DISTANCE-CORRECTED PRIVATE PROTOTYPES FOR LONG-TAILED CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

With a frozen public encoder, private prototype classification can use a single release of class sums and counts. Adding noise at the same scale to every class sum makes rare-class means less accurate and inflates squared distances to them. We introduce projected distance correction (PDC), which projects noisy class means onto the feature ball and corrects their distance scores to account for the added noise. When a fixed public feature bank is available, public-regularized distance correction (PRDC) instead retrieves features aligned with each released class sum and uses their mean to regularize each private prototype before correcting its score. Both methods preserve record-level differential privacy without additional privacy cost. Our analysis bounds deviations from non-private class scores and identifies when regularization preserves or changes the predicted class. It shows how accurate public means let PRDC reduce noise sensitivity while limiting the score changes caused by regularization. Experiments on three image datasets and two imbalance ratios show that, under the three strongest privacy settings, PDC improves balanced accuracy by an average of 20.0 percentage points over direct nearest-prototype classification. PRDC adds 3.0 points over PDC in this range, with positive average gains in five of six dataset-imbalance settings.

1 INTRODUCTION

Publicly pretrained encoders provide an effective starting point for differentially private image classification. The encoder remains fixed while the private learner operates on bounded features (Tramèr & Boneh, 2021; Mehta et al., 2023). Prototype classifiers represent each class by its mean feature vector and classify new examples by their distances to these prototypes (Snell et al., 2017; Wahdany et al., 2025). Long-tailed data expose a specific weakness of private prototypes. A record-level Gaussian mechanism adds noise at the same scale to every class sum, although rare classes contain far fewer records. Dividing these sums by class counts leaves rare-class means with much larger errors. Because sum noise increases squared distances on average, classes with fewer examples receive a larger systematic penalty.

We address this systematic penalty by correcting distance scores for the noise added to the class sums. The corrected scores can still fluctuate widely for rare classes, sometimes favoring them even for examples from other classes. To control both effects, we propose two classifiers that combine distance correction with prototype regularization using one joint release of class sums and counts. *Projected distance correction* (PDC) projects the noisy means onto the ball that bounds the encoder's features. When a public feature bank is available, *public-regularized distance correction* (PRDC) instead moves each noisy mean toward the average of public features most aligned with its released class sum. Both methods adjust the distance correction for the noise retained from the noisy mean after regularization.

Our analysis compares PDC and PRDC with the classifier based on the original class means, before privacy noise is added. We bound the changes in class scores caused by regularization and the remaining privacy noise. We then examine how regularization changes the score differences between competing classes, giving conditions for preserving or changing the original prediction. Projection reduces noise sensitivity by contracting toward the origin. An accurate public mean allows PRDC to reduce noise sensitivity while causing little change to the original class scores.

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.

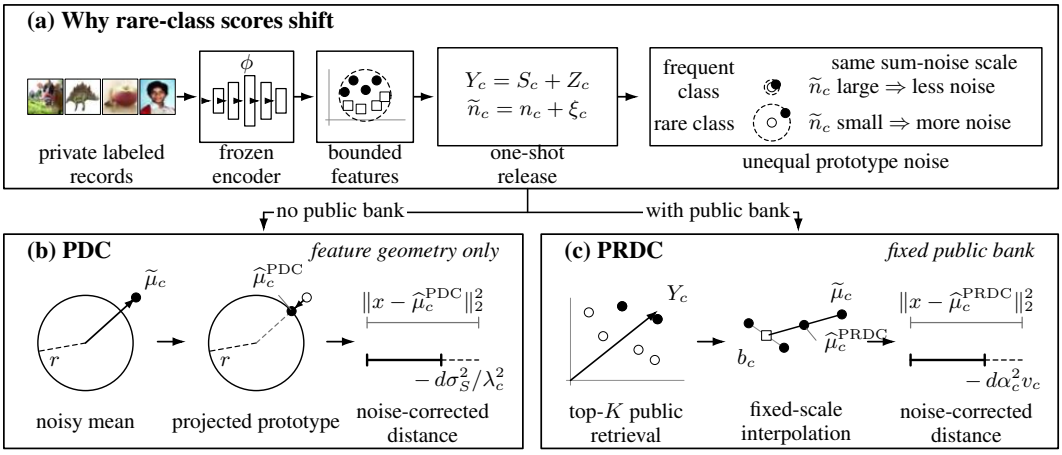


Figure 1: **(a)** With the same sum-noise scale, count normalization produces more uncertain prototypes and larger squared distances for rare classes. **(b)** PDC projects the noisy class mean onto the feature ball, then corrects the distance for added noise. **(c)** PRDC retrieves a public neighborhood, interpolates its mean with the noisy class mean, and corrects the distance score.

We evaluate CIFAR-10, CIFAR-100, and FOOD-101 with largest-to-smallest class ratios of 10 and 100. Averaged over the three smallest evaluated zCDP budgets, $\rho \in \{0.005, 0.02, 0.125\}$, PDC improves balanced accuracy over direct nearest-prototype classification with noisy class means in all six combinations. Its mean gain is 20.0 percentage points, and PRDC adds 3.0 points over PDC, with positive gains in five of the six combinations.

2 One-Shot Private Class Statistics

2.1 Learning setting

Let $D = \{(u_i, y_i)\}_{i=1}^N$ be a private labeled dataset with $y_i \in [C]$. Inputs, labels, and class counts are private. The class set, feature radius $r > 0$, and any public feature bank are fixed independently of D. A frozen encoder ϕ, trained only on public data, maps each input to $\mathbb{R}^d$. We normalize nonzero features to radius r as $x_i = r\phi(u_i)/\|\phi(u_i)\|_2$, and set $x_i = 0$ otherwise, so $\|x_i\|_2 \le r$. Test inputs are encoded and normalized in the same way. The private class statistics are

$$S_c = \sum_{i:y_i=c} x_i, \quad n_c = \sum_i \mathbf{1}\{y_i = c\}. \tag{1}$$

For $n_c > 0$, we write $\mu_c = S_c/n_c$ for the clean class mean.

2.2 Joint private release

Neighboring datasets differ by adding or removing one labeled record. The concatenated sum and count vectors have ℓ_2 sensitivities at most r and one, respectively. We release

$$Y_c = S_c + Z_c, \quad Z_c \sim \mathcal{N}(0, \sigma_S^2 I_d), \quad \sigma_S^2 = \frac{r^2}{2\rho_S}, \tag{2}$$

$$\widetilde{n}_c = n_c + \xi_c, \quad \xi_c \sim \mathcal{N}(0, \sigma_n^2), \quad \sigma_n^2 = \frac{1}{2\rho_n}. \tag{3}$$

The classwise noises are mutually independent, and the sum noises are independent of the count noises.

Proposition 1 (Privacy)**.** *Under record-level add/remove adjacency, the joint release $(Y, \widetilde{n})$ satisfies $(\rho_S + \rho_n)$-zCDP. Every deterministic function of this release and fixed public information has the same guarantee.*

The result follows from the Gaussian mechanism, composition, and postprocessing (Bun & Steinke, 2016; Dwork et al., 2006). We report the corresponding approximate-DP value $\varepsilon(\rho, \delta) = \rho + 2\sqrt{\rho \log(1/\delta)}$.

We use the fixed allocation

$$\rho_S = \frac{\sqrt{d}}{1+\sqrt{d}}\rho, \quad \rho_n = \frac{1}{1+\sqrt{d}}\rho. \tag{4}$$

This allocation minimizes a first-order mean-error bound (Appendix A.1). In experiments at IR 100, varying the count-budget share from half to twice this value changes the mean balanced accuracy of PDC and PRDC by less than 0.4 percentage points on each dataset, averaged over $\rho \in \{.005, .02, .125\}$ (Appendix D.5).

3 Distance-Corrected Private Prototypes

Both classifiers regularize the noisy class means before correcting their squared-distance scores (Figure 1). As postprocessing of the joint release, they preserve the privacy guarantee in Proposition 1.

We lower-bound each released count by one and form the noisy class mean:

$$\widehat{n}_c = \max\{\widetilde{n}_c, 1\}, \quad \widetilde{\mu}_c = \frac{Y_c}{\widehat{n}_c}, \quad v_c = \frac{\sigma_S^2}{\widehat{n}_c^2}. \tag{5}$$

Here v_c is the per-coordinate variance from sum noise, conditional on $\widehat{n}_c$.

3.1 Projected distance correction

A small released count can produce a noisy mean far outside the feature ball. PDC projects the noisy mean onto this ball:

$$\lambda_c = \max\left\{\widehat{n}_c, \frac{\|Y_c\|_2}{r}\right\}, \quad \widehat{\mu}_c^{\mathrm{PDC}} = \frac{Y_c}{\lambda_c}. \tag{6}$$

Outside the ball, projection multiplies $\widetilde{\mu}_c$ by $\widehat{n}_c/\lambda_c < 1$, contracting it toward the origin. The projected mean is no farther from the clean mean μ_c than the original noisy mean. PDC adds a correction for sum noise to the negative squared distance:

$$g_c^{\mathrm{PDC}}(x) = -\|x - \widehat{\mu}_c^{\mathrm{PDC}}\|_2^2 + \frac{d\sigma_S^2}{\lambda_c^2}. \tag{7}$$

When the noisy mean lies outside the ball, $\lambda_c = \|Y_c\|_2/r$ and $\widehat{n}_c$ drops out of the score.

3.2 Public-regularized distance correction

PRDC draws on a bank of feature vectors extracted from public data using the same encoder. It finds features relevant to each class and uses them to guide estimation when the noisy mean is inaccurate. Let $a_1, \ldots, a_M$ be the public features, with $\|a_j\|_2 = r$. PRDC retrieves the K features most aligned with the released sum Y_c and averages them:

$$J_c = \mathrm{TopK}_{j\in[M]}\, Y_c^\top a_j, \quad b_c = \frac{1}{K}\sum_{j\in J_c} a_j. \tag{8}$$

PRDC estimates the class mean by balancing squared distances to the noisy and public means:

$$\widehat{\mu}_c^{\mathrm{PRDC}} = \arg\min_m \left\{\frac{\|m - \widetilde{\mu}_c\|_2^2}{v_c} + \frac{d}{r^2}\|m - b_c\|_2^2\right\}. \tag{9}$$

The solution is a weighted average of the two means:

$$\widehat{\mu}_c^{\mathrm{PRDC}} = \alpha_c\widetilde{\mu}_c + (1-\alpha_c)b_c, \quad \alpha_c = \frac{r^2}{r^2 + dv_c}. \tag{10}$$

The noisy class mean receives more weight than the public mean when its total variance from sum noise, dv_c, is less than r^2. Smaller released counts also shift more weight to the public mean.

The weight α_c on the noisy mean enters the distance correction as α_c^2:

$$g_c^{\mathrm{PRDC}}(x) = -\|x - \widehat{\mu}_c^{\mathrm{PRDC}}\|_2^2 + d\alpha_c^2 v_c. \tag{11}$$

Both classifiers predict $\arg\max_c g_c(x)$.

4 RECOVERY OF CLEAN PROTOTYPE DECISIONS

We first bound the deviations of PDC and PRDC from scores based on the original class means. We then analyze how regularization changes the differences between class scores and, in turn, the predicted class.

4.1 DEVIATION FROM CLEAN CLASS SCORES

For a fixed query x, define the non-private empirical-mean score and margin

$$s_c^\star(x) = -\|x - \mu_c\|_2^2, \qquad \Delta_{cj}^\star(x) = s_c^\star(x) - s_j^\star(x). \tag{12}$$

This reference separates release distortion from sampling error. Before regularization, sum noise shifts the squared distance by

$$\mathbb{E}_{Z_c}\left[\left\|x - \frac{S_c + Z_c}{\widehat{n}_c}\right\|_2^2\right] = \left\|x - \frac{S_c}{\widehat{n}_c}\right\|_2^2 + dv_c. \tag{13}$$

An uncorrected score consequently penalizes smaller released counts. PDC and PRDC scale the correction by the private noise retained after regularization.

For comparison, let correction-only use the unprojected noisy mean,

$$g_c^{\mathrm{corr}}(x) = -\|x - \widetilde{\mu}_c\|_2^2 + dv_c. \tag{14}$$

For a sum-noise coefficient $t > 0$ and logarithmic factor $L > 0$, define the fluctuation bound

$$\mathcal{R}(t, L) = 4tr(\sigma_S + r\sigma_n)\sqrt{2L} + 2t^2\left[\sigma_S^2\sqrt{dL} + (\sigma_S + r\sigma_n)^2 L\right]. \tag{15}$$

Theorem 2 (Clean-score deviation). *Assume $n_c > 0$ for every class, fix $\delta \in (0, 1)$, and put*

$$L_0 = \log(10C/\delta), \quad L_M = \log(10CM/\delta).$$

With probability at least $1 - \delta$, jointly over classes and methods,

$$|g_c^{\mathrm{corr}}(x) - s_c^\star(x)| \le \mathcal{R}(1/\widehat{n}_c, L_0), \tag{16}$$

$$|g_c^{\mathrm{PDC}}(x) - s_c^\star(x)| \le 4r\left(1 - \frac{\widehat{n}_c}{\lambda_c}\right)\|\mu_c\|_2 + \mathcal{R}(1/\lambda_c, L_0), \tag{17}$$

$$|g_c^{\mathrm{PRDC}}(x) - s_c^\star(x)| \le 4r(1 - \alpha_c)\|b_c - \mu_c\|_2 + \mathcal{R}(\alpha_c/\widehat{n}_c, L_M). \tag{18}$$

In Equations (17) and (18), the displacement term bounds the score shift from moving the clean mean. The fluctuation term $\mathcal{R}$ bounds the remaining error from sum noise and the noisy count, including its lower bound. For PDC, the displacement bound is small for class means near the origin and grows as $\widehat{n}_c/\lambda_c$ falls. For PRDC, when b_c is close to μ_c, a small coefficient $\alpha_c/\widehat{n}_c$ incurs little displacement.

The regularized sum-noise coefficients are bounded in terms of r and d alone:

$$\frac{\sigma_S}{\lambda_c} \le r\sqrt{\frac{2}{d}}, \quad \frac{\sigma_S\alpha_c}{\widehat{n}_c} \le \frac{r}{2\sqrt{d}}. \tag{19}$$

The PDC inequality holds jointly over classes with probability at least $1 - Ce^{-d/16}$ over the Gaussian sum noise; the PRDC inequality holds for every $\widehat{n}_c \ge 1$. When $dv_c > 2r^2$, both upper bounds are smaller than the sum-noise scale of the unregularized mean, $\sigma_S/\widehat{n}_c = \sqrt{v_c}$. Appendix A gives the proofs and moment identities.

4.2 REGULARIZATION SHIFTS AND CLASS DECISIONS

Individual class scores may change substantially while their ordering remains unchanged. To analyze their differences, we retain the signed score changes from moving the clean means:

$$\begin{aligned} B_c^{\mathrm{PDC}}(x) &= \|x - \mu_c\|_2^2 - \left\|x - \frac{\widehat{n}_c}{\lambda_c}\mu_c\right\|_2^2, \\ B_c^{\mathrm{PRDC}}(x) &= \|x - \mu_c\|_2^2 - \left\|x - \left[\alpha_c\mu_c + (1 - \alpha_c)b_c\right]\right\|_2^2. \end{aligned} \tag{20}$$

Write $R_c^{\mathrm{PDC}} = \mathcal{R}(1/\lambda_c, L_0)$ and $R_c^{\mathrm{PRDC}} = \mathcal{R}(\alpha_c/\widehat{n}_c, L_M)$ for the corresponding fluctuation bounds.

Corollary 3 (Class margins under regularization)**.** *Under the assumptions of Theorem 2, with probability at least* $1-\delta$*, its bounds and the following inequality hold simultaneously for* $h \in \{\mathrm{PDC}, \mathrm{PRDC}\}$ *and all* $c \neq j$*:*

$$\left|g_c^h(x) - g_j^h(x) - \left[\Delta_{cj}^\star(x) + B_c^h(x) - B_j^h(x)\right]\right| \leq R_c^h + R_j^h. \tag{21}$$

On this event, let c maximize $s_c^\star(x)$. Method h retains c if $\Delta_{cj}^\star(x) + B_c^h(x) - B_j^h(x) > R_c^h + R_j^h$ for every $j \neq c$. If this adjusted margin is less than $-(R_c^h + R_j^h)$ for some $j \neq c$, then $g_j^h(x) > g_c^h(x)$, so h does not select c.

