Distributional Unlearning: Forgetting Distributions, Not Just Samples

Youssef Allouah [1] **Rachid Guerraoui** [1] **Sanmi Koyejo** [2]

Abstract

Machine unlearning seeks to remove unwanted information from trained models, initially at the individual-sample level, but increasingly at the level of entire sub-populations. In many deployments, models must delete whole topical domains to satisfy privacy, legal, or quality requirements, e.g., removing several users' posts under GDPR or copyrighted web content. Existing unlearning tools remain largely sample-oriented, and straightforward point deletion often leaves enough residual signal for downstream learners to recover the unwanted domain. We introduce *distributional unlearning*, a data-centric, model-agnostic framework that asks: *Given examples from an unwanted distribution and a retained distribution, what is the smallest set of points whose removal makes the edited dataset far from the unwanted domain yet close to the retained one?* Using Kullback–Leibler divergence to quantify removal and preservation, we derive the exact Pareto frontier in the Gaussian case and prove that any model retrained on the edited data incurs log-loss shifts bounded by the divergence thresholds. We propose a simple distance-based selection rule satisfying these constraints with a quadratic reduction in deletion budget compared to random removal. Experiments on synthetic Gaussians, Jigsaw Toxic Comments, SMS spam, and CIFAR-10 show 15–72% fewer deletions than random, with negligible impact on retained performance.

1. Introduction

Long-lived models meet short-lived data. Machine learning models often live in production far longer than the data on which they were trained. Over time, significant portions of the training data may become legally or ethically objectionable—think of a large text classifier that scrapes web forums, only to face a GDPR (gdp, 2016) takedown demand requiring the removal of several users' posting history, or an image-recognition system that must excise every photo from a now-copyrighted online archive (European Commission, 2023). In these scenarios, deleting a handful of flagged samples is often insufficient: the remaining dataset may still carry the statistical signature of the unwanted domain, allowing downstream models to relearn or exploit it. For example, recent works show that large language models can verbatim reproduce sequences that were removed from their training sets, purely from overlapping context (Liu et al., 2025). While "approximate retraining" methods, like influence-function updates (Ginart et al., 2019) or certified unlearning via convex optimization (Guo et al., 2020; Sekhari et al., 2021), reduce computational overhead, they do not minimize the number of samples that must be removed to erase a domain's influence, and hence do not address the core *statistical* question of sample-efficiency in unlearning.

Existing research on sample-level unlearning has mainly focused on per-record deletions with guarantees on parameter-space closeness to a retrained model (Guo et al., 2020; Bourtoule et al., 2021). These approaches typically assume that all unwanted samples have been perfectly identified and focus on how quickly the model parameters can be updated. By contrast, our work asks "which samples—and how many—must be deleted to sever the unwanted distribution's statistical footprint entirely?" Recent advances in class-level unlearning (Tarun et al., 2023; Kodge et al., 2024) and spurious-feature pruning (Mulchandani & Kim, 2025) demonstrate effective forms of domain-level removal, yet neither provides an explicit, divergence-based deletion budget for minimal sample removal. Additionally, concept erasure methods, e.g., INLP (Ravfogel et al., 2020), LEACE (Belrose et al., 2023), provide complementary, model-internal tools for suppressing unwanted signals post hoc, but because they do not edit the raw data, any downstream retraining can potentially reintroduce those concepts. Finally, robust optimization approaches such as distributionally-robust risk minimization (Namkoong & Duchi, 2016) seek to protect against worst-case shifts, rather than remove a known shift with provable guarantees on how many and which samples must be deleted.

While these lines of work tackle important aspects of unlearning or robustness, they do not provide a unified, distribution-level forgetting criterion coupled with minimal deletion guarantees. This gap raises a precise statistical question:

[1]EPFL, Switzerland [2]Stanford University, USA. Correspondence to: Youssef Allouah <youssef.allouah@epfl.ch>.

ICML 2025 Workshop on Machine Unlearning for Generative AI.
Copyright 2025 by the author(s).

