{Z\sim\pi_{u,r}}\mathbb{E}_{\theta\sim P_u}[A(Z, \theta)] + \delta.$$

By Tonelli's theorem (Folland, 1999, Chapter 2),

$$p_{\mathrm{succ}} \le e^{\varepsilon}\mathbb{E}_{\theta\sim P_u} \Pr_{Z\sim\pi_{u,r}}[\rho(Z, R(\theta, D^-)) \le \eta] + \delta.$$

For each fixed θ, the output constraint gives $R(\theta, D^-) \in \mathcal{A}_{u,r}$. Therefore

$$\begin{aligned}&\Pr_{Z\sim\pi_{u,r}}[\rho(Z, R(\theta, D^-)) \le \eta] \\ &\le \sup_{a\in\mathcal{A}_{u,r}} \Pr_{Z\sim\pi_{u,r}}[\rho(Z, a) \le \eta] = \kappa.\end{aligned}$$

Substituting this bound yields

$$p_{\mathrm{succ}} \le e^{\varepsilon}\kappa + \delta,$$

which is the claimed Ball-ReRo guarantee. □

F.2 BALL-RDP EXTENSION

For completeness, we adapt Rényi-DP (RDP) (Mironov, 2017) to the radius-restricted adjacency relation and state the resulting reconstruction certificate.

Definition F.1 (Ball-RDP)**.** Fix an order $\alpha > 1$. A mechanism M satisfies $(\alpha, \varepsilon_\alpha)$ *Ball-RDP at radius* r if, for every $D \sim_r D'$,

$$D_\alpha\big(M(D) \,\|\, M(D')\big) \le \varepsilon_\alpha.$$

Lemma F.2 (Gaussian Ball-RDP)**.** *Let $M(D) = f(D) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I)$, and let $\Delta_2(r) := \sup_{D\sim_r D'} \|f(D) - f(D')\|_2$. Then, for every $\alpha > 1$ and $D \sim_r D'$,*

$$D_\alpha\big(M(D) \,\|\, M(D')\big) \le \frac{\alpha\Delta_2(r)^2}{2\sigma^2}.$$

Theorem F.3 (Ball-RDP implies Ball-ReRo)**.** *Fix $D^- \in \mathcal{Z}^{n-1}$, a Ball-supported prior $\pi_{u,r}$, a loss ρ, a tolerance η, and a feasible output set $\mathcal{A}_{u,r} \subseteq \mathcal{Z}$. Let $\kappa := \kappa_{\pi_{u,r},\rho}(\eta; \mathcal{A}_{u,r})$. If M satisfies $(\alpha, \varepsilon_\alpha)$ Ball-RDP at radius r for some $\alpha > 1$, then every reconstructor $R : \Theta \times \mathcal{Z}^{n-1} \to \mathcal{A}_{u,r}$ has success probability at most*

$$\gamma_\alpha := \min\left\{1, \kappa^{(\alpha-1)/\alpha} \exp\left(\frac{\alpha-1}{\alpha}\varepsilon_\alpha\right)\right\}.$$

F.3 Proof of Lemma F.2

Proof. Fix a pair of r-adjacent datasets $D \sim_r D'$. Write

$$\mu_0 := f(D), \qquad \mu_1 := f(D').$$

Then the two released distributions are

$$M(D) \sim \mathcal{N}(\mu_0, \sigma^2 I), \qquad M(D') \sim \mathcal{N}(\mu_1, \sigma^2 I).$$

For two Gaussians with the same covariance matrix, the order-α Rényi divergence is

$$D_\alpha\big(\mathcal{N}(\mu_0, \sigma^2 I) \,\|\, \mathcal{N}(\mu_1, \sigma^2 I)\big) = \frac{\alpha}{2\sigma^2}\|\mu_0 - \mu_1\|_2^2.$$

Applying this identity gives

$$D_\alpha\big(M(D) \,\|\, M(D')\big) = \frac{\alpha}{2\sigma^2}\|f(D) - f(D')\|_2^2.$$

By the definition of r-adjacency sensitivity,

$$\|f(D) - f(D')\|_2 \le \Delta_2(r).$$

Substituting this into the previous display proves

$$D_\alpha\big(M(D) \,\|\, M(D')\big) \le \frac{\alpha\, \Delta_2(r)^2}{2\sigma^2}.$$

□

F.4 Proof of Theorem F.3

Proof. For each candidate record z, let

$$D_z := D^- \cup \{z\}, \qquad P_z := \mathrm{Law}\,\big(M(D_z)\big).$$

Define

$$A(z, \theta) := \mathbf{1}\big\{\rho(z, R(\theta, D^-)) \le \eta\big\}.$$

The reconstruction success probability is

$$p_{\mathrm{succ}} = \int_{\mathcal{Z}} \int_{\Theta} A(z, \theta)\, P_z(d\theta)\, \pi_{u,r}(dz).$$

Because $\pi_{u,r}$ is supported on $\mathcal{B}_d(u, r)$, every z in its support satisfies $D_z \sim_r D_u$. By Ball-RDP,

$$D_\alpha(P_z \| P_u) \le \varepsilon_\alpha.$$

In particular, P_z is absolutely continuous with respect to P_u for all relevant z. Let

$$\Lambda_z(\theta) := \frac{dP_z}{dP_u}(\theta).$$

Rewrite p_{succ} using P_u as the reference measure:

$$p_{\mathrm{succ}} = \int_{\Theta} \left[\int_{\mathcal{Z}} A(z, \theta) \Lambda_z(\theta)\, \pi_{u,r}(dz) \right] P_u(d\theta).$$