For PDC, contraction toward the origin gives

$$B_c^{\mathrm{PDC}}(x) = \left(1 - \frac{\widehat{n}_c}{\lambda_c}\right)\left[\left(1 + \frac{\widehat{n}_c}{\lambda_c}\right)\|\mu_c\|_2^2 - 2x^\top \mu_c\right].$$

The shift vanishes when $\widehat{n}_c = \lambda_c$. When $\widehat{n}_c < \lambda_c$, its sign depends on the query's alignment with the class mean and the mean's norm.

For PRDC, movement toward the public mean gives

$$B_c^{\mathrm{PRDC}}(x) = 2(1-\alpha_c)(x-\mu_c)^\top(b_c - \mu_c) - (1-\alpha_c)^2\|b_c - \mu_c\|_2^2.$$

The inner product compares the directions from μ_c to the query and to the public mean b_c; the quadratic term penalizes the distance moved. When $b_c = \mu_c$, the shift is zero for any α_c; a nearby public mean keeps it small even under strong regularization.

Count noise. The term $S_c/\widehat{n}_c = (n_c/\widehat{n}_c)\mu_c$ expands or contracts the clean mean along its own direction. This change can increase or decrease the squared distance, depending on the query's position relative to the class mean. Through λ_c and α_c, the released count also affects the regularization shifts B_c^h. The terms involving σ_n in $\mathcal{R}$ bound the remaining count error and its interaction with sum noise.

Noisy retrieval. Sum noise can change the selected public neighborhood. Let b_c^0 be the public mean retrieved with the unnoised sum S_c. With probability at least $1-\delta$, jointly with Theorem 2, for every class,

$$\|b_c - \mu_c\|_2 \leq \sqrt{\|b_c^0 - \mu_c\|_2^2 + r^2 - \|b_c^0\|_2^2} + \frac{4\sigma_S}{n_c}\sqrt{L_M}.$$

The quantity under the square root is the mean squared distance of the noiseless neighbors from μ_c. If these neighbors concentrate near the class mean and $\sigma_S\sqrt{L_M}/n_c$ is small, the bound keeps the magnitude of $B_c^{\mathrm{PRDC}}(x)$ small (Appendix B.1). Noisy retrieval also affects the score through the interaction $Z_c^\top b_c$. Its contribution remains after distance correction and is controlled by the terms involving L_M in $\mathcal{R}$. Appendix D.3 reports the measured residual after correction.

5 Experiments

5.1 Experimental setup

Data and features. We use CIFAR-10, CIFAR-100, and FOOD-101 (Bossard et al., 2014). We subsample training data to obtain long-tailed distributions with imbalance ratios $\mathrm{IR} \in \{10, 100\}$, keeping the test sets unchanged. Class sizes decrease geometrically with rank. For example, at IR 100, CIFAR-100 has 500 examples in its largest class and five in its smallest. All methods start from a frozen SWAG ViT-B/16 encoder (Singh et al., 2022) and 768-dimensional unit-length features. Appendix D.8 also evaluates ResNet-18 on CIFAR-10.

Evaluation and privacy. For each dataset and IR, all methods use the same 20 independently sampled training sets and ten randomized runs per set and budget. We average the ten runs within each set, then report the mean and standard deviation across the 20 set-level averages. Balanced accuracy (BA) averages classwise recall and equals accuracy on our class-balanced test sets. Macro-F1 averages classwise F1 scores, which combine precision and recall. We use $r = 1$ and $\rho \in \{.005, .02, .125, .5, 2, 8\}$, with $\delta = 10^{-5} < 1/N$ for every training set of size N. Conversion to approximate DP gives $\varepsilon \approx .485, .980, 2.524, 5.299, 11.597, 27.194$, respectively.

Baselines. As a direct baseline, *Noisy mean* classifies by distance to $Y_c/\widehat{n}_c$, without prototype regularization or distance correction. DPPL-Mean adjusts feature dimension by coordinate averaging with a chosen width, then releases Gaussian-perturbed class means. We use fixed public class counts and calibrate its noise to our add/remove adjacency (Wahdany et al., 2025). Its averaging width is selected on public development data for each source and budget, then fixed across target datasets and IRs. DPPL-Public clips feature similarities and uses the exponential mechanism to select one public feature per class (Wahdany et al., 2025). Its clipping interval is selected on public development tasks and held fixed across target datasets, IRs, and budgets for each public source.

We also compare with the private least-squares classifier DP-LS and the feature-covariance-based linear classifier DP-FC (Mehta et al., 2023). DP-SGD trains only a linear softmax head on features from the frozen, publicly pretrained encoder (Abadi et al., 2016). DP-SGD+LA applies post-hoc logit adjustment using privately released class frequencies (Menon et al., 2021). DP-LS uses fixed settings; DP-FC and DP-SGD settings are selected on public development data. For DP-SGD, these tasks contain 10,000 training examples, and one setting per public source and budget is applied to both target IRs. LA's adjustment strength is selected on the same public tasks. Both SGD variants use Opacus's PRV accountant (Gopi et al., 2021) to target $(\varepsilon(\rho, 10^{-5}), 10^{-5})$-DP, including the count release in DP-SGD+LA. Appendix C.1 gives the baseline settings.

Public data and neighborhood selection. PRDC and DPPL-Public use the same full bank: 1,281,167 ImageNet features for CIFAR-10/100 (Deng et al., 2009) and 31,395 UECFood-256 features for FOOD-101 (Kawano & Yanai, 2015). We select K by validation BA on classification tasks constructed from each public source. The resulting $K = 64$ for ImageNet and $K = 256$ for UECFood remain fixed in the main comparisons across target datasets, IRs, and budgets. Across the tested neighborhood sizes, larger K modestly benefits CIFAR-100, whereas FOOD-101 favors smaller neighborhoods (Appendix D.6). Appendix D.7 varies bank size and compares sources at matched size. Appendix C gives the development tasks.

5.2 Private long-tailed classification

PDC exceeds DPPL-Mean and DP-LS in both metrics on all six dataset–IR combinations, averaged over the three smallest budgets (Table 1). Relative to Noisy mean, its gains average 20.0 percentage points (pp) in BA and 23.6 pp in Macro-F1.

PRDC adds 3.0 pp in BA and 3.6 pp in Macro-F1 over PDC. The additional gains are larger on CIFAR-100 and FOOD-101 than on CIFAR-10. Both metrics improve in five combinations; CIFAR-10 at IR 10 changes by -0.1 pp. PRDC also exceeds the same-bank DPPL-Public in both metrics across all six settings.

Over these budgets, both proposed methods exceed DP-FC and DP-SGD in both metrics on all six settings. Logit adjustment improves both metrics of DP-SGD in every setting. Compared with DP-SGD+LA, PDC has higher BA in both CIFAR-10 settings and trails by less than 1.0 pp in the other four, with higher Macro-F1 in all six. PDC achieves competitive performance from a single statistics release, without additional public tuning data or iterative head training. PRDC exceeds DP-SGD+LA in both metrics across all six settings. Appendix D reports the results at every privacy level.

5.3 Contributions of projection and correction

Figure 2 isolates the two operations in PDC using the same noisy sums and counts. On CIFAR-100 and FOOD-101, adding correction to the projected prototype raises accuracy substantially, while projection alone stays much closer to Noisy mean. Bounding the prototype does not by itself remove the class-dependent distance penalty.

Correction without projection nearly matches PDC on CIFAR-10 at IR 10, but its BA stays near 1% on CIFAR-100 and FOOD-101 at IR 100 for $\rho \leq .125$. Removing the average distance penalty is insufficient in these small-class regimes: the realized scores still fluctuate, with small released counts amplifying the variance in Equation (25). Projection reduces the noise coefficient in Theorem 2.

Table 1: Long-tailed classification: balanced accuracy and Macro-F1 (%), averaged over $\rho \in \{.005, .02, .125\}$. All methods use the same frozen public encoder; groups describe the use of additional public data. Entries are mean ± SD across training sets. Within each IR, bold marks the highest mean per column among methods without a retrieval bank (first two groups) and with one (last group).

Method	CIFAR-10		CIFAR-100		FOOD-101	
	BA	Macro-F1	BA	Macro-F1	BA	Macro-F1
(a) IR 100						
No additional public data						
Noisy mean	60.44 ± 1.06	51.21 ± 1.53	17.37 ± 0.52	10.83 ± 0.55	22.47 ± 0.83	13.87 ± 0.59
DP-LS	64.12 ± 0.80	60.87 ± 1.02	18.69 ± 0.61	14.32 ± 0.54	21.02 ± 0.77	15.75 ± 0.57
PDC	$\mathbf{85.61} \pm 0.64$	$\mathbf{84.70} \pm 0.77$	36.63 ± 0.76	$\mathbf{30.37} \pm 0.70$	41.93 ± 0.88	$\mathbf{34.85} \pm 0.88$
Public data for tuning						
DPPL-Mean	73.41 ± 0.71	70.75 ± 1.00	30.39 ± 0.79	22.46 ± 0.69	35.86 ± 0.87	27.80 ± 0.86
DP-FC	56.86 ± 3.37	48.05 ± 4.11	25.27 ± 1.01	16.50 ± 1.22	24.97 ± 1.11	15.65 ± 0.92
DP-SGD	78.56 ± 1.39	76.03 ± 1.74	35.82 ± 0.89	26.62 ± 0.93	40.63 ± 1.00	31.25 ± 1.03
DP-SGD+LA	80.69 ± 1.11	78.50 ± 1.44	$\mathbf{36.86} \pm 0.83$	28.29 ± 0.85	$\mathbf{42.12} \pm 0.91$	33.26 ± 0.93
Public data for prototype construction						
DPPL-Public	81.84 ± 1.08	81.10 ± 1.28	37.73 ± 1.04	32.99 ± 0.98	32.45 ± 1.00	27.75 ± 0.92
PRDC	$\mathbf{86.41} \pm 0.86$	$\mathbf{85.57} \pm 1.04$	$\mathbf{41.46} \pm 0.79$	$\mathbf{36.17} \pm 0.90$	$\mathbf{46.42} \pm 0.78$	$\mathbf{40.81} \pm 0.75$
(b) IR 10						
No additional public data						
Noisy mean	90.63 ± 0.44	90.28 ± 0.58	30.82 ± 0.70	24.37 ± 0.78	39.53 ± 0.77	32.24 ± 0.57
DP-LS	79.74 ± 0.49	79.62 ± 0.55	29.82 ± 0.42	26.68 ± 0.33	31.97 ± 0.53	28.23 ± 0.35
PDC	$\mathbf{93.40} \pm 0.07$	$\mathbf{93.41} \pm 0.07$	58.63 ± 0.38	$\mathbf{57.00} \pm 0.25$	65.30 ± 0.28	$\mathbf{63.82} \pm 0.31$
Public data for tuning						
DPPL-Mean	82.14 ± 0.09	82.08 ± 0.09	48.80 ± 0.31	46.54 ± 0.27	55.61 ± 0.24	54.20 ± 0.30
DP-FC	87.85 ± 1.57	87.38 ± 1.86	43.34 ± 1.10	36.39 ± 1.11	42.47 ± 1.47	34.57 ± 1.75
DP-SGD	92.49 ± 0.19	92.48 ± 0.20	57.83 ± 0.25	54.28 ± 0.31	65.18 ± 0.46	62.21 ± 0.65
DP-SGD+LA	92.77 ± 0.16	92.76 ± 0.17	$\mathbf{58.66} \pm 0.24$	55.35 ± 0.29	$\mathbf{66.27} \pm 0.41$	63.51 ± 0.58
Public data for prototype construction						
DPPL-Public	86.36 ± 0.30	86.28 ± 0.31	54.90 ± 0.50	54.19 ± 0.53	45.90 ± 0.36	43.72 ± 0.42
PRDC	$\mathbf{93.28} \pm 0.12$	$\mathbf{93.30} \pm 0.13$	$\mathbf{63.61} \pm 0.21$	$\mathbf{62.85} \pm 0.26$	$\mathbf{68.10} \pm 0.35$	$\mathbf{67.08} \pm 0.46$

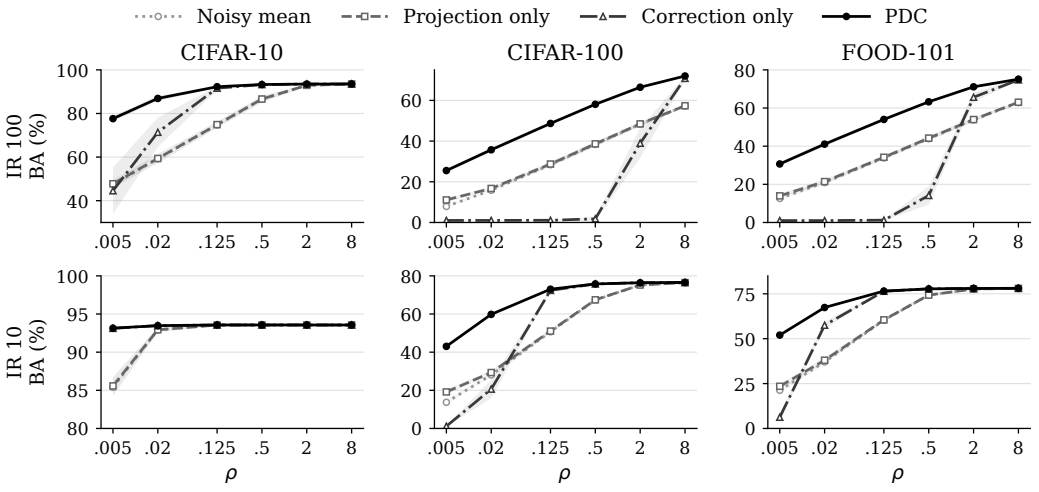


Figure 2: PDC components across six privacy budgets. All variants share the same noisy sums and counts. Points show means across training sets; bands show one standard deviation.

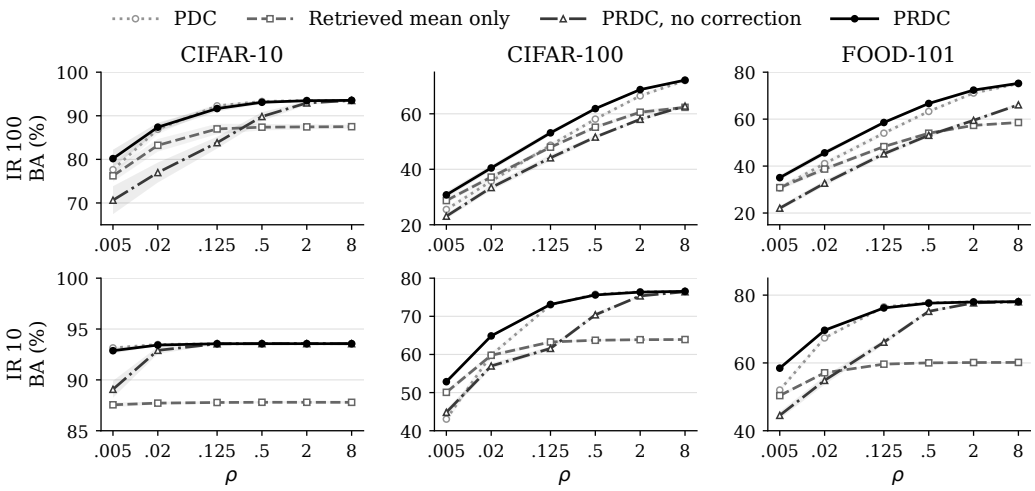


Figure 3: PRDC components under the same protocol as Figure 2. Retrieved mean only classifies by distance to b_c. Points show means across training sets; bands show one standard deviation.

Table 2: Regularization-target controls: BA (%), averaged over $\rho \in \{.005, .02, .125\}$. The three PRDC variants share weights, correction, and private releases; only the target changes. Entries are mean $\pm$ SD across training sets.