Given samples from an unwanted distribution p_1 and a retained distribution p_2, how can we select and minimize the set of points to delete so that the edited empirical distribution is information-theoretically far from p_1 but close to p_2?

To answer this, we introduce *distributional unlearning*, a data-centric, model-agnostic framework that reframes forgetting as a purely statistical operation at the distribution level. We quantify "far" and "close" via the forward Kullback–Leibler (KL) divergence. This choice enables control over downstream log-loss performance with respect to both distributions.

Contributions. We summarize our contributions in tackling the above question as follows:

- *Formal framework:* We formalize distributional unlearning via forward-KL constraints; α for removal, ε for preservation, enabling strong guarantees on downstream log-loss.
- *Pareto frontier & downstream guarantees:* We derive the closed-form Pareto frontier of achievable (α, ε) pairs for Gaussians and characterize its geometry in terms of divergence between the reference distributions for exponential families. We prove that any predictor retrained on the edited dataset incurs log-loss shifts bounded precisely by α and ε.
- *Algorithm & sample efficiency:* We study the sample complexity of distributional unlearning with both random and selective removal (of the most divergent samples) mechanisms. We show that selective removal improves sample-efficiency quadratically, in terms of number of removed samples, compared to random removal.
- *Empirical validation:* Our selective removal strategy significantly reduces the removal budget versus random deletion. We report on experiments on synthetic Gaussians, Jigsaw toxic comments, SMS spam, and CIFAR-10 images, where we demonstrate 15–72% reductions in deletion budgets over random removal, with little loss in utility on retained domains.

2. Distributional Unlearning: Definition and Implications

We consider probability distributions over an input space $\mathcal{X}$, and we denote by $\mathcal{P}$ a class of distributions on $\mathcal{X}$. We focus on two distributions:

- $p_1 \in \mathcal{P}$: the *target* distribution to forget,
- $p_2 \in \mathcal{P}$: the *reference* distribution to preserve.

The goal is to construct a new distribution $p \in \mathcal{P}$ that removes the statistical influence of p_1 while retaining the properties of p_2. We formalize this via Kullback–Leibler (KL) divergence, which we recall for two absolutely-continuous distributions q, p on $\mathcal{X}$ is $\mathrm{KL}(q\|p) := \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)}\, dx$. Throughout this section, we defer all proofs to Appendix D.

Definition 1 ((α, ε)-Distributional Unlearning)**.** *For tolerances $\alpha, \varepsilon > 0$, a distribution $p \in \mathcal{P}$ satisfies (α, ε)-distributional unlearning with respect to (p_1, p_2) if:*

$$\mathrm{KL}(p_1 \parallel p) \geq \alpha \quad \textit{(removal)}, \tag{1}$$

$$\mathrm{KL}(p_2 \parallel p) \leq \varepsilon \quad \textit{(preservation)}. \tag{2}$$

The first inequality forces the edited data to be information-theoretically distant from the population we wish to forget. The second inequality upper bounds collateral damage to the population we preserve. We adopt the forward KL divergence as it directly controls the expected log-loss under the edited distribution p, which is critical for bounding downstream predictive performance (Prop. 2).[1]

Feasibility and the pareto frontier. The pair (α, ε) captures a trade-off: how far we can move from p_1 while remaining close to p_2. To understand which (α, ε) pairs are jointly achievable, we characterize the feasible region and its boundary. Formally, the *Pareto frontier* $\mathrm{PF}(p_1, p_2; \mathcal{P})$ consists of those pairs (α, ε) for which no strictly better trade-off exists: there is no $p' \in \mathcal{P}$ satisfying $\mathrm{KL}(p_1\|p') \geq \alpha'$ and $\mathrm{KL}(p_2\|p') \leq \varepsilon'$ with $\alpha' > \alpha$ and $\varepsilon' < \varepsilon$. That is, every point on the frontier is optimal in the sense that one objective cannot be improved without worsening the other.