Fix θ and apply Hölder's inequality with conjugate exponents α and $\alpha' = \alpha/(\alpha - 1)$:

$$\begin{aligned} &\int_{\mathcal{Z}} A(z, \theta) \Lambda_z(\theta)\, \pi_{u,r}(dz) \\ &\quad \le \left(\int_{\mathcal{Z}} A(z, \theta)\, \pi_{u,r}(dz) \right)^{1/\alpha'} \left(\int_{\mathcal{Z}} \Lambda_z(\theta)^\alpha\, \pi_{u,r}(dz) \right)^{1/\alpha}. \end{aligned}$$

For fixed θ, the output constraint gives $R(\theta, D^-) \in \mathcal{A}_{u,r}$, so the first factor is bounded by

$$\int_{\mathcal{Z}} A(z, \theta)\, \pi_{u,r}(dz) = \Pr_{Z \sim \pi_{u,r}}[\rho(Z, R(\theta, D^-)) \le \eta] \le \kappa.$$

Therefore

$$p_{\mathrm{succ}} \le \kappa^{1/\alpha'} \int_\Theta \left(\int_{\mathcal{Z}} \Lambda_z(\theta)^\alpha \, \pi_{u,r}(dz) \right)^{1/\alpha} P_u(d\theta).$$

Let

$$H(\theta) := \int_{\mathcal{Z}} \Lambda_z(\theta)^\alpha \, \pi_{u,r}(dz).$$

Since $x \mapsto x^{1/\alpha}$ is concave, Jensen's inequality gives

$$\left(\int_\Theta H(\theta)^{1/\alpha} P_u(d\theta) \right)^\alpha \le \int_\Theta H(\theta) P_u(d\theta).$$

Consequently,

$$\left(\frac{p_{\mathrm{succ}}}{\kappa^{1/\alpha'}} \right)^\alpha \le \int_{\mathcal{Z}} \left[\int_\Theta \Lambda_z(\theta)^\alpha P_u(d\theta) \right] \pi_{u,r}(dz).$$

By the definition of Rényi divergence,

$$\int_\Theta \Lambda_z(\theta)^\alpha P_u(d\theta) = \exp\big((\alpha-1) D_\alpha(P_z \| P_u)\big) \le e^{(\alpha-1)\varepsilon_\alpha}.$$

Thus

$$\left(\frac{p_{\mathrm{succ}}}{\kappa^{1/\alpha'}} \right)^\alpha \le e^{(\alpha-1)\varepsilon_\alpha}.$$

Taking the αth root and using $1/\alpha' = (\alpha-1)/\alpha$ yields

$$p_{\mathrm{succ}} \le \left(\kappa e^{\varepsilon_\alpha} \right)^{(\alpha-1)/\alpha}.$$

Since a probability is at most one, the claimed minimum with 1 follows. □

F.5 A Gaussian testing lemma

Lemma F.4 (Blow-up function for equal-covariance Gaussians)**.** *Let*

$$P = \mathcal{N}(\mu_1, \sigma^2 I), \qquad Q = \mathcal{N}(\mu_0, \sigma^2 I),$$
$$\Delta := \|\mu_1 - \mu_0\|_2.$$

Then for every $\kappa \in [0,1]$*,*

$$\sup\{P(E) : Q(E) \le \kappa\} = \Phi\left(\Phi^{-1}(\kappa) + \frac{\Delta}{\sigma} \right),$$

where the values at $\kappa = 0$ *and* $\kappa = 1$ *are understood as the continuous limits, namely* 0 *and* 1*, respectively.*

Proof. By translation and orthogonal invariance of Gaussian measures, it is enough to consider

$$Q = \mathcal{N}(0, \sigma^2 I), \qquad P = \mathcal{N}(\Delta e_1, \sigma^2 I),$$

where e_1 is the first standard basis vector. The log-likelihood ratio is

$$\log \frac{dP}{dQ}(x) = \frac{\Delta}{\sigma^2} x_1 - \frac{\Delta^2}{2\sigma^2},$$

which is increasing in x_1. By the Neyman–Pearson lemma (Neyman & Pearson, 1933), among all measurable events with a fixed Q-mass, an event that maximizes $P(E)$ is a likelihood-ratio upper level set. Hence, in this Gaussian shift case, an optimal event at fixed Q-mass is a half-space

$$E_t := \{x : x_1 \ge t\}.$$

Choose t such that $Q(E_t) = \kappa$. Under Q, $x_1 \sim \mathcal{N}(0, \sigma^2)$, so

$$\kappa = 1 - \Phi(t/\sigma) \quad \text{and} \quad t = \sigma \Phi^{-1}(1-\kappa).$$

Under P, $x_1 \sim \mathcal{N}(\Delta, \sigma^2)$, hence

$$P(E_t) = 1 - \Phi\left(\frac{t-\Delta}{\sigma} \right) = \Phi\left(\Phi^{-1}(\kappa) + \frac{\Delta}{\sigma} \right).$$

The Neyman–Pearson lemma gives optimality, proving the identity. □

F.6 Proof of Theorem 4.6

Proof. Fix $D^- \in \mathcal{Z}^{n-1}$ and an anchor $u \in \mathcal{Z}$. For each z in the support of $\pi_{u,r}$, define

$$D_z := D^- \cup \{z\}, \qquad D_u := D^- \cup \{u\}.$$

Let

$$P_z := \mathrm{Law}(M(D_z)), \qquad Q_u := \mathrm{Law}(M(D_u)).$$

Because the mechanism is Gaussian output perturbation,

$$P_z = \mathcal{N}(f(D_z), \sigma^2 I), \qquad Q_u = \mathcal{N}(f(D_u), \sigma^2 I).$$

Set

$$\Delta_z := \|f(D_z) - f(D_u)\|_2, \qquad \Delta_\star := \bar{\Delta}(D^-).$$

By definition, $\Delta_z \le \Delta_\star$ for all z in the support. Since $d(z,u) \le r$, also $D_z \sim_r D_u$, and hence $\Delta_\star \le \Delta_2(r)$.