Method	CIFAR-10		CIFAR-100		FOOD-101	
	IR 100	IR 10	IR 100	IR 10	IR 100	IR 10
PDC	85.61 ± 0.64	93.40 ± 0.07	36.63 ± 0.76	58.63 ± 0.38	41.93 ± 0.88	65.30 ± 0.28
Zero target	79.23 ± 0.85	93.42 ± 0.08	30.07 ± 0.78	51.07 ± 0.50	34.50 ± 0.86	57.80 ± 0.50
Wrong-class target	71.41 ± 1.27	93.39 ± 0.09	21.07 ± 0.64	39.14 ± 0.51	28.89 ± 0.81	53.04 ± 0.50
PRDC	86.41 ± 0.86	93.28 ± 0.12	41.46 ± 0.79	63.61 ± 0.21	46.42 ± 0.78	68.10 ± 0.35

5.4 When public information helps

Figure 3 separates the contributions of the retrieved public mean, interpolation with the noisy mean, and distance correction. The public mean is competitive at small budgets, when the noisy mean is least reliable. As the budget grows, its accuracy levels off while PRDC continues to improve by retaining private information.

The two PRDC variants use the same prototype, so their accuracy gap isolates the effect of score correction. On CIFAR-100 and FOOD-101 at IR 100, this gap persists across the budget range; at IR 10, it narrows as the budget grows.

The role of class-specific regularization targets. Table 2 tests whether PRDC benefits from the class information in its public mean beyond the amount of shrinkage. We replace b_c with zero or the mean retrieved for the preceding class in label order. The latter is rescaled to the original target's norm. On CIFAR-100 and FOOD-101 at both IRs, both replacements have lower BA than PDC, while PRDC exceeds it. The wrong-class target also has lower BA than the zero target in these four settings: preserving the target's norm does not preserve its benefit when its direction is changed.

Benefits across class sizes. On CIFAR-100 and FOOD-101 at IR 100, tail recall is near zero for all three methods at $\rho = .005$ (Figure 4). At $\rho = .125$, PDC reaches 7.5% and 13.6%, and PRDC reaches 14.5% and 26.8%, respectively, while Noisy mean remains at zero. For middle classes, PRDC leads already at $\rho = .005$; PDC approaches it by $\rho = .125$, while Noisy mean remains much lower.

On these two datasets, PDC and PRDC also improve head recall at the smallest budget. On CIFAR-10, Noisy mean has higher head recall, but PDC and PRDC substantially improve tail recall at small budgets. PRDC's additional benefit for the hardest tail groups is largest at intermediate budgets.

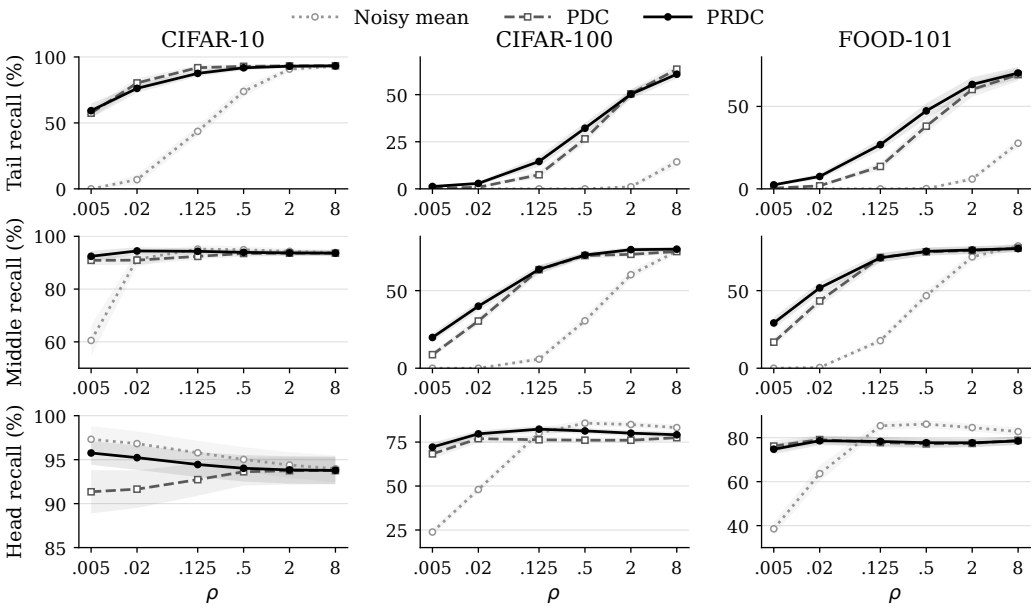


Figure 4: Recall at IR 100 for classes grouped by training size into tail, middle, and head thirds. Points show means across training sets; bands show one standard deviation.

Computational cost. With precomputed features and banks loaded, PDC/PRDC construction takes 1–10 ms on one NVIDIA RTX PRO 6000 GPU, including PRDC's exact GPU Top-K retrieval, versus 0.13–0.79 s for DP-SGD variants. PRDC's bank loading adds 0.05–2.05 s and can be shared across releases (Appendix D.9).

6 Related Work

Public features and private estimation. Private heads on fixed public features use closed-form, second-order, or iterative updates (Tramèr & Boneh, 2021; Mehta et al., 2023; Abadi et al., 2016). DPPL releases noisy means or privately selects public prototypes (Wahdany et al., 2025). Projection methods exploit known constraints after a private release (Ganesh et al., 2023; Lowy et al., 2024; Talwar et al., 2016; Balle & Wang, 2018; Cohen-Addad et al., 2024). PRDC combines shrinkage estimation (James & Stein, 1961; Stein, 1981) with a public mean retrieved from a private release. In non-private few-shot learning, prototype completion uses semantic priors (Zhang et al., 2021), while covariance-aware classifiers adapt the distance metric (Bateni et al., 2020).

Long tails and heterogeneous score distortion. Logit adjustment compensates for class-prior imbalance in non-private long-tailed classifiers (Menon et al., 2021). Privacy noise introduces class-dependent estimation error and can amplify disparities (Bagdasaryan et al., 2019; Rosenblatt et al., 2025), while heterogeneous estimation noise inflates distances by unequal amounts (Calmon et al., 2025; Hillebrand et al., 2026; Li et al., 2025). Distance-adjusted classifiers subtract estimated mean-squared error from squared distances (Aoshima & Yata, 2014). PDC and PRDC combine this correction with regularization to limit the score fluctuations amplified by small noisy counts.

7 Conclusion and Limitations

PDC and PRDC combine prototype regularization with distance correction for one-shot private classification. Both methods still struggle with rare classes under strong privacy. PRDC can underperform PDC with a mismatched public bank, and noisy retrieval can limit its gains even with a suitable bank.

AI USE STATEMENT

Generative AI tools assisted with research ideation and execution, drafting, and polishing manuscript text. The authors take responsibility for the final text, methods, claims, and results.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Makoto Aoshima and Kazuyoshi Yata. A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, 66(5): 983–1010, 2014.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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*, pp. 394–403, 2018.

Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

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

Flavio Calmon, Elbert Du, Cynthia Dwork, Brian Finley, and Grigory Franguridi. Debiasing Functions of Private Statistics in Postprocessing. In *6th Symposium on Foundations of Responsible Computing (FORC 2025)*, pp. 17:1–17:18, 2025.

Mark Cesar and Ryan Rogers. Bounding, concentrating, and truncating: Unifying privacy loss composition for data analytics. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 421–457, 2021.

Vincent Cohen-Addad, Tommaso D'Orsi, Alessandro Epasto, Vahab Mirrokni, and Peilin Zhong. Perturb-and-project: Differentially private similarities and marginals. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9161–9179, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10611–10627, 2023.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11631–11642, 2021.

Quentin Hillebrand, Jacob Imola, Rasmus Pagh, and Sia Sejer. Privacy by postprocessing the discrete laplace mechanism. *arXiv preprint arXiv:2605.06502*, 2026.

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, pp. 361–379, 1961.

Yoshiyuki Kawano and Keiji Yanai. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *Computer Vision - ECCV 2014 Workshops*, pp. 3–17, 2015.

Keyi Li, Yuval Kluger, and Boris Landa. Euclidean distance deflation under high-dimensional heteroskedastic noise. *arXiv preprint arXiv:2507.18520*, 2025.

Andrew Lowy, Zeman Li, Tianjian Huang, and Meisam Razaviyayn. Optimal differentially private model training with public data. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 32849–32903, 2024.

Harsh Mehta, Walid Krichene, Abhradeep Guha Thakurta, Alexey Kurakin, and Ashok Cutkosky. Differentially private image classification from features. *Transactions on Machine Learning Research*, 2023.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2021.

Lucas Rosenblatt, Yuliia Lut, Ethan Turok, Marco Avella Medina, and Rachel Cummings. Differential privacy under class imbalance: Methods and empirical insights. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 52065–52109, 2025.

Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 804–814, 2022.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6), 1981.

Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2016.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3754–3762, 2021.

A Proofs of the Main Results

We prove the privacy guarantee, the clean-score bound, and the margin corollary.

A.1 Privacy and the fixed budget allocation

Proof of Proposition 1. For the concatenated class-sum query, an added record changes exactly one class block by x with $\|x\|_2 \le r$; hence its ℓ_2 sensitivity is r. The count vector changes in one coordinate by one. Equations (2) and (3) are ρ_S-zCDP and ρ_n-zCDP Gaussian mechanisms. Sequential composition and postprocessing prove the claim. □

The public class set remains fixed when a class has no training examples. Such a class has $S_c = 0$ and $Y_c = Z_c$; the count lower bound keeps both classifiers defined, and the same sensitivity argument applies. The clean-score results assume $n_c > 0$ because their reference mean $\mu_c = S_c/n_c$ requires an observed class.

For counts away from the lower bound, the first-order ratio expansion is

$$\frac{Y_c}{\widehat{n}_c} - \mu_c \approx \frac{Z_c}{n_c} - \frac{\mu_c \xi_c}{n_c}. \tag{22}$$

The expected squared error of this first-order approximation is

$$\frac{1}{n_c^2}\left(d\sigma_S^2 + \|\mu_c\|_2^2 \sigma_n^2\right) \le \frac{r^2}{2n_c^2}\left(\frac{d}{\rho_S} + \frac{1}{\rho_n}\right). \tag{23}$$

For $\rho_S + \rho_n = \rho$, Cauchy–Schwarz yields

$$\frac{d}{\rho_S} + \frac{1}{\rho_n} \ge \frac{(\sqrt{d}+1)^2}{\rho}, \tag{24}$$

with equality exactly when $\rho_S/\rho_n = \sqrt{d}$. Equation (4) minimizes this upper bound.

A.2 Noisy-mean score identities

Conditional on the training data, test feature, and count $\widehat{n}_c$, write $\widetilde{\mu}_c = \frac{S_c}{\widehat{n}_c} + \sqrt{v_c} G_c$, where $G_c \sim \mathcal{N}(0, I_d)$. Expanding the corrected score gives

$$g_c^{\mathrm{corr}}(x) = -\left\| x - \frac{S_c}{\widehat{n}_c} \right\|_2^2 + 2\sqrt{v_c}\left(x - \frac{S_c}{\widehat{n}_c}\right)^\top G_c - v_c(\|G_c\|_2^2 - d).$$

The linear and centered quadratic terms have zero expectation and zero covariance, since the third moments of a centered Gaussian vanish. Their variances add to give

$$\mathrm{Var}_{Z_c}[g_c^{\mathrm{corr}}(x) \mid \widehat{n}_c] = 4v_c \left\| x - \frac{S_c}{\widehat{n}_c} \right\|_2^2 + 2dv_c^2. \tag{25}$$

Taking expectations in the expansion also gives Equation (13).

Pairwise decisions. Keep the training data and test feature fixed, and condition on the counts $\widehat{n}_c, \widehat{n}_j$. Subtracting the two independent noisy-distance expansions gives the noisy-mean classifier's expected margin in favor of class c:

$$\left\| x - \frac{S_j}{\widehat{n}_j} \right\|_2^2 - \left\| x - \frac{S_c}{\widehat{n}_c} \right\|_2^2 - d(v_c - v_j). \tag{26}$$

If $\widehat{n}_c < \widehat{n}_j$, then $v_c > v_j$: sum noise lowers the expected margin for class c and can reverse a decision that favors it without sum noise. Distance correction adds $d(v_c - v_j)$ to this margin, removing the sum-noise penalty in its expectation. Conditional on the counts, this adjustment is constant and leaves the variance unchanged. In particular, the $2dv_c^2$ term in Equation (25) scales as $\widehat{n}_c^{-4}$, explaining why small released counts can produce large fluctuations even after the average penalty is removed.

A.3 Clean-score decomposition and concentration

We prove Theorem 2 through a common score decomposition. Fix one class and suppress its index. Write

$$Y = n\mu + Z, \quad \widehat{n} = \max\{n + \xi, 1\}, \quad e = \widehat{n} - n.$$

The count error obeys

$$|e| = |\max\{n + \xi, 1\} - n| \le |\xi|. \tag{27}$$

The inequality is immediate when the lower bound is inactive. Otherwise, $n - 1 \le n - (n + \xi)$.

Consider coefficients $a, \beta \ge 0$ with $a + \beta \le 1$, a vector b in the feature ball, and

$$\widehat{\mu} = a\frac{Y}{\widehat{n}} + \beta b, \quad g(x) = -\|x - \widehat{\mu}\|_2^2 + d\sigma_S^2 k^2, \quad k = \frac{a}{\widehat{n}}. \tag{28}$$

The coefficients and b may depend on the release. Put $\bar{\mu} = a\mu + \beta b$. The exact identity

$$\widehat{\mu} = \bar{\mu} + k(Z - e\mu)$$

gives

$$g(x) - s^\star(x) = \underbrace{\|x-\mu\|_2^2 - \|x-\bar{\mu}\|_2^2}_{\text{score shift}} + \underbrace{2k(x-\bar{\mu})^\top(Z-e\mu) - k^2\big(\|Z-e\mu\|_2^2 - d\sigma_S^2\big)}_{\text{noise remainder}}. \tag{29}$$

Correction-only has $(a,\beta,k) = (1,0,1/\widehat{n})$. PDC has $(a,\beta,k) = (\widehat{n}/\lambda, 0, 1/\lambda)$. PRDC has $(a,\beta,k) = (\alpha, 1-\alpha, \alpha/\widehat{n})$.

Lemma 4 (Uniform noise remainder)**.** *Let b be any release-dependent convex combination of a fixed bank of M vectors with norm at most r. For $\delta \in (0,1)$, put $u = \log(10/\delta)$ and $u_M = \log(10M/\delta)$. With probability at least $1-\delta$,*

$$\begin{aligned}&\big|2k(x-\bar{\mu})^\top(Z-e\mu) - k^2\big(\|Z-e\mu\|_2^2 - d\sigma_S^2\big)\big| \\ &\quad\le 2k\sigma_S r\big[(1+a)\sqrt{2u} + \beta\sqrt{2u_M}\big] + 4kr^2\sigma_n\sqrt{2u} \\ &\qquad + 2k^2\sigma_S^2(\sqrt{du}+u) + 4k^2 r\sigma_S\sigma_n u + 2k^2 r^2\sigma_n^2 u.\end{aligned} \tag{30}$$

The event is simultaneous over all admissible a,β,b. When $\beta = 0$, the bank term and the union bound over its entries are unnecessary.

Proof. The Gaussian tail bound and a union bound give

$$|x^\top Z|,\ |\mu^\top Z| \le \sigma_S r\sqrt{2u}, \quad |b^\top Z| \le \max_{j\in[M]} |a_j^\top Z| \le \sigma_S r\sqrt{2u_M}.$$

These inequalities hold simultaneously for every convex combination of the bank, including one selected with Z. Equation (27) and the count-noise tail bound give $|e| \le \sigma_n\sqrt{2u}$. The Laurent–Massart inequality gives

$$\big|\|Z\|_2^2 - d\sigma_S^2\big| \le 2\sigma_S^2(\sqrt{du}+u).$$

Expand the noise remainder in Equation (29). The directional term is bounded by

$$2k\sigma_S r\big[(1+a)\sqrt{2u} + \beta\sqrt{2u_M}\big],$$

and the term linear in e is at most $4kr^2\sigma_n\sqrt{2u}$. Expanding $\|Z-e\mu\|_2^2$ gives the final three terms in Equation (30). The five Gaussian events have total failure probability at most δ. □

Proof of Theorem 2. Use the logarithmic factors L_0 and L_M defined in the theorem. The bounds on directional inner products, count noise, and the quadratic term fail with total probability at most $8Ce^{-L_0}$; the bank inner-product bound contributes at most $2CMe^{-L_M}$. Their sum is at most δ. Since $1+a+\beta \le 2$, the right-hand side of Lemma 4 is bounded by $\mathcal{R}(k, L_0)$ for correction-only and PDC. For PRDC, bounding both logarithmic levels by L_M gives $\mathcal{R}(k, L_M)$. Substituting the three choices of (a,β,k) gives the noise-fluctuation bounds in Theorem 2.