The following result gives the closed-form Pareto frontier in the Gaussian case:

Proposition 1 (Pareto Frontier)**.** *Let p_1, p_2 be two distributions in $\mathcal{P}$, the class of Gaussian distributions with shared covariance, with $\mathrm{KL}(p_1\|p_2) < \infty$. The Pareto frontier of (α, ε) values achievable by any $p \in \mathcal{P}$ is given by:*

$$\mathrm{PF}(p_1, p_2; \mathcal{P}) = \Big\{ \big(\alpha, \big(\sqrt{\alpha} - \sqrt{\mathrm{KL}(p_1\|p_2)}\big)^2\big) : \alpha \geq \mathrm{KL}(p_1\|p_2) \Big\}.$$

This frontier characterizes the minimal preservation loss ε required to achieve a given removal level α. In particular, the value $\mathrm{KL}(p_1\|p_2)$ plays a critical role: no distribution can forget more than this amount ($\alpha > \mathrm{KL}(p_1\|p_2)$) while remaining arbitrarily close to p_2. The shape of the frontier reflects how intertwined p_1 and p_2 are. While this curve was derived analytically for Gaussians, the same removal–preservation tradeoff arises more broadly, e.g., regular exponential families (Proposition 3). Empirically, the frontier matches closely in synthetic Gaussian experiments (Fig. 1), confirming both its shape and the threshold behavior.

3. Algorithms and Sample Complexity

We now instantiate the distributional unlearning framework in a concrete setting where we only have access to samples

[1] Reverse KL or Wasserstein could be used to enforce other tail behaviors, which we leave for future work.

Mechanism	Sample Complexity	
	Removal	Preservation
Random Removal (Thm. 1)	$n_1\left(1-\sqrt{1-\alpha}\right)$	$n_1\left(1-\sqrt{\varepsilon}\right)$
Selective Removal (Thm. 2)	$n_1\left(1-(1-\alpha)^{1/4}\right)$	$n_1\left(1-\varepsilon^{1/4}\right)$

Table 1. Summary of simplified sample complexity bounds: the formulas give the minimal number of p_1 samples (out of $n_1 \geq 1$) to remove in order to achieve (α, ε)-distributional unlearning, and show the quadratic improvement of selective removal over random in ε, α. We focus on dependences on $\alpha, \varepsilon \in (0,1)$ and ignore constants, $\mathrm{KL}(p_1 \| p_2)$, and $\frac{n_2}{n_1}$, and assume that n_2 is large enough. Full expressions and assumptions are given in Corollary 8 in the appendix.

from p_1 and p_2, and we must decide which points to remove and how many in order to achieve (α, ε)-distributional unlearning. In this section, we *(i)* introduce two deletion strategies—random removal and selective removal—and *(ii)* derive high-probability bounds on the number of samples that must be deleted under each.

In the following, we focus on the univariate Gaussian case, since it captures the essential phenomena. That is, denoting $\mathcal{P} := \{\mathcal{N}(\mu, \sigma^2) \colon \mu \in \mathbb{R}\}$, where $\sigma > 0$, we have $p_1, p_2 \in \mathcal{P}$. We consider n_1 i.i.d. samples from p_1 and n_2 from p_2. We let $0 \leq f \leq n_1$ denote the number of points removed from the p_1 samples. We defer all proofs to Appendix D.

3.1. Random Removal

We begin with a baseline deletion strategy that treats every sample equally, deleting f points chosen uniformly at random from the n_1 samples of p_1. For clarity, we first state the random removal procedure formally below:

Algorithm (Random Removal).

1. Randomly select f out of the n_1 samples of p_1 without replacement.
2. Remove those f samples.
3. Re-fit $\mathcal{N}(\hat{\mu}, \sigma^2)$ by MLE on the remaining data.

The following theorem provides a finite-sample guarantee for achieving (α, ε)-distributional unlearning using random removal with a deletion budget f.