For each z, define

$$E_z := \{\theta : \rho(z, R(\theta, D^-)) \le \eta\}.$$

The attack success probability is

$$p_{\mathrm{succ}} = \mathbb{E}_{Z\sim\pi_{u,r}}[P_Z(E_Z)].$$

By Lemma F.4,

$$\begin{aligned} P_z(E_z) &\le \Phi\left(\Phi^{-1}(Q_u(E_z)) + \frac{\Delta_z}{\sigma}\right) \\ &\le \Phi\left(\Phi^{-1}(Q_u(E_z)) + \frac{\Delta_\star}{\sigma}\right). \end{aligned}$$

Therefore

$$p_{\mathrm{succ}} \le \mathbb{E}_{Z\sim\pi_{u,r}}\left[\Phi\left(\Phi^{-1}(Q_u(E_Z)) + \frac{\Delta_\star}{\sigma}\right)\right].$$

Let

$$a := \frac{\Delta_\star}{\sigma}, \qquad g(t) := \Phi\big(\Phi^{-1}(t) + a\big).$$

For $t \in (0,1)$, with $x = \Phi^{-1}(t)$ and φ denoting the standard normal density,

$$g'(t) = \frac{\varphi(x+a)}{\varphi(x)} = \exp\left(-ax - \frac{a^2}{2}\right) > 0, \qquad g''(t) = -\frac{a}{\varphi(x)}\exp\left(-ax - \frac{a^2}{2}\right) \le 0.$$

Thus g is nondecreasing and concave on $(0,1)$, and the same conclusion holds on $[0,1]$ by continuous extension. Jensen's inequality gives

$$p_{\mathrm{succ}} \le \Phi\left(\Phi^{-1}\big(\mathbb{E}_{Z\sim\pi_{u,r}}[Q_u(E_Z)]\big) + \frac{\Delta_\star}{\sigma}\right).$$

It remains to bound the expectation inside Φ^{-1}. By Tonelli's theorem (Folland, 1999, Chapter 2),

$$\mathbb{E}_{Z\sim\pi_{u,r}}[Q_u(E_Z)] = \mathbb{E}_{\theta\sim Q_u} \Pr_{Z\sim\pi_{u,r}}[\rho(Z, R(\theta, D^-)) \le \eta].$$

For each fixed θ, the output constraint gives $R(\theta, D^-) \in \mathcal{A}_{u,r}$, so

$$\Pr_{Z\sim\pi_{u,r}}[\rho(Z, R(\theta, D^-)) \le \eta] \le \kappa_{\pi_{u,r},\rho}(\eta; \mathcal{A}_{u,r}) = \kappa.$$

Thus

$$\mathbb{E}_{Z\sim\pi_{u,r}}[Q_u(E_Z)] \le \kappa.$$

Using the monotonicity of g yields

$$p_{\mathrm{succ}} \le \Phi\left(\Phi^{-1}(\kappa) + \frac{\Delta_\star}{\sigma}\right).$$

Finally, $\Delta_\star \le \Delta_2(r)$ gives the second displayed inequality in the theorem. □

F.7 Proof of Corollary 4.7

Proof. Under the exact-identification loss $\rho_{\mathrm{ID}}(z, z') = \mathbf{1}\{z \neq z'\}$ with tolerance $\eta = 0$, the success event is exactly $R(\theta, D^-) = Z$. The allowed output set is $\mathcal{A}_{u,r} = S$. For the uniform prior

$$\pi_S = \frac{1}{m} \sum_{i=1}^{m} \delta_{z_i},$$

the best oblivious feasible-set strategy is to guess any one candidate, so

$$\kappa_{\pi_S, \rho_{\mathrm{ID}}}(0; S) = \sup_{a \in S} \Pr_{Z \sim \pi_S}[Z = a] = \frac{1}{m}.$$

Substituting this value into Theorem 4.6 and then using $\bar{\Delta}(D^-) \leq \Delta_2(r)$ gives

$$\Pr[R(\theta, D^-) = Z] \leq \Phi\left(\Phi^{-1}(1/m) + \frac{\Delta_2(r)}{\sigma}\right),$$

which is the claimed exact-identification bound. □

F.8 Support-specific finite-Gaussian exact-identification bound

The following bound is a sharper finite-prior statement for equal-covariance Gaussian releases. It is useful as a technical comparison point, but it is not the quantity plotted in the experiments; the main text uses the simpler direct certificate from Corollary 4.7.