It remains to bound the regularization-induced score shifts. Because $x, \mu, \bar{\mu}$ lie in the radius-r ball,

$$\big|\|x-\mu\|_2^2 - \|x-\bar{\mu}\|_2^2\big| \le 4r\|\bar{\mu}-\mu\|_2. \tag{31}$$

The prototype displacement $\|\bar{\mu}-\mu\|_2$ is zero for correction-only, $(1-\widehat{n}/\lambda)\|\mu\|_2$ for PDC, and $(1-\alpha)\|b-\mu\|_2$ for PRDC. This proves Equations (16)–(18). □

Proof of Equation (19). For each class, $Y_c = S_c + Z_c$ with $Z_c \sim \mathcal{N}(0, \sigma_S^2 I_d)$. Gaussian integration and Markov's inequality give

$$\begin{aligned}\Pr\left(\|Y_c\|_2^2 \le \frac{d\sigma_S^2}{2}\right) &\le e^{d/4}\,\mathbb{E}\exp\left(-\frac{\|Y_c\|_2^2}{2\sigma_S^2}\right) \\ &= \exp\left(\frac{d}{4} - \frac{d}{2}\log 2 - \frac{\|S_c\|_2^2}{4\sigma_S^2}\right) \le e^{-d/16}.\end{aligned}$$

A union bound over classes and $\lambda_c \geq \|Y_c\|_2/r$ yield the first inequality. For PRDC, the definition of α_c and the arithmetic–geometric mean inequality yield, for every $\widehat{n}_c \geq 1$,

$$\frac{\sigma_S \alpha_c}{\widehat{n}_c} = \frac{\sigma_S r^2 \widehat{n}_c}{r^2 \widehat{n}_c^2 + d\sigma_S^2} \leq \frac{r}{2\sqrt{d}}.$$

Substituting these bounds into $\mathcal{R}$ in Theorem 2 requires both the theorem's score bounds and the coefficient bounds above to hold. For PDC, a union bound gives probability at least $1 - \delta - Ce^{-d/16}$ for their intersection. For PRDC, the coefficient bound holds deterministically, so the substituted score bound retains the theorem's $1 - \delta$ guarantee. □

A.4 Projection geometry

Lemma 5 (Projection geometry)**.** *For every release,*

$$\|\widehat{\mu}_c^{\mathrm{PDC}} - \mu_c\|_2^2 \leq \|\widetilde{\mu}_c - \mu_c\|_2^2 - \big(\|\widetilde{\mu}_c\|_2 - r\big)_+^2. \tag{32}$$

Proof. Let $u = \widetilde{\mu}_c$ and let p be its projection onto the radius-r ball. Since μ_c belongs to this ball, the projection variational inequality gives $\|p - \mu_c\|_2^2 \leq \|u - \mu_c\|_2^2 - \|u - p\|_2^2$. Substitute $\|u - p\|_2 = (\|u\|_2 - r)_+$. □

A.5 Proof of the margin corollary

Proof of Corollary 3. For PDC, substitute $(a, \beta, k) = (\widehat{n}_c/\lambda_c, 0, 1/\lambda_c)$ into Equation (29). For PRDC, substitute $(a, \beta, k) = (\alpha_c, 1-\alpha_c, \alpha_c/\widehat{n}_c)$. The regularization-induced score shifts are exactly those in Equation (20). On the joint event constructed in the proof of Theorem 2, the uniform noise-remainder bound gives

$$|g_c^h(x) - s_c^\star(x) - B_c^h(x)| \leq R_c^h, \quad h \in \{\mathrm{PDC}, \mathrm{PRDC}\}. \tag{33}$$

The event holds simultaneously over classes and the release-dependent coefficients and public mean. Subtract the two classwise decompositions and apply the triangle inequality to obtain Equation (21) on that same event.

For a clean winner c, the margin obeys

$$g_c^h(x) - g_j^h(x) \geq \Delta_{cj}^\star(x) + B_c^h(x) - B_j^h(x) - (R_c^h + R_j^h),$$
$$g_c^h(x) - g_j^h(x) \leq \Delta_{cj}^\star(x) + B_c^h(x) - B_j^h(x) + (R_c^h + R_j^h).$$

If the lower bound is positive for every $j \neq c$, then c has the unique highest score. If the upper bound is negative for some $j \neq c$, then $g_j^h(x) > g_c^h(x)$ and c cannot be selected. □

Expanding the squared distances in Equation (20) gives the two expressions interpreted in Section 4.2. If the ratios $\widehat{n}_c/\lambda_c$ and mean norms are common across classes, the scores $s_c^\star(x) + B_c^{\mathrm{PDC}}(x)$ have the clean ordering. Equal true counts alone do not ensure equal released ratios.

B Quality of the Retrieved Public Mean

We bound the error of the retrieved public mean and quantify the distance increase caused by its dependence on the sum noise.

B.1 Error of the retrieved public mean

The distance $\|b_c - \mu_c\|_2$ controls PRDC's displacement bound in Theorem 2 and the magnitude of its signed shift in Corollary 3. Define the neighborhood retrieved with the unnoised sum and its mean by

$$J_c^0 = \mathrm{TopK}_{j \in [M]}\, S_c^\top a_j, \quad b_c^0 = \frac{1}{K} \sum_{j \in J_c^0} a_j. \tag{34}$$

We prove the bound in Section 4.2 by comparing each newly selected feature with one it replaces.

Proof of the public-mean bound. A Gaussian union bound over classes and pairs of bank features gives

$$Z_c^\top (a_j - a_i) \le 2\sigma_S \sqrt{L_M}\, \|a_j - a_i\|_2 \quad \text{for all } c, i, j,$$

except with probability at most $CM^2 e^{-2L_M}$. Pair a feature i removed from J_c^0 with a newly selected feature j. Noisy ranking and the equal feature norms give

$$\begin{aligned}\|a_j - \mu_c\|_2^2 - \|a_i - \mu_c\|_2^2 &= 2\mu_c^\top (a_i - a_j) \\ &\le \frac{4\sigma_S\sqrt{L_M}}{n_c}\|a_j - a_i\|_2 \\ &\le \frac{4\sigma_S\sqrt{L_M}}{n_c}\big(\|a_j - \mu_c\|_2 + \|a_i - \mu_c\|_2\big).\end{aligned}$$

Factoring the difference of squares gives $\|a_j - \mu_c\|_2 \le \|a_i - \mu_c\|_2 + 4\sigma_S\sqrt{L_M}/n_c$; the inequality also holds when both distances are zero. Pair each unchanged feature with itself. Averaging over the K pairs and applying the triangle and Cauchy–Schwarz inequalities yields

$$\begin{aligned}\|b_c - \mu_c\|_2 &\le \frac{1}{K}\sum_{i \in J_c^0} \|a_i - \mu_c\|_2 + \frac{4\sigma_S\sqrt{L_M}}{n_c} \\ &\le \sqrt{\frac{1}{K}\sum_{i \in J_c^0} \|a_i - \mu_c\|_2^2} + \frac{4\sigma_S\sqrt{L_M}}{n_c}.\end{aligned}$$

Expanding the average gives $K^{-1}\sum_{i \in J_c^0}\|a_i - \mu_c\|_2^2 = \|b_c^0 - \mu_c\|_2^2 + r^2 - \|b_c^0\|_2^2$.

To obtain the joint probability with Theorem 2, use $\Pr\{|G| > \sqrt{2L}\} \le e^{-L}$ for a standard normal G and $L \ge \log 10$. The scalar Gaussian and chi-square events in that theorem's proof then fail with probability at most $5Ce^{-L_0} + CMe^{-L_M}$. Adding the pairwise event above gives total failure probability at most

$$5Ce^{-L_0} + CMe^{-L_M} + CM^2 e^{-2L_M} = \frac{3\delta}{5} + \frac{\delta^2}{100C} < \delta. \qquad \square$$

Changes in the selected neighborhood. The same pairing also relates the change in the mean to the distances between replaced features:

$$\|b_c - b_c^0\|_2 \le \frac{1}{K}\sum_{(i,j)\,\text{paired}} \|a_j - a_i\|_2.$$

Here the sum includes only pairs with $i \in J_c^0 \setminus J_c$ and $j \in J_c \setminus J_c^0$. Thus, a small number of replacements, or replacements by nearby features, can leave the mean stable even when the selected set changes.

B.2 Distance increase from shared sum noise

The public mean and the noisy class mean share the same sum noise. We isolate how this dependence changes squared distances by fixing the training data, query x, and released count $\widehat{n}_c$, hence also α_c. For each sum-noise draw, define the reference prototype

$$m_c(Z_c) = \alpha_c \frac{S_c}{\widehat{n}_c} + (1 - \alpha_c) b_c(S_c + Z_c), \quad \widehat{\mu}_c^{\mathrm{PRDC}} = m_c(Z_c) + \frac{\alpha_c}{\widehat{n}_c} Z_c.$$

Here $b_c(S_c + Z_c)$ denotes the mean retrieved with that noisy sum. The reference retains the selected public mean but removes the direct sum-noise contribution to the private mean. Expanding the two distances gives

$$\begin{aligned}&\mathbb{E}_{Z_c}\big[\|x - \widehat{\mu}_c^{\mathrm{PRDC}}\|_2^2 - \|x - m_c(Z_c)\|_2^2\big] - d\alpha_c^2 v_c \\ &\quad = \frac{2\alpha_c(1 - \alpha_c)}{\widehat{n}_c}\mathbb{E}_{Z_c}\big[Z_c^\top b_c(S_c + Z_c)\big].\end{aligned} \tag{35}$$

Indeed, the quadratic term has expectation $d\alpha_c^2 v_c$ and the inner product with $x - \alpha_c S_c/\widehat{n}_c$ has expectation zero. The remaining term couples sum noise with the selected public mean. For exact top-K retrieval it is nonnegative in expectation: optimality of the noisy and noiseless selections gives $Z_c^\top (b_c - b_c^0) \ge S_c^\top (b_c^0 - b_c) \ge 0$, and $\mathbb{E}[Z_c^\top b_c^0] = 0$. This identity measures distance increase relative to $m_c(Z_c)$; the error from moving the clean mean μ_c is a separate contribution.

C EXPERIMENTAL PROTOCOL

Encoders and preprocessing. We use torchvision's ViT-B/16 with `IMAGENET1K_SWAG_E2E_V1` weights and their standard preprocessing: bicubic resizing of the shorter side to 384 pixels, a 384×384 center crop, and ImageNet channel normalization. Features are the final 768-dimensional class-token representation before the classifier. The ResNet-18 comparison uses `IMAGENET1K_V1` weights, bilinear resizing of the shorter side to 256 pixels, a 224×224 center crop, and the same channel normalization. Its features are the 512-dimensional output of global average pooling. Both encoders remain frozen, and feature vectors are normalized to unit length.

Training-set sampling and evaluation. The CIFAR datasets each contain 50,000 training and 10,000 test images; FOOD-101 contains 75,750 training and 25,250 test images. For each sampled training set, we randomly assign classes to size ranks and sample the prescribed number of images from each class. With C classes and imbalance ratio IR, rank c receives

$$n_c = \max\left\{1, \left\lfloor n_{\max} \, \mathrm{IR}^{-(c-1)/(C-1)} \right\rfloor\right\}, \tag{36}$$

where $n_{\max}$ is the number of examples per class in the original training set: 5,000 for CIFAR-10, 500 for CIFAR-100, and 750 for FOOD-101. All methods use the same 20 sampled training sets at each dataset–IR combination, with ten randomized runs per set and privacy level. PDC, PRDC, and their components share the noisy sums and counts.

We first average runs within each training set, then compute the mean and sample standard deviation across the 20 sets. This standard deviation reflects variation across training sets after averaging within-set randomness. For summaries over several budgets, we also average those budgets within each set before computing these statistics. Balanced accuracy is mean per-class recall, and Macro-F1 is mean classwise F1. Tail, middle, and head groups divide classes sorted by training count into three nearly equal-sized parts.

The balanced comparison uses the same total training size as IR 100: 12,406, 10,847, and 16,454 examples for CIFAR-10, CIFAR-100, and FOOD-101, respectively. These examples are distributed as evenly as possible across classes, so class counts differ by at most one. We independently sample 20 balanced training sets and use ten shared PDC/PRDC noise draws per set. Regularization-target, neighborhood-size, privacy-allocation, and ResNet-18 comparisons use the same repetition scheme on their stated dataset and budget ranges.

Public banks and development tasks. We encode public images with the target encoder and normalize the feature directions. The ImageNet bank contains 1,281,167 images, and the UECFood-256 bank contains 31,395 images. Both PRDC and DPPL-Public use the full ImageNet bank for the CIFAR datasets and the full UECFood bank for FOOD-101. PRDC ranks features by inner product and averages the top K, without renormalizing their mean. Bank labels are used only for development tasks.

For each source, we reserve one fifth of its classes for development tasks. The retrieval bank for these tasks uses the remaining classes, with up to 128 examples per class. We construct four tasks, each with 100 ImageNet classes or 50 UECFood classes. Each class provides 64 candidate support examples and 24 disjoint query examples. The support examples simulate private training data, and the queries evaluate the resulting classifier. We sample long-tailed support sets with counts from 2 to 64 and balanced support sets at the same total size. Mean public balanced accuracy over $\rho \in \{.005, .125, 2\}$ selects one K per source from $\{16, 64, 256, 2048\}$: 64 for ImageNet and 256 for UECFood. DP-FC uses the same public tasks, evaluated at all six budgets. DPPL-Mean and both SGD variants use the larger development tasks described in Appendix C.1.

Direct score baselines and components. Noisy mean uses $-\|x - Y_c/\widehat{n}_c\|_2^2$, with the same count lower bound as PDC and PRDC. Correction only adds $d\sigma_S^2/\widehat{n}_c^2$ to this score. The normalized-sum baseline uses $x^\top Y_c/\|Y_c\|_2$. Projection only uses PDC's prototype without distance correction. Retrieved mean only uses $-\|x - b_c\|_2^2$. PRDC without correction retains its regularized prototype and omits $d\alpha_c^2 v_c$. The neighborhood-scale variant replaces r^2/d by $(r^2 - \|b_c\|_2^2)/d$, keeping the same bank and K.

Table 3: DPPL-Mean coordinate-averaging widths selected by public validation BA. Each source–budget setting is fixed across target datasets and IRs.

Source	.005	.02	.125	.5	2	8
ImageNet	32	16	16	16	8	8
UECFood	32	16	16	8	8	4

Computation. Public features are encoded once and reused. Forming private class sums costs $O(Nd)$, while exact public retrieval costs $O(CMd)$ without an index. The resulting prototypes are cached, leaving $O(Cd)$ computation and no bank access per prediction. Appendix D.9 reports construction and inference times, bank-loading costs, and memory use.

C.1 Baseline settings

DPPL. We evaluate both mechanisms of Wahdany et al. (2025). DPPL-Mean receives each class's training quota q_c as a public, fixed denominator. For a class sum, its Gaussian noise has coordinate variance $r^2/(2q_c^2\rho)$ after division by q_c. This add/remove calibration differs from the original variance $2r^2/(n_c^2\rho)$; at $q_c = n_c$, it is one quarter as large. For optional dimension reduction, we average adjacent feature coordinates (mean pooling) with widths $\{1, 2, 4, 8, 16, 32, 64\}$. We additionally normalize these vectors to unit length, following our shared feature convention. Width one leaves the original features unchanged. We select one width per public source and budget using the same 10,000-example development tasks as DP-SGD, described below. Selection maximizes validation BA averaged over the four tasks and ten noise draws per task. The resulting widths, listed in Table 3, remain fixed across target datasets and IRs.