Theorem 1 (Random Removal)**.** *Let $p_1, p_2 \in \mathcal{P}$ and $\delta \in (0,1)$. We observe n_1 samples from p_1 and n_2 samples from p_2, and randomly remove f samples from p_1 before fitting. Then with probability at least $1-\delta$, the resulting MLE distribution satisfies (α, ε)-distributional unlearning with:*

$$\alpha \geq \left(\frac{1}{2} - 3\left(\tfrac{n_1-f}{n_2}\right)^2\right)\mathrm{KL}(p_1\|p_2) - \frac{3\ln(4/\delta)}{2n_2}\left(1+\tfrac{n_1-f}{n_2}\right),$$

$$\varepsilon \leq 3\left(\tfrac{n_1-f}{n_2}\right)^2\mathrm{KL}(p_1\|p_2) + \frac{3\ln(4/\delta)}{n_2}\left(1+\tfrac{n_1-f}{n_2}\right).$$

This result shows that random removal can achieve both forgetting and preservation as long as f is large enough, so that the term $\frac{n_1-f}{n_2}$, i.e., ratio of number of non-deleted p_1 samples to p_2 samples, is small enough. First, we remark that the last term in both inequalities can be ignored if n_2 is large enough compared to $\log(1/\delta)$. Next, the removal guarantee improves linearly in $\mathrm{KL}(p_1\|p_2)$, but quadratically in $\frac{n_1-f}{n_2}$, indicating diminishing returns as more data is deleted. Note that preservation improves as n_2 increases, since the estimation of p_2 becomes more accurate, reducing collateral damage. While conceptually simple, this method does not prioritize which samples contribute most to the divergence from p_2, and may thus be inefficient in certain regimes.

3.2. Selective Removal

We next consider a selective removal strategy that uses knowledge of the data to delete the most distinguishable samples from p_2, specifically the lowest likelihood under the empirical estimation of the preserved distribution p_2. We outline the exact procedure below:

Algorithm (Selective Removal).

1. Compute the mean $\hat{\mu}_2$ of the n_2 samples from p_2.
2. For each of the n_1 samples x_i from p_1, compute the score $s_i = |x_i - \hat{\mu}_2|$.
3. Delete the f samples with the largest scores s_i.
4. Re-fit $\mathcal{N}(\hat{\mu}, \sigma^2)$ by MLE on the remaining data.

The following theorem provides a finite-sample guarantee for achieving (α, ε)-distributional unlearning using selective removal with a deletion budget f.

Theorem 2 (Selective Removal)**.** *Let $p_1, p_2 \in \mathcal{P}$ and $\delta \in (0,1)$. Let f samples from p_1 be removed according to Selective Removal. Then with probability at least $1-\delta$, the resulting estimate satisfies (α, ε)-distributional unlearning with:*

$$\alpha \geq \frac{1}{2}\mathrm{KL}(p_1\|p_2) - \frac{1}{2}\left(\tfrac{n_1-f}{n_2}\right)^2 \Gamma(f, n_1, \delta, \mathrm{KL}(p_1\|p)) - \tfrac{\ln(4/\delta)}{n_2},$$

$$\varepsilon \leq \left(\tfrac{n_1-f}{n_2}\right)^2 \Gamma(f, n_1, \delta, \mathrm{KL}(p_1\|p)) + \tfrac{2\ln(4/\delta)}{n_2},$$

where $\Gamma(f, n_1, \delta, \mathrm{KL}(p_1\|p)) := g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1\|p_2)\Big)^2$ *with* $g(u;\kappa) := \Phi(u - \sqrt{2\kappa}) + \Phi(u + \sqrt{2\kappa}) - 1$, *for* $u, \kappa > 0$, *and* Φ *is the standard normal CDF.*

While this result also shows that larger deletion budgets improve both removal and preservation, the main difference

Domain	Target on p_1	Random	Selective	Saving
Gaussians (low)	$\mathrm{KL}(p_1 \| p)$	65	18	**72 %**
Gaussians (high)	$\mathrm{KL}(p_1 \| p)$	65	50	**23 %**
Jigsaw toxic comments	Recall	100	85	**15 %**
SMS Spam	Recall	90	75	**17 %**
CIFAR-10	$\mathrm{Accuracy}_{\mathrm{cat}}$	80	50	**38 %**