Theorem F.5 (Support-specific finite-Gaussian exact-ID bound)**.** *Let $P_i = \mathcal{N}(\mu_i, \sigma^2 I)$ be the release law under candidate z_i, with prior weights $\pi_i > 0$ and $\sum_i \pi_i = 1$. Fix any reference Gaussian $Q = \mathcal{N}(\mu_0, \sigma^2 I)$ and define*

$$d_i := \frac{\|\mu_i - \mu_0\|_2}{\sigma}.$$

Then every exact-identification rule satisfies

$$\Pr[\widehat{Z} = Z] \leq \max_{q_1, \ldots, q_m \geq 0:\, \sum_i q_i \leq 1} \sum_{i=1}^{m} \pi_i \, \Phi\big(\Phi^{-1}(q_i) + d_i\big),$$

with boundary cases interpreted by continuity. In particular, if $\pi_i = 1/m$ and $d_i \leq R$ for all i, then

$$\Pr[\widehat{Z} = Z] \leq \Phi\big(\Phi^{-1}(1/m) + R\big).$$

Proof. Let A_i be the event that the attack outputs candidate z_i. The events $A_1, \ldots, A_m$ are disjoint, and therefore

$$\sum_{i=1}^{m} Q(A_i) \leq 1.$$

Write $q_i := Q(A_i)$. By Lemma F.4, applied to $P_i = \mathcal{N}(\mu_i, \sigma^2 I)$ and $Q = \mathcal{N}(\mu_0, \sigma^2 I)$,

$$P_i(A_i) \leq \Phi\big(\Phi^{-1}(q_i) + d_i\big), \qquad d_i = \frac{\|\mu_i - \mu_0\|_2}{\sigma}.$$

The exact-identification success probability is

$$p_{\mathrm{succ}} = \sum_{i=1}^{m} \pi_i P_i(A_i).$$

Combining the previous two displays gives

$$p_{\mathrm{succ}} \leq \sum_{i=1}^{m} \pi_i \Phi\big(\Phi^{-1}(q_i) + d_i\big)$$

for some nonnegative q_i with $\sum_i q_i \leq 1$. Maximizing over all such q_i proves the first claim.

For the uniform-prior specialization, assume $d_i \le R$ for all i and define

$$h_R(q) := \Phi\big(\Phi^{-1}(q) + R\big).$$

For $q \in (0,1)$, with $x = \Phi^{-1}(q)$ and φ denoting the standard normal density,

$$h_R'(q) = \frac{\varphi(x+R)}{\varphi(x)} = \exp\left(-Rx - \frac{R^2}{2}\right) > 0, \qquad h_R''(q) = -\frac{R}{\varphi(x)} \exp\left(-Rx - \frac{R^2}{2}\right) \le 0.$$

Thus h_R is nondecreasing and concave on $(0,1)$, and the same conclusion holds on $[0,1]$ by continuous extension. Jensen's inequality gives

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} \Phi\big(\Phi^{-1}(q_i) + d_i\big) &\le \frac{1}{m}\sum_{i=1}^{m} h_R(q_i) \\ &\le h_R\left(\frac{1}{m}\sum_i q_i\right) \\ &= \Phi\left(\Phi^{-1}\left(\frac{1}{m}\sum_i q_i\right) + R\right).\end{aligned}$$

Because $\sum_i q_i \le 1$ and the same map is nondecreasing,

$$p_{\mathrm{succ}} \le \Phi\big(\Phi^{-1}(1/m) + R\big).$$

This proves the stated finite-Gaussian support bound. □

G Proof of Proposition 4.1

Proof. Fix a finite candidate set $S = \{z_1, \dots, z_m\}$ and prior weights $\pi_1, \dots, \pi_m$. For each candidate z_i, the Gaussian release density is

$$p(\widetilde{\theta} \mid z_i, D^-) = (2\pi\sigma^2)^{-p/2} \exp\left(-\frac{1}{2\sigma^2}\big\|\widetilde{\theta} - \widehat{\theta}(D^- \cup \{z_i\})\big\|_2^2\right),$$

where p is the parameter dimension. By Bayes' rule, for each $i = 1, \dots, m$,

$$p(z_i \mid \widetilde{\theta}, D^-) = \frac{\pi_i\, p(\widetilde{\theta} \mid z_i, D^-)}{\sum_{j=1}^m \pi_j\, p(\widetilde{\theta} \mid z_j, D^-)} = \frac{\pi_i \exp\left(-\frac{1}{2\sigma^2}\big\|\widetilde{\theta} - \widehat{\theta}(D^- \cup \{z_i\})\big\|_2^2\right)}{\sum_{j=1}^m \pi_j \exp\left(-\frac{1}{2\sigma^2}\big\|\widetilde{\theta} - \widehat{\theta}(D^- \cup \{z_j\})\big\|_2^2\right)}.$$

Take logarithms:

$$\log p(z_i \mid \widetilde{\theta}, D^-) = \log \pi_i - \frac{1}{2\sigma^2}\big\|\widetilde{\theta} - \widehat{\theta}(D^- \cup \{z_i\})\big\|_2^2 + C(\widetilde{\theta}),$$

where $C(\widetilde{\theta})$ is a constant independent of the candidate index i. Since the MAP rule chooses the candidate with the largest posterior probability, and since $C(\widetilde{\theta})$ is the same for every candidate, any MAP estimator is of the form

$$\arg\max_{z_i \in S} \left\{\log \pi_i - \frac{1}{2\sigma^2}\big\|\widetilde{\theta} - \widehat{\theta}(D^- \cup \{z_i\})\big\|_2^2\right\}.$$

If the prior is uniform, then $\log \pi_i$ is the same for every candidate and can be dropped, leaving the least-squares objective. □

H Derivations for Secondary Convex Head Constants

Proof of Proposition A.6. We prove the three constants separately. Throughout this proof, the record metric is the label-preserving metric used in the statement. Thus finite-distance substitutions have the same label, and for a same-label pair $(z, y), (z', y)$ the distance is $\|z - z'\|_2$. Different-label pairs have infinite distance, so they impose no additional finite Lipschitz constraint in Assumption 3.4.