DPPL-Public uses the same full feature bank as PRDC and selects one public prototype per class. For clipping interval $[\ell, u]$, its utility is

$$U_{cj} = \sum_{i:y_i=c} \left[\mathrm{clip}(1 + x_i^\top a_j, \ell, u) - \ell\right]. \tag{37}$$

We sample with weights proportional to $\exp(\varepsilon_{\mathrm{pure}} U_{cj}/(u-\ell))$, following the original monotone exponential mechanism. Its bounded-range guarantee gives $\rho = \varepsilon_{\mathrm{pure}}^2/8$, so we use $\varepsilon_{\mathrm{pure}} = \sqrt{8\rho}$ (Wahdany et al., 2025; Cesar & Rogers, 2021).

We select DPPL-Public's clipping interval using the four public tasks per source described in Appendix C. For each candidate in $\{[0, 2], [1.1, 1.5], [1.1, 1.75]\}$, we average validation BA over the balanced and long-tailed conditions, $\rho \in \{.005, .125, 2\}$, and five runs per condition. Selection gives $[1.1, 1.5]$ for ImageNet and $[1.1, 1.75]$ for UECFood. These intervals remain fixed across target datasets, IRs, and all six budgets.

DP-LS and DP-FC. DP-LS uses unit feature clipping, $\alpha = 2$, and regularization 12. DP-FC uses the complete-gradient, one-versus-all head of Mehta et al. (2023); its covariance and gradient releases compose to the stated zCDP budget. Public tasks select 25 steps at $\rho = .005$ and five at the other budgets, with regularization 100 for $\rho \le .5$ and 1 for $\rho \in \{2, 8\}$. These settings are the same for both public sources. The learning rate is .2, gradient clipping is .005, and the bias parameter is -10.

DP-SGD. We train a linear softmax head with Poisson sampling, clipped per-example gradients, and momentum .9 (Abadi et al., 2016). The head's weights and biases are initialized to zero; the encoder remains frozen. For each public source, ImageNet or UECFood, we construct 256-class development tasks with 10,000 training examples and ten separate validation examples per class. Training counts follow geometric long-tailed distributions at IR 10 and 100, with two sampled tasks per IR.

We search nominal batch sizes $\{256, 512, 1024\}$, learning rates $\{.1, 1, 10\}$, epoch counts $\{25, 50, 100\}$, and clipping thresholds $\{.25, 1\}$, giving 54 settings. For each source and budget, selection maximizes validation BA averaged over the four tasks and two independent training runs per task. The selected batch size, learning rate, epochs, and clipping threshold are transferred unchanged:

Table 4: DP-SGD settings selected by public validation BA. Each source–budget setting is shared across target datasets and imbalance ratios; B is the nominal batch size.

Source	ρ	Epochs	Learning rate	Clip	Batch B
ImageNet	0.005	100	0.1	1	512
UECFood	0.005	100	0.1	1	512
ImageNet	0.02	100	0.1	1	256
UECFood	0.02	100	0.1	1	256
ImageNet	0.125	50	10	0.25	1024
UECFood	0.125	50	10	0.25	1024
ImageNet	0.5	100	10	0.25	1024
UECFood	0.5	100	10	0.25	1024
ImageNet	2	100	10	0.25	1024
UECFood	2	100	10	0.25	512
ImageNet	8	100	10	0.25	512
UECFood	8	100	10	0.25	512

ImageNet settings serve both CIFAR datasets and UECFood settings serve FOOD-101, for both IRs. Table 4 lists the selected settings.

For training size N and nominal batch size B, each epoch has $\lceil N/B \rceil$ updates with Poisson sampling rate $1/\lceil N/B \rceil$. The Opacus PRV accountant calibrates noise for this rate and the total update count to meet $(\varepsilon(\rho, 10^{-5}), 10^{-5})$-DP (Gopi et al., 2021). All trained heads follow the same 20-set, ten-run evaluation as the one-shot methods.

DP-SGD with logit adjustment. DP-SGD+LA subtracts $\tau \log(\widehat{n}_c / \sum_j \widehat{n}_j)$ from the learned logit of class c (Menon et al., 2021). The counts use the Gaussian mechanism in Equation (3), with $\sigma_n^2 = (1 + \sqrt{d})/(2\rho)$ and the same lower bound of one. We retain DP-SGD's publicly selected optimizer settings and recalibrate its gradient noise to include the independent count release. The PRV accountant composes all Poisson-sampled training updates with one unsubsampled Gaussian count release, meeting the same total $(\varepsilon(\rho, 10^{-5}), 10^{-5})$-DP target as DP-SGD.

For each public source and budget, we select $\tau \in \{0, .25, .5, 1, 2\}$ by validation BA on the same four 10,000-example tasks, using two independent head-and-count runs per task. Both sources select $\tau = 0, .25, .5, .5, 1, 1$ at the six budgets in increasing order. These strengths remain fixed across target datasets and IRs. All reported DP-SGD+LA results use this public selection, with the count-release cost included even when $\tau = 0$.

D Additional Experimental Results

D.1 Long-tailed results by privacy level

Tables 5– 10 report BA and Macro-F1 for all comparison methods at each privacy level, including the normalized-sum baseline defined in Appendix C. The budget-wise results show where the gains persist beyond the strong-privacy average in the main table. On CIFAR-100 and FOOD-101 at IR 100, PDC retains a clear BA advantage over Noisy mean even at $\rho = 8$. At IR 10, these classifiers largely converge as the budget grows. PRDC's additional gains are concentrated at smaller budgets, where the noisy means are less accurate.

At $\rho = 8$ and IR 10, DP-SGD+LA has higher BA than PDC and PRDC on all three datasets. At IR 100, PRDC remains ahead of DP-SGD+LA on CIFAR-100 and FOOD-101 at this budget.

D.2 Components and regularization targets

Table 11 supplements the main component curves with both metrics, averaged over all six budgets separately for each IR. It also compares PRDC's fixed scale r^2/d with the neighborhood-based scale $(r^2 - \|b_c\|_2^2)/d$, keeping the bank and K unchanged. The neighborhood scale puts more weight on the public mean when the retrieved features cluster tightly together. The fixed scale retains more

Table 5: CIFAR-10 BA (%) by privacy budget. Entries are mean ± SD across training sets.

IR	Method	.005	.02	.125	.5	2	8
100	DPPL-Mean	61.8 ± 1.1	74.7 ± 1.3	83.8 ± 0.7	85.0 ± 0.4	90.0 ± 0.2	90.1 ± 0.2
	DP-LS	50.6 ± 1.4	63.3 ± 1.3	78.5 ± 1.6	84.6 ± 2.0	86.3 ± 2.1	86.4 ± 2.4
	DP-FC	47.9 ± 3.9	59.9 ± 3.2	62.8 ± 4.0	63.2 ± 4.3	78.8 ± 1.2	84.6 ± 1.1
	DP-SGD	70.0 ± 1.8	78.7 ± 1.8	87.0 ± 1.1	89.3 ± 0.8	91.0 ± 0.7	91.2 ± 0.7
	DP-SGD+LA	70.0 ± 1.8	82.2 ± 1.6	89.9 ± 0.7	91.5 ± 0.4	94.3 ± 0.2	94.1 ± 0.2
	Noisy mean	47.4 ± 1.5	59.2 ± 2.0	74.7 ± 1.3	86.5 ± 1.5	92.9 ± 0.4	93.5 ± 0.2
	Normalized sum	53.8 ± 2.4	64.6 ± 1.7	80.3 ± 1.7	90.4 ± 1.1	93.4 ± 0.4	93.7 ± 0.2
	Correction only	44.6 ± 10.6	71.4 ± 6.8	91.6 ± 0.9	93.3 ± 0.3	93.5 ± 0.2	93.5 ± 0.2
	PDC	77.6 ± 1.2	86.9 ± 1.2	92.3 ± 0.5	93.3 ± 0.2	93.5 ± 0.2	93.5 ± 0.2
	DPPL-Public	76.8 ± 1.8	83.2 ± 1.3	85.5 ± 0.6	86.2 ± 0.5	86.4 ± 0.5	86.5 ± 0.5
	PRDC	80.2 ± 2.0	87.4 ± 1.0	91.7 ± 0.7	93.1 ± 0.4	93.5 ± 0.2	93.5 ± 0.2
10	DPPL-Mean	75.4 ± 0.1	85.5 ± 0.1	85.6 ± 0.1	85.6 ± 0.1	90.2 ± 0.1	90.2 ± 0.1
	DP-LS	70.7 ± 1.0	79.8 ± 0.5	88.7 ± 0.2	92.3 ± 0.2	93.6 ± 0.3	94.1 ± 0.3
	DP-FC	81.5 ± 3.0	90.7 ± 1.1	91.4 ± 0.8	91.2 ± 0.9	73.7 ± 0.3	85.4 ± 0.4
	DP-SGD	91.1 ± 0.3	92.5 ± 0.2	93.9 ± 0.1	94.4 ± 0.1	95.0 ± 0.2	94.8 ± 0.1
	DP-SGD+LA	91.1 ± 0.3	92.9 ± 0.2	94.3 ± 0.1	94.7 ± 0.1	95.5 ± 0.0	95.3 ± 0.1
	Noisy mean	85.4 ± 1.2	92.9 ± 0.3	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1
	Normalized sum	90.0 ± 1.2	93.3 ± 0.3	93.7 ± 0.1	93.8 ± 0.0	93.8 ± 0.0	93.7 ± 0.0
	Correction only	93.1 ± 0.1	93.5 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1
	PDC	93.1 ± 0.1	93.5 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1
	DPPL-Public	86.2 ± 0.4	86.4 ± 0.3	86.5 ± 0.3	86.5 ± 0.3	86.5 ± 0.3	86.5 ± 0.3
	PRDC	92.9 ± 0.3	93.4 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1

Table 6: CIFAR-10 Macro-F1 (%) by privacy budget. Entries are mean ± SD across training sets.

IR	Method	.005	.02	.125	.5	2	8
100	DPPL-Mean	56.9 ± 1.5	71.6 ± 1.7	83.7 ± 0.8	85.0 ± 0.4	90.0 ± 0.2	90.1 ± 0.2
	DP-LS	46.2 ± 1.4	59.7 ± 1.5	76.7 ± 2.2	83.8 ± 2.4	85.9 ± 2.5	86.0 ± 2.8
	DP-FC	37.3 ± 4.3	51.4 ± 4.2	55.4 ± 5.1	56.2 ± 5.5	79.0 ± 1.4	84.6 ± 1.2
	DP-SGD	64.9 ± 2.2	76.6 ± 2.4	86.7 ± 1.2	89.2 ± 0.8	91.0 ± 0.8	91.2 ± 0.7
	DP-SGD+LA	64.9 ± 2.2	80.9 ± 2.2	89.8 ± 0.7	91.5 ± 0.4	94.3 ± 0.2	94.1 ± 0.2
	Noisy mean	35.6 ± 1.7	49.2 ± 2.7	68.7 ± 2.0	85.1 ± 2.2	92.9 ± 0.5	93.5 ± 0.2
	Normalized sum	42.9 ± 2.8	56.0 ± 2.5	76.8 ± 2.3	90.2 ± 1.3	93.4 ± 0.4	93.7 ± 0.2
	Correction only	40.7 ± 11.3	70.7 ± 7.2	91.5 ± 1.0	93.3 ± 0.2	93.5 ± 0.2	93.6 ± 0.1
	PDC	75.2 ± 1.5	86.6 ± 1.3	92.3 ± 0.4	93.3 ± 0.2	93.5 ± 0.2	93.6 ± 0.1
	DPPL-Public	74.8 ± 2.2	83.0 ± 1.6	85.4 ± 0.6	86.1 ± 0.5	86.3 ± 0.5	86.4 ± 0.5
	PRDC	78.1 ± 2.6	87.0 ± 1.3	91.6 ± 0.7	93.1 ± 0.4	93.5 ± 0.2	93.6 ± 0.2
10	DPPL-Mean	75.2 ± 0.1	85.5 ± 0.1	85.6 ± 0.1	85.6 ± 0.1	90.2 ± 0.1	90.2 ± 0.1
	DP-LS	70.4 ± 1.2	79.8 ± 0.5	88.7 ± 0.2	92.3 ± 0.2	93.6 ± 0.3	94.1 ± 0.3
	DP-FC	80.2 ± 3.8	90.6 ± 1.1	91.3 ± 0.9	91.2 ± 0.9	73.8 ± 0.3	85.5 ± 0.3
	DP-SGD	91.0 ± 0.3	92.5 ± 0.2	93.9 ± 0.2	94.4 ± 0.1	95.0 ± 0.2	94.8 ± 0.1
	DP-SGD+LA	91.0 ± 0.3	92.9 ± 0.2	94.3 ± 0.1	94.7 ± 0.1	95.5 ± 0.0	95.3 ± 0.1
	Noisy mean	84.3 ± 1.7	92.9 ± 0.3	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1
	Normalized sum	89.9 ± 1.3	93.3 ± 0.3	93.7 ± 0.1	93.8 ± 0.0	93.8 ± 0.0	93.8 ± 0.0
	Correction only	93.1 ± 0.1	93.5 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1
	PDC	93.2 ± 0.1	93.5 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1
	DPPL-Public	86.1 ± 0.4	86.3 ± 0.3	86.4 ± 0.3	86.4 ± 0.3	86.4 ± 0.3	86.4 ± 0.3
	PRDC	92.9 ± 0.3	93.4 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1	93.6 ± 0.1

private information and gives higher mean BA in all six dataset–IR combinations. The comparison distinguishes the concentration of a neighborhood from its usefulness as a regularization target.

Distance correction. Table 12 reports how often adding correction to the same prototype repairs an incorrect prediction or introduces an error. The difference between these rates is the net accuracy gain. Averaged over $\rho \in \{.005, .02, .125\}$, repairs exceed harms in all six dataset–IR combinations for both methods. On CIFAR-100 at IR 100, PDC's correction repairs 19.8% of examples and harms

Table 7: CIFAR-100 BA (%) by privacy budget. Entries are mean ± SD across training sets.