Table 2. Deletion budget (%) needed to reach half of the initial value of the removal metric (no deletion) on each dataset. "Selective"=best-performing selective removal score; "Saving"=relative reduction versus random deletion. Gaussians (low) and (high) are the scenarios of the top leftmost and rightmost plots of Fig. 2, respectively.

with random removal is that the quadratic factor $(\frac{n_1-f}{n_2})^2$ is amplified by a multiplicative factor involving the inverse of $g(\cdot\,;\kappa)$, which represents the CDF of the folded normal shifted by divergence $\kappa = \mathrm{KL}(p_1\|p_2)$. Therefore, for $p \in (0,1)$, $g^{-1}(p;\kappa)$ is the p-th quantile of the folded normal shifted by the aforementioned divergence κ. Intuitively, this term arises as scores computed by selective removal, which are of the form $s_i = |x_i - \hat{\mu}_2|$ with x_i being a sample from p_1, are distributed following a folded normal. Thus, we are effectively selecting samples from p_1 whose distance to the mean of p_2 is at most the $\approx (1-\frac{f}{n_1})$-th quantile of the aforementioned folded normal, which is fortunately a decreasing function in f. That is, this term strictly amplifies the quadratic decrease in f, which was the best we could previously obtain with random removal.

Simplified bounds. While the above theorem is general and gives exact quantities, the term involving $g^{-1}(\cdot\,;\kappa)$ can be upper bounded with a simpler analytical expression. As we mentioned previously, this term represents a quantile of the folded normal distribution. We can bound this quantile with a linear approximation, when the divergence $\mathrm{KL}(p_1\|p_2)$ is small enough, i.e., p_1 and p_2 are similar, and $\frac{f}{n_1}$ is large enough, i.e., we delete enough samples. The result of this simplification is given in Table 1, and further developed in Corollary 8. Essentially, focusing only on the dependence on $\alpha \in (0,1)$ for removal, selective removal only requires $\mathcal{O}(n_1(1-\sqrt{1-\alpha}))$, while random removal requires $\mathcal{O}(n_1(1-(1-\alpha)^{1/4}))$. Analogous conclusions hold for ε for preservation. Therefore, we show that selective removal strictly improves, at least quadratically, over random removal with respect to ε, α in low-divergence regimes.

We empirically validate the improvement over random removal in Figure 2. Specifically, as discussed earlier, when $\mathrm{KL}(p_1\|p_2)$ is small (top left plot), selective removal significantly improves over random removal. When the distributions are very far apart (top right plot), the difference narrows and both mechanisms perform comparably.

4. Empirical Validation

We now evaluate distributional unlearning empirically across synthetic as well as real-world data. Our goals are threefold:

1. Compare the (deletion) sample-efficiency of random versus selective deletion strategies.
2. Assess how distributional unlearning shapes downstream predictive performance.
3. Validate the theoretical Pareto frontier of achievable (α, ε) values.

We consider four qualitatively different settings: synthetic Gaussians, Jigsaw toxic-comment moderation, SMS spam filtering, and image classification on CIFAR-10. In each case, we define a distribution p_1 to forget and a distribution p_2 to preserve, apply several deletion methods to remove samples from p_1, and measure the resulting trade-off in statistical proximity and model performance. All results are averaged over multiple random seeds. We defer experimental details to Appendix E.

Results overview. We summarize our main findings in Table 2 for convenience, before detailing them in Appendix C due to space constraints. Across Gaussians, toxic text, short-message spam, and natural images, the same selective removal recipe cuts the deletion budget by $15-72\,\%$ relative to random removal while reducing accuracy or recall on the forget distribution by at least half. The gains are largest in the low-divergence Gaussian regime ($72\,\%$), where theory predicts the greatest advantage for selective deletion. The same trends persist in high-diversity real data ($15-38\,\%$) confirming that distributional unlearning controls downstream predictive performance as predicted by Proposition 2. Crucially, we achieve these savings without harming utility on the retained distribution. Table 2 distills the central empirical message: selective removal delivers the desired distributional unlearning guarantees with substantially less data removal than naïve approaches consistently across domains and models.