The derivations in the three subsections below establish all bounded-parameter and ERM-minimizer constants in Proposition A.6. Substituting any of them into Theorem 3.5 gives the stated radius-r sensitivity bound. □

H.1 SQUARED HINGE

Fix a label $y \in \{-1, 1\}$ and a parameter vector $w \in \mathbb{R}^d$. For a feature vector z, define

$$q(z) := 1 - yw^\top z, \qquad h(z) := q(z)_+ = \max\{0, q(z)\}.$$

The squared-hinge loss is

$$\ell(w; z, y) = h(z)^2.$$

The scalar map $u \mapsto (u_+)^2$ is differentiable everywhere: its derivative is 0 for $u < 0$, is $2u$ for $u > 0$, and is also 0 at $u = 0$ because the one-sided derivatives both equal 0. Hence, by the chain rule,

$$\nabla_w \ell(w; z, y) = 2h(z)\nabla_w(1 - yw^\top z) = 2h(z)(-yz) = -2yzh(z).$$

Write

$$g(z) := \nabla_w \ell(w; z, y) = -2yzh(z).$$

We first bound the scalar hinge difference. The map $m(a) := a_+ = \max\{0, a\}$ is 1-Lipschitz on $\mathbb{R}$. Indeed, if $a, b \le 0$, then $m(a) = m(b) = 0$; if $a, b \ge 0$, then $|m(a) - m(b)| = |a - b|$; and if $a \ge 0 \ge b$, then $|m(a) - m(b)| = a \le a - b = |a - b|$, with the remaining mixed case symmetric. Therefore,

$$\begin{aligned} |h(z) - h(z')| &\le |q(z) - q(z')| \\ &= |(1 - yw^\top z) - (1 - yw^\top z')| \\ &= |yw^\top (z' - z)| \\ &= |w^\top (z' - z)| \\ &\le \|w\|_2 \, \|z' - z\|_2 = \|w\|_2 \, \|z - z'\|_2. \end{aligned}$$

Next, for any admissible feature vector u with $\|u\|_2 \le B$,

$$\begin{aligned} h(u) &= \max\{0, 1 - yw^\top u\} \\ &\le 1 + |w^\top u| \\ &\le 1 + \|w\|_2 \|u\|_2 \\ &\le 1 + B\|w\|_2. \end{aligned}$$

In particular, $h(z') \le 1 + B\|w\|_2$.

Now subtract the gradients and add and subtract $zh(z')$:

$$\begin{aligned} g(z) - g(z') &= -2yzh(z) + 2yz'h(z') \\ &= -2y\big(zh(z) - z'h(z')\big) \\ &= -2y\big(z(h(z) - h(z')) + (z - z')h(z')\big). \end{aligned}$$

Taking Euclidean norms and using $|y| = 1$ gives

$$\begin{aligned} \|g(z) - g(z')\|_2 &\le 2\Big(\|z\|_2 |h(z) - h(z')| + \|z - z'\|_2 |h(z')|\Big) \\ &\le 2\Big(B\|w\|_2 \|z - z'\|_2 + (1 + B\|w\|_2)\|z - z'\|_2\Big) \\ &= \big(2 + 4B\|w\|_2\big)\|z - z'\|_2. \end{aligned}$$

Thus, on $\Theta_R = \{w : \|w\|_2 \le R\}$,

$$\|\nabla_w \ell(w; z, y) - \nabla_w \ell(w; z', y)\|_2 \le (2 + 4BR)\|z - z'\|_2,$$

so $L_z(R) = 2 + 4BR$ is valid.

It remains to bound the radius of ERM minimizers. Let

$$F_D(w) = \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i w^\top z_i\}^2 + \frac{\lambda}{2}\|w\|_2^2,$$

and let $w^\star \in \arg\min_w F_D(w)$. Since $w^\star$ minimizes F_D,

$$F_D(w^\star) \le F_D(0).$$

At $w = 0$, every hinge value is $\max\{0, 1\} = 1$, so

$$F_D(0) = \frac{1}{n}\sum_{i=1}^{n} 1^2 + \frac{\lambda}{2}\|0\|_2^2 = 1.$$

The empirical squared-hinge term is nonnegative, hence

$$F_D(w^\star) = \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(w^\star)^\top z_i\}^2 + \frac{\lambda}{2}\|w^\star\|_2^2 \geq \frac{\lambda}{2}\|w^\star\|_2^2.$$

Combining the last three displays,

$$\frac{\lambda}{2}\|w^\star\|_2^2 \leq F_D(w^\star) \leq F_D(0) = 1,$$

so

$$\|w^\star\|_2 \leq \sqrt{\frac{2}{\lambda}}.$$

Substituting $R = \sqrt{2/\lambda}$ into $L_z(R) = 2 + 4BR$ gives the ERM-minimizer constant

$$L_z \leq 2 + 4B\sqrt{\frac{2}{\lambda}}.$$

H.2 Binary Logistic

Let $y \in \{-1, 1\}$, let

$$\widetilde{z} := \begin{pmatrix} z \\ 1 \end{pmatrix} \in \mathbb{R}^{d+1}, \qquad \widetilde{\theta} := \begin{pmatrix} w \\ b \end{pmatrix} \in \mathbb{R}^{d+1}, \qquad \widetilde{B} := \sqrt{B^2 + 1}.$$