IR	Method	.005	.02	.125	.5	2	8
100	DPPL-Mean	20.2 ± 0.8	30.0 ± 1.0	41.0 ± 1.0	48.3 ± 0.8	57.3 ± 0.7	64.1 ± 0.5
	DP-LS	7.6 ± 0.5	16.8 ± 0.8	31.7 ± 1.0	41.8 ± 1.0	48.3 ± 0.9	50.8 ± 0.9
	DP-FC	19.9 ± 0.9	26.3 ± 1.2	29.6 ± 1.3	31.1 ± 1.4	45.8 ± 1.0	50.3 ± 0.8
	DP-SGD	26.8 ± 1.0	34.9 ± 1.0	45.8 ± 1.0	53.1 ± 0.8	57.2 ± 0.9	61.7 ± 0.5
	DP-SGD+LA	26.8 ± 1.0	36.1 ± 1.0	47.8 ± 0.9	55.8 ± 0.6	64.5 ± 0.6	68.4 ± 0.6
	Noisy mean	7.9 ± 0.7	15.8 ± 0.7	28.4 ± 1.1	38.4 ± 1.2	48.3 ± 1.0	57.2 ± 0.7
	Normalized sum	15.3 ± 1.1	21.8 ± 0.9	32.8 ± 1.3	42.2 ± 1.2	51.5 ± 0.9	60.0 ± 0.6
	Correction only	1.0 ± 0.0	1.0 ± 0.0	1.1 ± 0.1	1.8 ± 1.0	39.0 ± 7.3	70.8 ± 1.1
	PDC	25.5 ± 1.0	35.7 ± 1.1	48.7 ± 0.8	58.1 ± 0.7	66.5 ± 0.5	72.0 ± 0.5
	DPPL-Public	28.7 ± 1.2	37.2 ± 1.3	47.3 ± 1.0	53.9 ± 0.7	58.2 ± 0.6	59.9 ± 0.5
	PRDC	30.8 ± 1.1	40.5 ± 1.0	53.1 ± 0.8	61.9 ± 0.4	68.7 ± 0.5	72.1 ± 0.4
10	DPPL-Mean	33.0 ± 0.7	51.5 ± 0.5	61.9 ± 0.3	63.5 ± 0.2	70.9 ± 0.2	71.1 ± 0.2
	DP-LS	11.8 ± 0.6	26.9 ± 0.9	50.8 ± 0.5	63.9 ± 0.5	69.7 ± 0.6	71.4 ± 0.7
	DP-FC	34.6 ± 1.2	44.9 ± 1.2	50.5 ± 1.2	52.3 ± 1.1	64.5 ± 0.6	68.0 ± 0.5
	DP-SGD	45.4 ± 0.8	58.3 ± 0.4	69.8 ± 0.4	74.0 ± 0.3	77.0 ± 0.3	77.2 ± 0.2
	DP-SGD+LA	45.2 ± 0.8	59.6 ± 0.4	71.2 ± 0.3	75.0 ± 0.3	79.0 ± 0.2	78.8 ± 0.2
	Noisy mean	13.7 ± 1.1	28.1 ± 1.1	50.7 ± 0.9	67.4 ± 0.6	75.2 ± 0.4	76.5 ± 0.2
	Normalized sum	26.4 ± 1.2	37.5 ± 1.5	56.7 ± 0.7	70.1 ± 0.7	75.6 ± 0.4	76.6 ± 0.2
	Correction only	1.2 ± 0.2	20.6 ± 4.8	72.4 ± 0.4	75.8 ± 0.2	76.4 ± 0.2	76.6 ± 0.2
	PDC	43.0 ± 1.0	59.8 ± 0.7	73.0 ± 0.3	75.8 ± 0.2	76.4 ± 0.2	76.6 ± 0.2
	DPPL-Public	48.1 ± 0.8	56.1 ± 0.5	60.6 ± 0.5	62.1 ± 0.3	62.7 ± 0.3	62.9 ± 0.4
	PRDC	52.9 ± 0.7	64.9 ± 0.4	73.1 ± 0.3	75.6 ± 0.2	76.3 ± 0.2	76.5 ± 0.1

Table 8: CIFAR-100 Macro-F1 (%) by privacy budget. Entries are mean ± SD across training sets.

IR	Method	.005	.02	.125	.5	2	8
100	DPPL-Mean	12.7 ± 0.7	21.1 ± 1.0	33.6 ± 0.9	43.0 ± 0.9	52.9 ± 0.9	62.7 ± 0.8
	DP-LS	6.1 ± 0.4	12.7 ± 0.7	24.1 ± 1.0	33.6 ± 1.2	40.6 ± 1.1	43.5 ± 1.2
	DP-FC	11.8 ± 0.9	17.3 ± 1.4	20.4 ± 1.7	21.9 ± 1.8	38.6 ± 1.2	44.1 ± 0.9
	DP-SGD	17.1 ± 0.9	25.3 ± 1.1	37.4 ± 1.1	46.6 ± 0.9	51.1 ± 1.0	57.7 ± 0.7
	DP-SGD+LA	17.1 ± 0.9	27.0 ± 1.1	40.7 ± 1.0	50.5 ± 0.6	61.3 ± 0.8	66.6 ± 0.8
	Noisy mean	4.2 ± 0.5	9.1 ± 0.7	19.1 ± 1.4	28.9 ± 1.5	40.0 ± 1.2	51.1 ± 0.9
	Normalized sum	8.6 ± 0.9	13.5 ± 1.1	23.0 ± 1.7	33.0 ± 1.5	43.7 ± 1.2	54.7 ± 0.9
	Correction only	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.1	0.9 ± 1.3	42.2 ± 7.2	70.7 ± 0.9
	PDC	18.5 ± 1.0	28.9 ± 0.9	43.8 ± 0.7	55.3 ± 0.6	65.9 ± 0.6	72.1 ± 0.5
	DPPL-Public	23.3 ± 1.2	31.9 ± 1.3	43.8 ± 1.0	52.2 ± 0.7	57.4 ± 0.7	59.3 ± 0.5
	PRDC	25.5 ± 1.3	34.7 ± 1.2	48.4 ± 1.0	58.9 ± 0.6	67.8 ± 0.7	72.0 ± 0.5
10	DPPL-Mean	29.0 ± 0.7	48.9 ± 0.5	61.7 ± 0.3	63.4 ± 0.2	71.0 ± 0.2	71.2 ± 0.2
	DP-LS	10.2 ± 0.5	23.2 ± 0.8	46.7 ± 0.7	61.2 ± 0.7	67.7 ± 0.8	69.5 ± 0.9
	DP-FC	27.0 ± 1.3	38.0 ± 1.2	44.1 ± 1.2	46.3 ± 1.2	63.0 ± 0.9	67.0 ± 0.7
	DP-SGD	39.1 ± 0.7	55.0 ± 0.6	68.7 ± 0.6	73.4 ± 0.5	76.6 ± 0.4	77.0 ± 0.3
	DP-SGD+LA	39.0 ± 0.7	56.7 ± 0.6	70.4 ± 0.5	74.6 ± 0.4	78.8 ± 0.2	78.7 ± 0.2
	Noisy mean	8.6 ± 0.9	20.4 ± 1.3	44.1 ± 1.1	65.6 ± 1.0	75.3 ± 0.4	76.6 ± 0.2
	Normalized sum	18.8 ± 1.5	29.6 ± 1.8	51.7 ± 0.9	68.8 ± 1.1	75.5 ± 0.4	76.7 ± 0.2
	Correction only	0.1 ± 0.1	22.7 ± 5.2	72.8 ± 0.4	76.0 ± 0.2	76.6 ± 0.2	76.8 ± 0.2
	PDC	38.6 ± 0.9	58.9 ± 0.6	73.5 ± 0.3	76.0 ± 0.2	76.6 ± 0.2	76.8 ± 0.2
	DPPL-Public	46.5 ± 0.8	55.8 ± 0.6	60.3 ± 0.5	61.6 ± 0.4	62.1 ± 0.4	62.2 ± 0.4
	PRDC	50.7 ± 0.6	64.4 ± 0.6	73.4 ± 0.4	75.9 ± 0.2	76.6 ± 0.2	76.8 ± 0.2

2.0%; PRDC's correction repairs 10.5% and harms 2.6%. Distance correction changes predictions substantially even after public regularization, with more errors removed than introduced.

Regularization targets. Table 13 gives both metrics for the public-target controls in Table 2 and adds a class-agnostic target: the arithmetic mean of the full public bank, without rescaling. All classes use this same mean, with PRDC's interpolation weights, correction, and private releases unchanged. On CIFAR-100 and FOOD-101, the global mean improves on the zero target, but remains below

Table 9: FOOD-101 BA (%) by privacy budget. Entries are mean ± SD across training sets.

IR	Method	.005	.02	.125	.5	2	8
100	DPPL-Mean	25.3 ± 0.9	35.6 ± 1.0	46.6 ± 1.1	54.8 ± 0.9	62.8 ± 0.4	69.5 ± 0.7
	DP-LS	10.2 ± 0.8	19.2 ± 0.9	33.6 ± 1.0	42.3 ± 1.3	46.8 ± 1.5	48.2 ± 1.6
	DP-FC	20.2 ± 1.3	26.0 ± 1.2	28.8 ± 1.0	29.7 ± 1.0	43.4 ± 1.3	47.3 ± 1.5
	DP-SGD	31.4 ± 0.9	39.7 ± 1.2	50.8 ± 1.2	58.3 ± 1.0	63.1 ± 0.7	66.6 ± 0.8
	DP-SGD+LA	31.4 ± 0.9	41.3 ± 1.2	53.7 ± 1.1	62.0 ± 0.8	69.3 ± 0.8	74.6 ± 0.8
	Noisy mean	12.6 ± 1.0	20.9 ± 1.2	33.9 ± 0.9	43.9 ± 1.2	53.8 ± 1.1	62.9 ± 0.5
	Normalized sum	17.9 ± 1.5	25.0 ± 1.4	37.0 ± 1.4	46.6 ± 1.2	55.9 ± 1.0	64.8 ± 0.6
	Correction only	1.0 ± 0.0	1.0 ± 0.0	1.2 ± 0.3	14.2 ± 4.8	65.6 ± 2.4	74.8 ± 0.5
	PDC	30.7 ± 1.0	41.1 ± 1.2	54.0 ± 1.0	63.3 ± 0.6	71.1 ± 0.8	75.1 ± 0.5
	DPPL-Public	24.9 ± 1.1	32.1 ± 1.4	40.4 ± 0.8	45.2 ± 0.6	48.4 ± 0.5	49.8 ± 0.5
	PRDC	35.1 ± 1.1	45.6 ± 1.1	58.6 ± 0.6	66.7 ± 0.7	72.4 ± 0.7	75.2 ± 0.4
10	DPPL-Mean	41.5 ± 0.5	59.6 ± 0.4	65.8 ± 0.2	72.4 ± 0.1	72.7 ± 0.1	76.3 ± 0.1
	DP-LS	14.7 ± 0.9	29.1 ± 0.8	52.1 ± 0.6	64.6 ± 0.8	70.4 ± 1.0	71.9 ± 1.1
	DP-FC	34.6 ± 1.4	44.3 ± 1.6	48.5 ± 1.7	49.7 ± 1.7	62.3 ± 1.0	67.0 ± 0.9
	DP-SGD	53.0 ± 0.8	65.7 ± 0.7	76.9 ± 0.5	81.0 ± 0.4	81.6 ± 0.2	83.6 ± 0.3
	DP-SGD+LA	53.0 ± 0.8	67.3 ± 0.6	78.5 ± 0.5	82.2 ± 0.3	83.1 ± 0.2	85.2 ± 0.2
	Noisy mean	21.3 ± 1.3	37.0 ± 1.4	60.3 ± 0.7	74.3 ± 0.6	77.7 ± 0.2	78.0 ± 0.1
	Normalized sum	30.4 ± 1.4	43.4 ± 1.5	63.6 ± 0.7	74.8 ± 0.6	77.6 ± 0.2	77.9 ± 0.1
	Correction only	6.2 ± 1.9	57.5 ± 2.0	76.4 ± 0.2	77.8 ± 0.1	78.0 ± 0.1	78.1 ± 0.1
	PDC	52.0 ± 0.7	67.4 ± 0.5	76.5 ± 0.2	77.8 ± 0.1	78.0 ± 0.1	78.1 ± 0.1
	DPPL-Public	40.3 ± 0.7	47.0 ± 0.4	50.4 ± 0.3	51.1 ± 0.3	51.3 ± 0.4	51.3 ± 0.4
	PRDC	58.5 ± 0.5	69.6 ± 0.6	76.2 ± 0.2	77.6 ± 0.1	78.0 ± 0.1	78.1 ± 0.1

Table 10: FOOD-101 Macro-F1 (%) by privacy budget. Entries are mean ± SD across training sets.

IR	Method	.005	.02	.125	.5	2	8
100	DPPL-Mean	17.3 ± 0.7	26.4 ± 1.1	39.7 ± 1.2	48.6 ± 0.9	59.6 ± 0.7	67.7 ± 1.1
	DP-LS	7.8 ± 0.6	14.2 ± 0.7	25.2 ± 1.0	33.2 ± 1.4	38.0 ± 1.7	39.5 ± 1.8
	DP-FC	11.4 ± 1.1	16.4 ± 1.0	19.1 ± 1.0	20.1 ± 1.0	35.4 ± 1.6	39.8 ± 1.6
	DP-SGD	21.3 ± 0.8	30.0 ± 1.3	42.5 ± 1.4	52.1 ± 1.2	58.7 ± 1.1	62.8 ± 1.2
	DP-SGD+LA	21.3 ± 0.8	31.9 ± 1.2	46.6 ± 1.1	56.9 ± 1.0	67.2 ± 1.0	73.1 ± 1.0
	Noisy mean	6.2 ± 0.7	11.8 ± 1.0	23.6 ± 0.9	34.2 ± 1.4	45.9 ± 1.3	57.9 ± 0.6
	Normalized sum	9.6 ± 1.1	15.3 ± 1.2	26.9 ± 1.6	37.5 ± 1.4	48.7 ± 1.3	60.6 ± 1.0
	Correction only	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.3	15.1 ± 5.4	66.5 ± 1.9	74.6 ± 0.6
	PDC	22.3 ± 0.9	33.4 ± 1.1	48.9 ± 0.9	60.6 ± 0.8	70.6 ± 0.9	75.1 ± 0.6
	DPPL-Public	19.6 ± 1.0	27.1 ± 1.3	36.6 ± 0.8	42.5 ± 0.8	46.5 ± 0.6	48.2 ± 0.6
	PRDC	28.2 ± 1.2	39.5 ± 1.2	54.8 ± 0.6	65.0 ± 0.9	72.0 ± 0.8	75.2 ± 0.5
10	DPPL-Mean	38.9 ± 0.4	58.2 ± 0.6	65.5 ± 0.2	72.3 ± 0.1	72.6 ± 0.1	76.3 ± 0.1
	DP-LS	12.4 ± 0.7	25.0 ± 0.7	47.4 ± 0.8	61.3 ± 1.3	67.8 ± 1.5	69.5 ± 1.6
	DP-FC	26.3 ± 1.6	36.4 ± 1.9	41.0 ± 2.0	42.4 ± 2.1	60.0 ± 1.5	65.3 ± 1.4
	DP-SGD	47.4 ± 0.9	63.1 ± 1.0	76.2 ± 0.7	80.8 ± 0.5	81.4 ± 0.3	83.5 ± 0.4
	DP-SGD+LA	47.4 ± 0.9	65.1 ± 1.0	78.0 ± 0.6	82.0 ± 0.4	83.0 ± 0.2	85.2 ± 0.2
	Noisy mean	13.3 ± 0.9	28.0 ± 1.5	55.4 ± 0.9	73.9 ± 0.8	77.8 ± 0.2	78.2 ± 0.1
	Normalized sum	21.8 ± 1.2	35.4 ± 1.6	60.1 ± 1.1	74.4 ± 0.7	77.6 ± 0.2	78.0 ± 0.1
	Correction only	5.7 ± 2.1	59.1 ± 1.8	76.5 ± 0.2	77.9 ± 0.1	78.1 ± 0.1	78.2 ± 0.1
	PDC	48.0 ± 0.7	66.8 ± 0.7	76.6 ± 0.2	77.9 ± 0.1	78.1 ± 0.1	78.2 ± 0.1
	DPPL-Public	37.3 ± 0.7	45.1 ± 0.6	48.8 ± 0.4	49.4 ± 0.4	49.6 ± 0.5	49.7 ± 0.5
	PRDC	55.9 ± 0.7	69.1 ± 0.7	76.3 ± 0.2	77.7 ± 0.1	78.1 ± 0.1	78.2 ± 0.1

PDC in both metrics at both IRs. Class-specific retrieval adds 7.3 and 8.3 pp in BA over the global mean at IR 100, respectively. The public bank's overall center is useful, but does not capture the full benefit of a class-specific target.

Table 11: Component BA and Macro-F1 (%). We average all six budgets within each training set, separately for each IR. Entries are mean ± SD across training sets.