Then $\|\widetilde{z}\|_2 \leq \widetilde{B}$, and for same-label pairs,

$$\|\widetilde{z} - \widetilde{z}'\|_2 = \left\| \begin{pmatrix} z - z' \\ 0 \end{pmatrix} \right\|_2 = \|z - z'\|_2.$$

Define the score $s(\widetilde{z}) := \widetilde{\theta}^\top \widetilde{z}$ and the sigmoid $\sigma(u) := (1 + e^{-u})^{-1}$. The binary logistic loss is

$$\ell(\widetilde{\theta}; \widetilde{z}, y) = \log\big(1 + \exp(-y s(\widetilde{z}))\big).$$

With $u = -y s(\widetilde{z})$, we have

$$\frac{d}{du}\log(1 + e^u) = \frac{e^u}{1 + e^u} = \sigma(u), \qquad \nabla_{\widetilde{\theta}} u = -y\widetilde{z}.$$

Therefore,

$$\nabla_{\widetilde{\theta}} \ell(\widetilde{\theta}; \widetilde{z}, y) = -y\sigma(-y\widetilde{\theta}^\top \widetilde{z})\widetilde{z}.$$

For two same-label examples, set

$$\alpha := \sigma(-y\widetilde{\theta}^\top \widetilde{z}), \qquad \beta := \sigma(-y\widetilde{\theta}^\top \widetilde{z}').$$

Since $0 \leq \sigma \leq 1$, $0 \leq \alpha, \beta \leq 1$. The gradient difference satisfies

$$\begin{aligned} &\left\| \nabla_{\widetilde{\theta}} \ell(\widetilde{\theta}; \widetilde{z}, y) - \nabla_{\widetilde{\theta}} \ell(\widetilde{\theta}; \widetilde{z}', y) \right\|_2 \\ &\quad = \|-y\alpha\widetilde{z} + y\beta\widetilde{z}'\|_2 \\ &\quad = \|-\alpha\widetilde{z} + \beta\widetilde{z}'\|_2 . \end{aligned}$$

Add and subtract $\alpha\widetilde{z}'$:

$$-\alpha\widetilde{z} + \beta\widetilde{z}' = \alpha(\widetilde{z}' - \widetilde{z}) + (\beta - \alpha)\widetilde{z}'.$$

Thus,

$$\begin{aligned} \|-\alpha\widetilde{z} + \beta\widetilde{z}'\|_2 &\leq |\alpha| \, \|\widetilde{z}' - \widetilde{z}\|_2 + |\beta - \alpha| \, \|\widetilde{z}'\|_2 \\ &\leq \|z - z'\|_2 + \widetilde{B}|\beta - \alpha|. \end{aligned}$$

The sigmoid derivative is

$$\sigma'(u) = \sigma(u)(1 - \sigma(u)).$$

For $t \in [0, 1]$,

$$t(1-t) = \frac{1}{4} - \left(t - \frac{1}{2}\right)^2 \leq \frac{1}{4},$$

so $0 \leq \sigma'(u) \leq 1/4$ and σ is $1/4$-Lipschitz. Therefore,

$$\begin{aligned}
|\beta - \alpha| &\leq \frac{1}{4}\left|-y\widetilde{\theta}^\top \widetilde{z}' + y\widetilde{\theta}^\top \widetilde{z}\right| \\
&= \frac{1}{4}\left|\widetilde{\theta}^\top(\widetilde{z} - \widetilde{z}')\right| \\
&\leq \frac{1}{4}\|\widetilde{\theta}\|_2 \, \|\widetilde{z} - \widetilde{z}'\|_2 \\
&= \frac{1}{4}\|\widetilde{\theta}\|_2 \, \|z - z'\|_2.
\end{aligned}$$

Substituting this into the gradient-difference bound gives

$$\left\|\nabla_{\widetilde{\theta}}\ell(\widetilde{\theta}; \widetilde{z}, y) - \nabla_{\widetilde{\theta}}\ell(\widetilde{\theta}; \widetilde{z}', y)\right\|_2 \leq \left(1 + \frac{\widetilde{B}\|\widetilde{\theta}\|_2}{4}\right)\|z - z'\|_2.$$

Hence, on $\Theta_R = \{\widetilde{\theta} : \|\widetilde{\theta}\|_2 \leq R\}$,

$$L_z(R) = 1 + \frac{\widetilde{B}R}{4}$$

is valid.

Now consider the regularized objective, with both weights and intercept regularized,

$$F_D(\widetilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\widetilde{\theta}; \widetilde{z}_i, y_i) + \frac{\lambda}{2}\|\widetilde{\theta}\|_2^2.$$

Let $\widetilde{\theta}^\star$ be an ERM minimizer. First-order optimality gives

$$0 = \nabla F_D(\widetilde{\theta}^\star) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\widetilde{\theta}}\ell(\widetilde{\theta}^\star; \widetilde{z}_i, y_i) + \lambda\widetilde{\theta}^\star,$$

so

$$\lambda\widetilde{\theta}^\star = -\frac{1}{n}\sum_{i=1}^{n} \nabla_{\widetilde{\theta}}\ell(\widetilde{\theta}^\star; \widetilde{z}_i, y_i).$$