Method	CIFAR-10		CIFAR-100		FOOD-101	
	BA	Macro-F1	BA	Macro-F1	BA	Macro-F1
IR 100						
Noisy mean	75.7 ± 0.5	70.9 ± 0.7	32.7 ± 0.5	25.4 ± 0.6	38.0 ± 0.7	29.9 ± 0.6
Projection only	75.9 ± 0.5	71.1 ± 0.8	33.5 ± 0.6	26.3 ± 0.7	38.5 ± 0.7	30.5 ± 0.6
Correction only	81.3 ± 2.2	80.5 ± 2.3	19.1 ± 1.3	19.0 ± 1.3	26.3 ± 1.0	26.1 ± 1.0
PDC	89.5 ± 0.3	89.1 ± 0.4	51.1 ± 0.3	47.4 ± 0.3	55.9 ± 0.4	51.8 ± 0.4
Normalized sum	79.4 ± 0.6	75.5 ± 0.8	37.3 ± 0.6	29.4 ± 0.7	41.2 ± 0.7	33.1 ± 0.6
Retrieved mean only	84.8 ± 0.7	85.4 ± 0.7	48.7 ± 0.3	49.6 ± 0.3	48.0 ± 0.4	45.3 ± 0.4
PRDC, no correction	84.6 ± 1.0	83.8 ± 1.1	45.5 ± 0.4	43.3 ± 0.5	46.4 ± 0.5	42.6 ± 0.6
Neighborhood scale	89.1 ± 0.6	88.8 ± 0.6	53.6 ± 0.3	51.0 ± 0.4	58.2 ± 0.4	55.0 ± 0.4
PRDC	89.9 ± 0.5	89.5 ± 0.5	54.5 ± 0.3	51.2 ± 0.4	58.9 ± 0.4	55.8 ± 0.4
IR 10						
Noisy mean	92.1 ± 0.2	91.9 ± 0.3	51.9 ± 0.3	48.4 ± 0.3	58.1 ± 0.4	54.4 ± 0.3
Projection only	92.1 ± 0.2	92.0 ± 0.3	53.1 ± 0.3	49.8 ± 0.4	58.7 ± 0.4	55.1 ± 0.3
Correction only	93.5 ± 0.1	93.5 ± 0.1	53.8 ± 0.8	54.2 ± 0.9	62.3 ± 0.6	62.6 ± 0.6
PDC	93.5 ± 0.1	93.5 ± 0.1	67.4 ± 0.2	66.7 ± 0.1	71.6 ± 0.2	70.9 ± 0.2
Normalized sum	93.1 ± 0.2	93.0 ± 0.2	57.2 ± 0.3	53.5 ± 0.4	61.3 ± 0.4	57.9 ± 0.4
Retrieved mean only	87.7 ± 0.1	87.9 ± 0.1	60.8 ± 0.2	63.2 ± 0.2	57.9 ± 0.2	57.2 ± 0.2
PRDC, no correction	92.7 ± 0.2	92.7 ± 0.2	64.3 ± 0.3	64.4 ± 0.3	66.1 ± 0.5	64.9 ± 0.5
Neighborhood scale	93.1 ± 0.2	93.2 ± 0.2	69.0 ± 0.1	69.2 ± 0.2	72.5 ± 0.2	71.9 ± 0.3
PRDC	93.4 ± 0.1	93.4 ± 0.1	69.9 ± 0.1	69.6 ± 0.2	73.0 ± 0.2	72.5 ± 0.2

Table 12: Prediction changes from distance correction (%), averaged over $\rho \in \{.005, .02, .125\}$. Repairs change an incorrect prediction to a correct one; harms make the reverse change. Each comparison keeps the prototype fixed and includes all competing classes. Entries are mean ± SD across training sets.

Dataset	IR	PDC		PRDC	
		Repair	Harm	Repair	Harm
CIFAR-10	100	27.4 ± 1.4	2.4 ± 0.4	10.4 ± 1.4	1.2 ± 0.3
	10	3.5 ± 0.5	0.8 ± 0.2	2.1 ± 0.3	0.6 ± 0.2
CIFAR-100	100	19.8 ± 0.7	2.0 ± 0.2	10.5 ± 0.4	2.6 ± 0.2
	10	29.0 ± 0.8	3.6 ± 0.2	13.1 ± 0.3	4.0 ± 0.2
FOOD-101	100	21.0 ± 0.9	2.3 ± 0.2	16.9 ± 0.6	3.8 ± 0.2
	10	28.4 ± 1.0	3.7 ± 0.2	17.5 ± 0.6	4.6 ± 0.2

D.3 Noisy retrieval and predictions

We examine how sum noise changes the retrieved mean, the distance scores, and the resulting predictions.

Retrieved directions. Table 14 compares means retrieved with noisy and unnoised class sums. We subtract the global bank mean from both before computing their cosine similarity. The centered cosine similarity measures how noise changes the retrieved direction relative to the global bank mean. Head-class directions are already well aligned at the smallest budget. Tail-class agreement improves over a wider budget range: on CIFAR-100, for example, it rises from 0.02 at $\rho = .005$ to 0.97 at $\rho = 8$. The class-size dependence is consistent with the retrieval perturbation scale σ_S/n_c in Appendix B.1.

Distance increase from noisy retrieval. We measure the distance increase relative to the reference prototype $m_c(Z_c)$ in Appendix B.2, which retains the retrieved public mean but removes the direct sum-noise component. Equation (35) separates this increase into the correction $d\alpha_c^2 v_c$ and a selection-dependent residual.

Table 13: Regularization-target controls: BA and Macro-F1 (%), averaged over $\rho \in \{.005, .02, .125\}$. The four PRDC variants share weights, correction, and private releases; only the target changes. Entries are mean $\pm$ SD across training sets.

Method	CIFAR-10		CIFAR-100		FOOD-101	
	BA	Macro-F1	BA	Macro-F1	BA	Macro-F1
IR 100						
PDC	85.61 ± 0.64	84.70 ± 0.77	36.63 ± 0.76	30.37 ± 0.70	41.93 ± 0.88	34.85 ± 0.88
Zero target	79.23 ± 0.85	75.17 ± 1.21	30.07 ± 0.78	20.34 ± 0.84	34.50 ± 0.86	24.38 ± 0.82
Global mean	84.35 ± 0.87	82.85 ± 1.17	34.21 ± 0.84	24.77 ± 0.84	38.09 ± 0.83	28.83 ± 0.91
Wrong-class target	71.41 ± 1.27	67.48 ± 1.25	21.07 ± 0.64	15.62 ± 0.51	28.89 ± 0.81	25.15 ± 0.79
PRDC	86.41 ± 0.86	85.57 ± 1.04	41.46 ± 0.79	36.17 ± 0.90	46.42 ± 0.78	40.81 ± 0.75
IR 10						
PDC	93.40 ± 0.07	93.41 ± 0.07	58.63 ± 0.38	57.00 ± 0.25	65.30 ± 0.28	63.82 ± 0.31
Zero target	93.42 ± 0.08	93.44 ± 0.08	51.07 ± 0.50	45.38 ± 0.52	57.80 ± 0.50	52.96 ± 0.55
Global mean	93.39 ± 0.11	93.40 ± 0.11	55.76 ± 0.40	51.27 ± 0.35	61.26 ± 0.42	57.77 ± 0.46
Wrong-class target	93.39 ± 0.09	93.40 ± 0.08	39.14 ± 0.51	35.87 ± 0.46	53.04 ± 0.50	51.01 ± 0.46
PRDC	93.28 ± 0.12	93.30 ± 0.13	63.61 ± 0.21	62.85 ± 0.26	68.10 ± 0.35	67.08 ± 0.46

Table 14: Cosine similarity between means retrieved with noisy and unnoised sums at IR 100, after subtracting the global bank mean. Tail, middle, and head are class-count thirds. Entries are mean $\pm$ SD across training sets.

Dataset	ρ	Tail	Middle	Head
CIFAR-10	0.005	0.65 ± 0.05	0.98 ± 0.01	1.00 ± 0.00
	0.02	0.88 ± 0.04	1.00 ± 0.00	1.00 ± 0.00
	0.125	0.98 ± 0.02	1.00 ± 0.00	1.00 ± 0.00
	0.5	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	2	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	8	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
CIFAR-100	0.005	0.02 ± 0.01	0.33 ± 0.02	0.93 ± 0.02
	0.02	0.08 ± 0.01	0.66 ± 0.02	0.98 ± 0.00
	0.125	0.35 ± 0.02	0.94 ± 0.01	1.00 ± 0.00
	0.5	0.67 ± 0.02	0.98 ± 0.00	1.00 ± 0.00
	2	0.91 ± 0.02	0.99 ± 0.00	1.00 ± 0.00
	8	0.97 ± 0.01	1.00 ± 0.00	1.00 ± 0.00
FOOD-101	0.005	0.19 ± 0.01	0.69 ± 0.02	0.97 ± 0.00
	0.02	0.36 ± 0.02	0.88 ± 0.01	0.99 ± 0.00
	0.125	0.69 ± 0.02	0.98 ± 0.00	1.00 ± 0.00
	0.5	0.88 ± 0.01	0.99 ± 0.00	1.00 ± 0.00
	2	0.97 ± 0.01	1.00 ± 0.00	1.00 ± 0.00
	8	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

We evaluate this increase on all three datasets, both IRs, and all six budgets. For each training set and budget, we fix one released count vector and use ten independent sum-noise draws, each with its retrieved public mean b_c. Distances are averaged over the test queries of each class and then over classes in each size group. We average the ten draws within each of the 20 training sets and report variation across sets. As a control, we replace b_c by the fixed noiseless-retrieval mean b_c^0 while retaining the same count and sum noise. The control's expected residual is zero; its measured mean has magnitude below 0.00034 in every dataset–IR–budget and size group.

Table 15 reports the correction, observed distance increase, and their difference at IR 100 for two small budgets. At $\rho = .125$, the tail residual is 0.0389 on CIFAR-100 and 0.0233 on FOOD-101, compared with corrections of 0.0557 and 0.0884. The corresponding head residuals are below 0.0006. Noisy retrieval contributes an appreciable distance shift in the tail, alongside the direct sum-noise term targeted by the correction. Its effect on predictions also depends on the relative shifts across competing classes, as in Corollary 3.

Table 15: PRDC distance increase relative to $m_c(Z_c)$ at IR 100. Correction is $d\alpha_c^2 v_c$; residual is the observed increase minus this correction. Entries are mean ± SD across training sets, in squared-distance units.

Dataset	ρ	Group	Correction	Increase	Residual
CIFAR-10	.005	Tail	0.1346 ± 0.0307	0.1739 ± 0.0365	0.0392 ± 0.0085
		Middle	0.1421 ± 0.0085	0.1463 ± 0.0094	0.0042 ± 0.0020
		Head	0.0118 ± 0.0005	0.0118 ± 0.0005	0.0000 ± 0.0001
	.125	Tail	0.1534 ± 0.0087	0.1611 ± 0.0090	0.0077 ± 0.0031
		Middle	0.0102 ± 0.0003	0.0102 ± 0.0004	0.0000 ± 0.0001
		Head	0.0005 ± 0.0000	0.0005 ± 0.0000	0.0000 ± 0.0000
CIFAR-100	.005	Tail	0.0201 ± 0.0051	0.0434 ± 0.0087	0.0233 ± 0.0039
		Middle	0.0582 ± 0.0096	0.0986 ± 0.0138	0.0404 ± 0.0046
		Head	0.1997 ± 0.0088	0.2215 ± 0.0095	0.0218 ± 0.0016
	.125	Tail	0.0557 ± 0.0082	0.0945 ± 0.0124	0.0389 ± 0.0048
		Middle	0.1982 ± 0.0073	0.2178 ± 0.0074	0.0196 ± 0.0013
		Head	0.0650 ± 0.0016	0.0655 ± 0.0016	0.0006 ± 0.0003
FOOD-101	.005	Tail	0.0225 ± 0.0053	0.0378 ± 0.0077	0.0153 ± 0.0027
		Middle	0.0881 ± 0.0087	0.1125 ± 0.0106	0.0244 ± 0.0024
		Head	0.1964 ± 0.0071	0.2041 ± 0.0073	0.0076 ± 0.0008
	.125	Tail	0.0884 ± 0.0108	0.1117 ± 0.0126	0.0233 ± 0.0024
		Middle	0.1946 ± 0.0040	0.2011 ± 0.0043	0.0065 ± 0.0008
		Head	0.0317 ± 0.0006	0.0318 ± 0.0006	0.0001 ± 0.0001

Table 16: Tail recall (%) at IR 100 and $\rho = .125$. Entries are mean ± SD across training sets. The noiseless-retrieval diagnostic changes only PRDC's public mean to b_c^0, retrieved with the unnoised private sum S_c.

Classifier	CIFAR-10	CIFAR-100	FOOD-101
Clean mean	93.3 ± 1.4	65.0 ± 2.4	70.7 ± 3.7
PDC	91.9 ± 1.9	7.5 ± 1.8	13.6 ± 2.2
PRDC	87.6 ± 2.4	14.5 ± 2.8	26.8 ± 3.4
PRDC, b_c^0	88.2 ± 2.5	48.0 ± 2.7	55.1 ± 4.3

Predictions with noiseless retrieval. To isolate the effect of neighborhood selection, we retrieve with the unnoised private sum S_c, replacing only PRDC's public mean by b_c^0. The noisy sums and counts, interpolation weights, and distance correction remain unchanged. Table 16 compares this diagnostic with PDC, PRDC, and the clean-mean classifier.

At IR 100 and $\rho = .125$, noiseless retrieval raises PRDC's tail recall from 14.5% to 48.0% on CIFAR-100 and from 26.8% to 55.1% on FOOD-101. The change on CIFAR-10 is only 0.6 pp. Thus, PRDC improves on PDC in the two many-class tasks even while noisy neighborhood selection still limits its tail performance.

D.4 Balanced training sets of the same total size

We next ask whether class imbalance is necessary for public information to help. Figure 5 compares IR 100 with balanced training sets of the same total size. At $\rho = .005$, PRDC improves BA over PDC by 22.1 pp on balanced CIFAR-100 and 9.4 pp on balanced FOOD-101. These gains shrink with increasing budget and become slightly negative from $\rho = .125$ onward. The IR-100 gains remain positive across this range on both datasets. CIFAR-10 changes little in the balanced setting. Public information can help even without imbalance when the available examples are spread across many classes. In this comparison, its benefit persists over a wider range of budgets under imbalance.

D.5 Privacy-allocation sensitivity

We multiply the count-budget share in Equation (4) by $\{.5, .75, 1, 1.5, 2\}$, assigning the remainder of the fixed total budget to the sums. The comparison uses IR 100, $\rho \in \{.005, .02, .125\}$, and

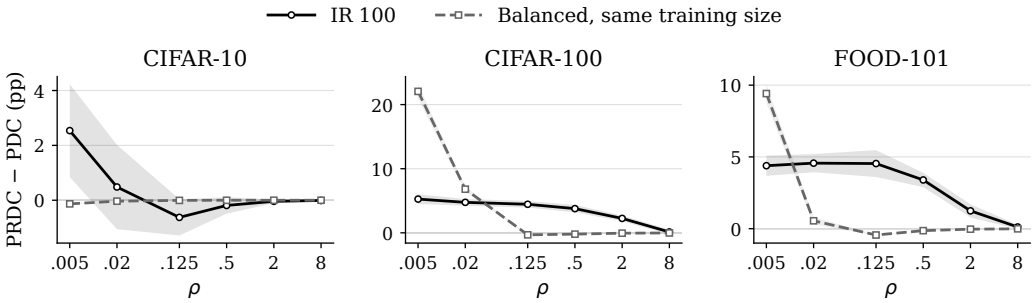


Figure 5: PRDC's BA gain over PDC for IR 100 and balanced training sets with the same total number of examples. Curves show means of the paired gains across training sets; bands show one standard deviation.

Table 17: Count-budget sensitivity: BA (%) averaged over $\rho \in \{.005, .02, .125\}$ at IR 100. Multipliers apply to the count-budget share in Equation (4). Entries are mean ± SD across training sets.