Taking norms and using the triangle inequality,

$$\lambda\|\widetilde{\theta}^\star\|_2 \leq \frac{1}{n}\sum_{i=1}^{n} \left\|\nabla_{\widetilde{\theta}}\ell(\widetilde{\theta}^\star; \widetilde{z}_i, y_i)\right\|_2.$$

For each example,

$$\begin{aligned}
\left\|\nabla_{\widetilde{\theta}}\ell(\widetilde{\theta}^\star; \widetilde{z}_i, y_i)\right\|_2 &= \left\|-y_i\sigma(-y_i(\widetilde{\theta}^\star)^\top \widetilde{z}_i)\widetilde{z}_i\right\|_2 \\
&\leq |y_i| \cdot 1 \cdot \|\widetilde{z}_i\|_2 \\
&\leq \widetilde{B}.
\end{aligned}$$

Therefore,

$$\lambda\|\widetilde{\theta}^\star\|_2 \leq \frac{1}{n}\sum_{i=1}^{n} \widetilde{B} = \widetilde{B},$$

and hence

$$\|\widetilde{\theta}^\star\|_2 \leq \frac{\widetilde{B}}{\lambda}.$$

Substituting $R = \widetilde{B}/\lambda$ into the fixed-parameter constant gives

$$L_z \leq 1 + \frac{\widetilde{B}}{4} \cdot \frac{\widetilde{B}}{\lambda} = 1 + \frac{\widetilde{B}^2}{4\lambda} = 1 + \frac{B^2 + 1}{4\lambda}.$$

H.3 SOFTMAX LOGISTIC

Let $y \in \{1, \dots, K\}$, let $\widetilde{z} = (z, 1) \in \mathbb{R}^{d+1}$, and let $\widetilde{W} \in \mathbb{R}^{K \times (d+1)}$. Define

$$p(\widetilde{z}) := \mathrm{softmax}(\widetilde{W}\widetilde{z}), \qquad a(\widetilde{z}) := p(\widetilde{z}) - e_y,$$

where e_y is the yth standard basis vector in $\mathbb{R}^K$. For scores $s := \widetilde{W}\widetilde{z}$, the softmax loss is

$$\ell(\widetilde{W}; \widetilde{z}, y) = -s_y + \log\left(\sum_{c=1}^{K} e^{s_c}\right).$$

For each class c,

$$\frac{\partial \ell}{\partial s_c} = -\mathbf{1}\{c = y\} + \frac{e^{s_c}}{\sum_{j=1}^{K} e^{s_j}} = p_c(\widetilde{z}) - (e_y)_c.$$

Thus $\nabla_s \ell = p(\widetilde{z}) - e_y = a(\widetilde{z})$. Since $s_c = \widetilde{w}_c^\top \widetilde{z}$ for the cth row $\widetilde{w}_c^\top$ of $\widetilde{W}$, the row-wise chain rule gives

$$\nabla_{\widetilde{W}} \ell(\widetilde{W}; \widetilde{z}, y) = a(\widetilde{z})\widetilde{z}^\top.$$

For a same-label pair, write

$$a := a(\widetilde{z}), \qquad a' := a(\widetilde{z}').$$

Subtract the gradients and add and subtract $a(\widetilde{z}')^\top$:

$$\begin{aligned}
&\nabla_{\widetilde{W}} \ell(\widetilde{W}; \widetilde{z}, y) - \nabla_{\widetilde{W}} \ell(\widetilde{W}; \widetilde{z}', y) \\
&\quad = a\widetilde{z}^\top - a'(\widetilde{z}')^\top \\
&\quad = a(\widetilde{z} - \widetilde{z}')^\top + (a - a')(\widetilde{z}')^\top.
\end{aligned}$$

Taking Frobenius norms and using the triangle inequality,

$$\begin{aligned}
&\left\| \nabla_{\widetilde{W}} \ell(\widetilde{W}; \widetilde{z}, y) - \nabla_{\widetilde{W}} \ell(\widetilde{W}; \widetilde{z}', y) \right\|_F \\
&\quad \le \left\| a(\widetilde{z} - \widetilde{z}')^\top \right\|_F + \left\| (a - a')(\widetilde{z}')^\top \right\|_F.
\end{aligned}$$

For vectors u, v, $\|uv^\top\|_F = \|u\|_2 \|v\|_2$, so

$$\begin{aligned}
&\left\| \nabla_{\widetilde{W}} \ell(\widetilde{W}; \widetilde{z}, y) - \nabla_{\widetilde{W}} \ell(\widetilde{W}; \widetilde{z}', y) \right\|_F \\
&\quad \le \|a\|_2 \, \|\widetilde{z} - \widetilde{z}'\|_2 + \|a - a'\|_2 \, \|\widetilde{z}'\|_2.
\end{aligned} \tag{5}$$

We bound the two factors separately.

First, since $p(\widetilde{z})$ is a probability vector,

$$\begin{aligned}
\|a\|_2^2 &= \|p(\widetilde{z}) - e_y\|_2^2 \\
&= (1 - p_y(\widetilde{z}))^2 + \sum_{c \ne y} p_c(\widetilde{z})^2 \\
&\le (1 - p_y(\widetilde{z}))^2 + \left(\sum_{c \ne y} p_c(\widetilde{z})\right)^2 \\
&= (1 - p_y(\widetilde{z}))^2 + (1 - p_y(\widetilde{z}))^2 \\
&\le 2.
\end{aligned}$$

Thus $\|a\|_2 \le \sqrt{2}$.