Count share	CIFAR-10		CIFAR-100		FOOD-101	
	PDC	PRDC	PDC	PRDC	PDC	PRDC
0.5×	85.7 ± 0.6	86.4 ± 0.8	36.7 ± 0.8	41.4 ± 0.8	42.0 ± 0.9	46.3 ± 0.8
0.75×	85.6 ± 0.6	86.4 ± 0.9	36.7 ± 0.8	41.5 ± 0.8	42.0 ± 0.9	46.4 ± 0.8
1×	85.6 ± 0.6	86.4 ± 0.9	36.6 ± 0.8	41.5 ± 0.8	41.9 ± 0.9	46.4 ± 0.8
1.5×	85.5 ± 0.6	86.4 ± 0.9	36.5 ± 0.8	41.4 ± 0.8	41.8 ± 0.9	46.4 ± 0.8
2×	85.4 ± 0.6	86.3 ± 0.9	36.4 ± 0.8	41.3 ± 0.8	41.7 ± 0.9	46.3 ± 0.8

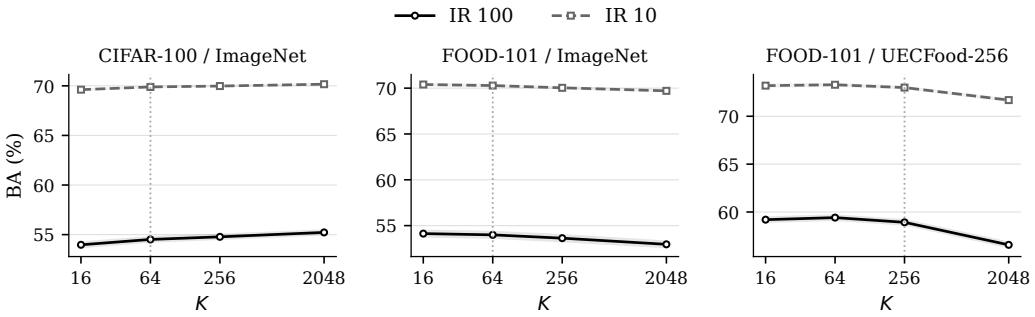


Figure 6: PRDC neighborhood-size sensitivity. Results are averaged over all six budgets within each training set. Curves show means across training sets; bands show one standard deviation. Vertical lines mark the source-specific K selected on public development tasks.

the same underlying Gaussian draws across allocations. Table 17 reports the three-budget average. Across the five allocations, each method's mean BA varies by less than 0.4 pp on each dataset, and PRDC remains above PDC at every allocation. The method comparison is stable over this range of count-budget shares.

D.6 Neighborhood-size sensitivity

Figure 6 varies K while keeping the source fixed. Larger neighborhoods modestly improve CIFAR-100, whereas FOOD-101 favors smaller neighborhoods. The source-specific choices selected on public development tasks are marked in the figure.

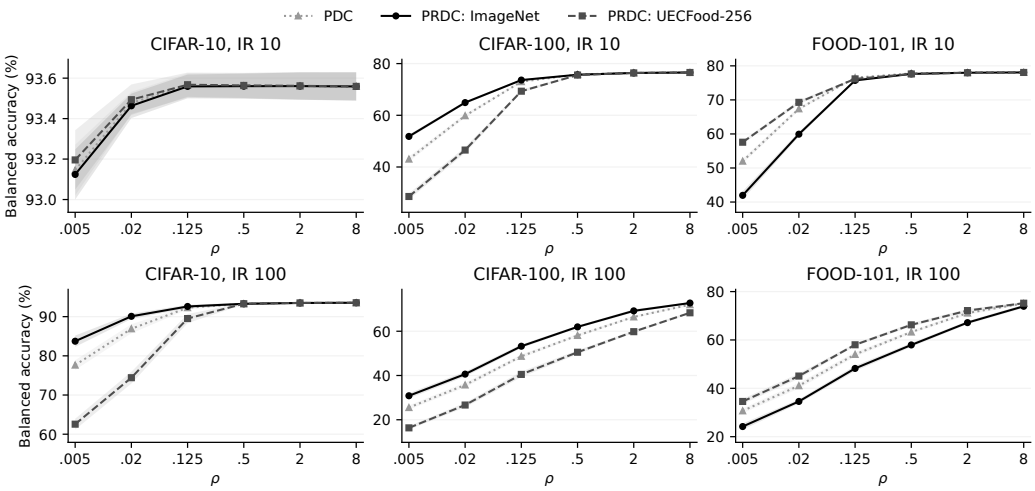


Figure 7: Public-bank source comparison with 20,000 features per bank. PDC does not use a bank. Curves show means across training sets; bands show one standard deviation.

Table 18: Matched-size bank comparison: BA and Macro-F1 (%) averaged over $\rho \in \{.005, .02, .125\}$. Entries are mean $\pm$ SD across training sets.

Dataset	IR	PDC		PRDC: ImageNet		PRDC: UECFood-256	
		BA	Macro-F1	BA	Macro-F1	BA	Macro-F1
CIFAR-10	10	93.40 ± 0.07	93.41 ± 0.07	93.38 ± 0.07	93.40 ± 0.07	93.42 ± 0.07	93.43 ± 0.07
	100	85.61 ± 0.64	84.70 ± 0.77	88.81 ± 0.57	88.33 ± 0.64	75.50 ± 0.44	70.05 ± 0.85
CIFAR-100	10	58.63 ± 0.38	57.00 ± 0.25	63.46 ± 0.25	61.63 ± 0.33	48.15 ± 0.53	42.19 ± 0.54
	100	36.63 ± 0.76	30.37 ± 0.70	41.56 ± 0.85	34.64 ± 0.91	27.83 ± 0.73	18.49 ± 0.83
FOOD-101	10	65.30 ± 0.28	63.82 ± 0.31	59.21 ± 0.48	54.71 ± 0.52	67.67 ± 0.34	66.54 ± 0.45
	100	41.93 ± 0.88	34.85 ± 0.88	35.67 ± 0.87	25.61 ± 0.87	45.88 ± 0.76	40.04 ± 0.73

D.7 Public-bank source and size

Source at matched size. To separate bank source from bank size, we draw five random subsets of 20,000 features from each of ImageNet and UECFood-256 and use both sources on every dataset. We retain $K = 64$ for CIFAR-10/100 and $K = 256$ for FOOD-101 when changing sources. Each comparison uses the same 20 training sets and ten noisy releases per set and budget. The five bank subsets are fixed across these releases, and their results are averaged within each training set.

Figure 7 shows the effect across budgets; Table 18 summarizes the three smallest budgets. At equal bank size, ImageNet gives higher BA on CIFAR-100 by 15.3 pp at IR 10 and 13.7 pp at IR 100, whereas UECFood-256 gives higher BA on FOOD-101 by 8.5 and 10.2 pp, respectively. On CIFAR-10, source choice has little effect at IR 10 but changes BA by 13.3 pp at IR 100. Macro-F1 follows the same pattern. With the less suitable source, PRDC falls below PDC in these five source-sensitive settings. Thus, a public bank can change the benefit of regularization even when its size and the interpolation weight are held fixed.

The source gaps narrow as the privacy budget grows and the private means receive more weight. This pattern is consistent with the factor $1 - \alpha_c$ multiplying the public-mean error in Theorem 2.

Bank size at fixed source. We next reduce each dataset's original bank to one tenth and one hundredth of its full size, keeping K unchanged. ImageNet then contains 128,116 or 12,811 features, and UECFood-256 contains 3,139 or 313. For each source, five random orderings define nested subsets at the two sizes. These subsets remain fixed across the same 20 training sets, ten releases, and six budgets; we average over subsets and releases within each training set.

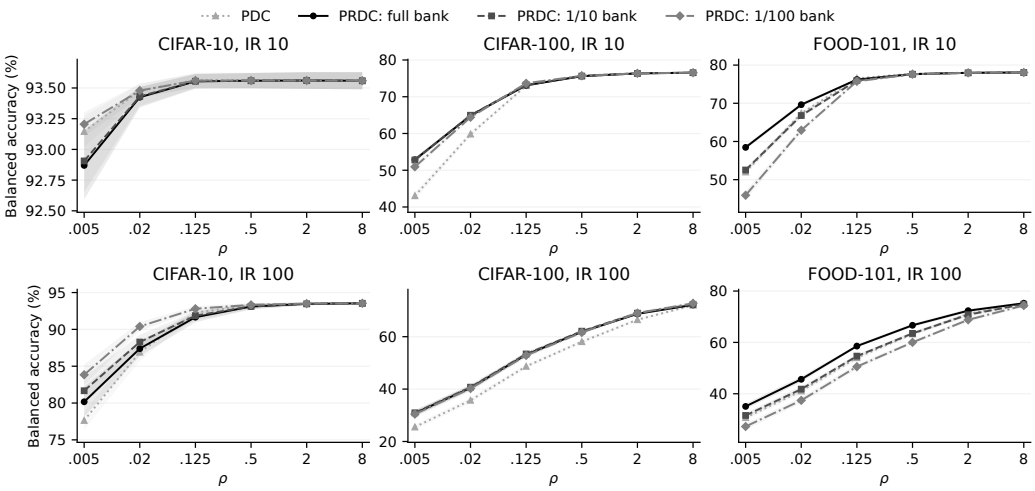


Figure 8: PRDC with smaller banks from the same source and fixed K. CIFAR-10/100 use ImageNet; FOOD-101 uses UECFood-256. Curves show means across training sets; bands show one standard deviation.

Table 19: PRDC bank-size comparison: BA and Macro-F1 (%), averaged over $\rho \in \{.005, .02, .125\}$. We fix $K = 64$ for CIFAR-10/100 and $K = 256$ for FOOD-101. Entries are mean $\pm$ SD across training sets.

Bank fraction	CIFAR-10		CIFAR-100		FOOD-101	
	BA	Macro-F1	BA	Macro-F1	BA	Macro-F1
IR 100						
Full bank	86.41 ± 0.86	85.57 ± 1.04	41.46 ± 0.79	36.17 ± 0.90	46.42 ± 0.78	40.81 ± 0.75
$1/10$	87.27 ± 0.79	86.64 ± 0.91	41.69 ± 0.78	36.11 ± 0.87	42.66 ± 0.79	35.68 ± 0.82
$1/100$	89.01 ± 0.51	88.52 ± 0.59	41.17 ± 0.89	33.85 ± 0.92	38.41 ± 0.84	29.36 ± 0.92
IR 10						
Full bank	93.28 ± 0.12	93.30 ± 0.13	63.61 ± 0.21	62.85 ± 0.26	68.10 ± 0.35	67.08 ± 0.46
$1/10$	93.30 ± 0.11	93.31 ± 0.11	63.65 ± 0.21	62.72 ± 0.27	65.10 ± 0.36	63.25 ± 0.46
$1/100$	93.41 ± 0.06	93.43 ± 0.06	63.00 ± 0.25	60.88 ± 0.33	61.53 ± 0.42	58.21 ± 0.45

Figure 8 and Table 19 show that the effect of reducing the bank differs across tasks. At one hundredth of ImageNet's size, CIFAR-100's three-budget mean BA changes by less than 0.7 pp at both IRs, although Macro-F1 falls by 2.0–2.3 pp. CIFAR-10 at IR 100 instead gains 2.6 pp in BA. On FOOD-101, the same relative reduction lowers BA by 6.6 pp at IR 10 and 8.0 pp at IR 100. Here $K = 256$ selects most of the remaining 313 features, leaving little scope for class-specific neighborhoods. Performance approaches the global-mean control in Table 13: the two differ by less than 0.4 pp in BA at both IRs.

D.8 A SECOND ENCODER

We replace SWAG with a frozen ResNet-18 encoder on CIFAR-10, preserving the training sets and using unit-normalized 512-dimensional features for both private data and the ImageNet bank. Table 20 reports every privacy level. At IR 100, PRDC adds 0.9–2.1 pp in BA over PDC across $\rho \in \{.005, .02, .125\}$. At IR 10, the difference stays within 0.3 pp across all six budgets. PDC improves over Noisy mean at the smallest budgets under both IRs, while PRDC's additional gain is concentrated at IR 100. At larger budgets the three methods approach the same accuracy.

Table 20: ResNet-18 / CIFAR-10: BA and Macro-F1 (%). Entries are mean ± SD across training sets.

IR	ρ	Noisy mean		PDC		PRDC	
		BA	Macro-F1	BA	Macro-F1	BA	Macro-F1
100	0.005	34.9 ± 2.5	22.5 ± 2.7	50.8 ± 1.7	42.7 ± 1.8	52.9 ± 1.2	45.8 ± 1.8
	0.02	43.6 ± 2.2	32.6 ± 2.5	58.8 ± 1.3	54.0 ± 1.6	60.8 ± 1.2	56.5 ± 1.8
	0.125	54.7 ± 1.4	46.3 ± 1.8	68.8 ± 1.0	68.2 ± 1.0	69.7 ± 0.7	68.8 ± 1.1
	0.5	63.0 ± 1.2	58.2 ± 1.8	72.9 ± 0.4	72.8 ± 0.5	73.0 ± 0.5	72.9 ± 0.6
	2	70.4 ± 1.0	69.1 ± 1.5	74.2 ± 0.3	74.2 ± 0.3	74.2 ± 0.3	74.2 ± 0.4
	8	74.1 ± 0.5	74.0 ± 0.6	74.6 ± 0.3	74.5 ± 0.4	74.6 ± 0.3	74.5 ± 0.4
10	0.005	56.5 ± 1.5	50.0 ± 2.2	72.3 ± 0.3	72.2 ± 0.3	72.5 ± 0.3	72.4 ± 0.4
	0.02	70.2 ± 1.1	69.3 ± 1.4	74.4 ± 0.2	74.4 ± 0.2	74.4 ± 0.2	74.4 ± 0.2
	0.125	74.8 ± 0.3	74.8 ± 0.3	75.0 ± 0.1	75.0 ± 0.1	75.0 ± 0.1	75.0 ± 0.1
	0.5	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2
	2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2
	8	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2	75.1 ± 0.2

Table 21: Classifier construction time (ms), mean ± SD across training sets at IR 100. Peak allocated GPU memory includes resident inputs and is the maximum over these runs; GB denotes 10^9 bytes. Bank loading is timed separately.

Dataset	Method	$\rho = .005$	$\rho = .125$	$\rho = 8$	Peak (GB)
CIFAR-100	PDC	0.86 ± 0.08	0.92 ± 0.17	0.95 ± 0.17	0.110
	PRDC	9.31 ± 0.22	9.37 ± 0.33	9.46 ± 0.31	4.552
	DP-SGD	534.40 ± 6.96	134.57 ± 1.08	526.89 ± 2.13	0.050
	DP-SGD+LA	529.27 ± 4.39	134.83 ± 0.74	527.35 ± 1.40	0.050
FOOD-101	PDC	1.03 ± 0.16	1.06 ± 0.16	1.05 ± 0.15	0.161
	PRDC	2.87 ± 0.21	2.90 ± 0.31	2.88 ± 0.27	0.406
	DP-SGD	792.39 ± 7.39	206.27 ± 1.12	788.81 ± 6.88	0.067
	DP-SGD+LA	791.28 ± 8.89	207.09 ± 2.54	787.87 ± 3.90	0.068

D.9 Classifier construction and cached inference

We measure PDC, PRDC, DP-SGD, and DP-SGD+LA on one NVIDIA RTX PRO 6000 GPU, running one model at a time. The comparison uses all 20 IR-100 training sets for CIFAR-100 and FOOD-101 at $\rho \in \{.005, .125, 8\}$, with one timed construction per set. All methods start from precomputed training features on the GPU. PDC and PRDC include class-sum and count aggregation, the private release, and prototype construction; PRDC also includes exact retrieval from the full bank used in the main experiments. The SGD variants use the selected settings in Table 4 and execute all training updates, including the count release for LA. We use the same analytic clipped-gradient implementation as in the accuracy experiments. Image encoding, public hyperparameter search, and privacy-accountant calibration are outside this feature-level comparison.

Table 21 reports synchronized wall-clock times after warm-up. PDC takes about 1 ms; PRDC takes about 9.4 ms with ImageNet and 2.9 ms with UECFood-256. The SGD variants take 0.13–0.79 s, depending on the dataset and selected training schedule. Their lower time at $\rho = .125$ reflects fewer epochs and a larger batch. PRDC's peak memory is dominated by its bank and retrieval workspace, whereas the other methods need no retrieval bank during construction.

Bank preparation and inference. Bank loading is separate from the construction times above. Across five warm-cache repetitions, reading and converting the stored features to FP32 takes 1.94 s for ImageNet and 0.047 s for UECFood-256; transfer to the GPU adds 0.113 s and 0.003 s, respectively. This preparation can be shared across releases using the same bank. After construction, the bank and training features are no longer needed for prediction. We store each classifier as FP32 linear-score weights and offsets, requiring 0.29–0.30 MiB for 100–101 classes in 768 dimensions. Across methods and budgets, median scoring-and-argmax latency is 20.1–20.4 μs for one feature vector and 24.5–25.8 μs for a batch of 1,024. These GPU timings use 1,000 repetitions after warm-up and exclude feature

extraction. The computational difference lies in classifier construction; cached inference uses the same linear scoring operation for all four methods.