Second, because $a - a' = p(\widetilde{z}) - p(\widetilde{z}')$, we need a Lipschitz bound for softmax. For $p = \mathrm{softmax}(x)$, the Jacobian is

$$J(x) = \mathrm{Diag}(p) - pp^\top.$$

It is symmetric. In row i,

$$\begin{aligned}\sum_{j=1}^{K}|J_{ij}(x)| &= p_i(1-p_i)+\sum_{j\neq i}p_ip_j\\ &= p_i(1-p_i)+p_i\sum_{j\neq i}p_j\\ &= 2p_i(1-p_i)\\ &\leq \frac{1}{2}.\end{aligned}$$

Therefore $\|J(x)\|_\infty \leq 1/2$. Since $J(x)$ is symmetric, $\|J(x)\|_2 \leq \|J(x)\|_\infty \leq 1/2$. Applying the mean-value theorem to $x \mapsto \mathrm{softmax}(x)$ along the segment between $\widetilde{W}\widetilde{z}$ and $\widetilde{W}\widetilde{z}'$ gives

$$\begin{aligned}\|p(\widetilde{z})-p(\widetilde{z}')\|_2 &\leq \frac{1}{2}\|\widetilde{W}\widetilde{z}-\widetilde{W}\widetilde{z}'\|_2\\ &= \frac{1}{2}\|\widetilde{W}(\widetilde{z}-\widetilde{z}')\|_2\\ &\leq \frac{1}{2}\|\widetilde{W}\|_F\,\|\widetilde{z}-\widetilde{z}'\|_2.\end{aligned}$$

Hence

$$\|a-a'\|_2 \leq \frac{1}{2}\|\widetilde{W}\|_F\,\|\widetilde{z}-\widetilde{z}'\|_2.$$

Since $\|\widetilde{z}'\|_2 \leq \widetilde{B} = \sqrt{B^2+1}$ and $\|\widetilde{z}-\widetilde{z}'\|_2 = \|z-z'\|_2$, substituting the two bounds into (5) yields

$$\begin{aligned}&\left\|\nabla_{\widetilde{W}}\ell(\widetilde{W};\widetilde{z},y)-\nabla_{\widetilde{W}}\ell(\widetilde{W};\widetilde{z}',y)\right\|_F\\ &\quad\leq\left(\sqrt{2}+\frac{\widetilde{B}\|\widetilde{W}\|_F}{2}\right)\|z-z'\|_2.\end{aligned}$$

Therefore, on $\Theta_R = \{\widetilde{W} : \|\widetilde{W}\|_F \leq R\}$,

$$L_z(R) = \sqrt{2}+\frac{\widetilde{B}R}{2}$$

is valid.

It remains to bound the Frobenius norm of regularized ERM minimizers. Let

$$F_D(\widetilde{W}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\widetilde{W};\widetilde{z}_i,y_i)+\frac{\lambda}{2}\|\widetilde{W}\|_F^2,$$

and let $\widetilde{W}^\star$ minimize this objective. First-order optimality gives

$$0 = \frac{1}{n}\sum_{i=1}^{n}\nabla_{\widetilde{W}}\ell(\widetilde{W}^\star;\widetilde{z}_i,y_i)+\lambda\widetilde{W}^\star,$$

so

$$\lambda\widetilde{W}^\star = -\frac{1}{n}\sum_{i=1}^{n}\nabla_{\widetilde{W}}\ell(\widetilde{W}^\star;\widetilde{z}_i,y_i).$$

Taking Frobenius norms and applying the triangle inequality,

$$\lambda\|\widetilde{W}^\star\|_F \leq \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_{\widetilde{W}}\ell(\widetilde{W}^\star;\widetilde{z}_i,y_i)\right\|_F.$$

For one example, define $p_\star(\widetilde{z}_i) := \mathrm{softmax}(\widetilde{W}^\star\widetilde{z}_i)$. Then

$$\nabla_{\widetilde{W}}\ell(\widetilde{W}^\star;\widetilde{z}_i,y_i) = \big(p_\star(\widetilde{z}_i)-e_{y_i}\big)\widetilde{z}_i^\top,$$

and hence, by the same probability-vector bound used above,

$$\left\|\nabla_{\widetilde{W}}\ell(\widetilde{W}^\star;\widetilde{z}_i,y_i)\right\|_F = \|p_\star(\widetilde{z}_i)-e_{y_i}\|_2\,\|\widetilde{z}_i\|_2 \le \sqrt{2}\,\widetilde{B}.$$

Therefore,

$$\lambda\|\widetilde{W}^\star\|_F \le \frac{1}{n}\sum_{i=1}^{n}\sqrt{2}\,\widetilde{B} = \sqrt{2}\,\widetilde{B},$$

so every ERM minimizer satisfies

$$\|\widetilde{W}^\star\|_F \le \frac{\sqrt{2}\,\widetilde{B}}{\lambda}.$$

Substituting $R = \sqrt{2}\,\widetilde{B}/\lambda$ into the bounded-parameter constant gives

$$\begin{aligned} L_z &\le \sqrt{2} + \frac{\widetilde{B}}{2}\cdot\frac{\sqrt{2}\,\widetilde{B}}{\lambda} \\ &= \sqrt{2}\left(1+\frac{\widetilde{B}^2}{2\lambda}\right) \\ &= \sqrt{2}\left(1+\frac{B^2+1}{2\lambda}\right). \end{aligned}$$