5. Conclusion and Discussion

While our approach operates by editing samples drawn from structured distributions, the core idea extends naturally to representation-level interventions, structured latent-variable models, or causal inference frameworks. Distributional unlearning may also enhance fairness, robustness, or interpretability by targeting and removing harmful subpopulation signals. By operating at the level of data distributions rather than model internals, it complements existing unlearning and debiasing techniques in a model-agnostic fashion. As a general-purpose statistical primitive, it offers a new lens for thinking about modularity and controllability in data-centric learning pipelines.

References

Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Official Journal of the European Union, May 2016.

Allouah, Y., Kazdan, J., Guerraoui, R., and Koyejo, S. The utility and complexity of in- and out-of-distribution machine unlearning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Almeida, T. A., Gómez Hidalgo, J. M., and Yamakami, A. Contributions to the study of sms spam filtering: New collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, pp. 3–10. ACM, 2011.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Bertsekas, D. P. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):38–39, 1997.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Chien, E., Wang, H. P., Chen, Z., and Li, P. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3LKuC8rbyV.

cjadams, Sorensen, J., Elliott, J., Dixon, L., McDonald, M., nithum, and Cukierski, W. Toxic comment classification challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge, 2017. Kaggle.

Elazar, Y. and Goldberg, Y. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21, 2018.

European Commission. Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (AI Act). COM(2021) 206 final, 2023.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. In *International Conference on Machine Learning*, pp. 3832–3842. PMLR, 2020.

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Kaushik, D., Hovy, E., and Lipton, Z. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Kodge, S., Saha, G., and Roy, K. Deep unlearning: Fast and efficient gradient-free class forgetting. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=BmI5p6wBi0.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kurmanji, M., Triantafillou, P., Hayes, J., and Triantafillou, E. Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

Liu, K. Z., Choquette-Choo, C. A., Jagielski, M., Kairouz, P., Koyejo, S., Liang, P., and Papernot, N. Language models may verbatim complete text they were not explicitly trained on. *arXiv preprint arXiv:2503.17514*, 2025.

Mulchandani, V. and Kim, J.-E. Severing spurious correlations with data pruning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Bk13Qfu8Ru.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Ravfogel, S., Elazar, Y., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7231–7245, 2020.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Tarun, A. K., Chundawat, V. S., Mandal, M., and Kankanhalli, M. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Thudi, A., Jia, H., Shumailov, I., and Papernot, N. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 4007–4022, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-31-1. URL https://www.usenix.org/conference/usenixsecurity22/presentation/thudi.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Zhang, B., Dong, Y., Wang, T., and Li, J. Towards certified unlearning for deep neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 58800–58818, 2024.

A. Related Work

Sample-level unlearning. Sample-level unlearning has made impressive strides in fast model updates and formal deletion guarantees (Neel et al., 2021; Thudi et al., 2022; Zhang et al., 2024; Chien et al., 2024; Kurmanji et al., 2024; Allouah et al., 2025). Izzo et al. (Izzo et al., 2021) use influence-function updates to approximate the effect of deleting a single point without full retraining; Guo et al. (Guo et al., 2020) provide certified bounds on how close the post-deletion model is to a scratch-retrained one; Bourtoule et al. (Bourtoule et al., 2021) employ data sharding to efficiently erase small batches. However, this class of methods does not address which or how many samples must be removed to eliminate a domain's overall statistical footprint. Our work complements these techniques by asking not just how to update a model once samples are flagged, but which samples to flag in the first place, and in what quantity.

Model-internal unlearning. Concept erasure methods tackle unwanted attributes in learned features. INLP (Ravfogel et al., 2020) repeatedly projects out directions predictive of a protected attribute; counterfactual augmentation (Kaushik et al., 2020) synthesizes targeted data to sever causal links; adversarial training (Elazar & Goldberg, 2018) trains encoders to remove specific signals. These operate post-hoc on a fixed model's representations—ideal for fairness use-cases such as removing gender or sentiment—but they rely on white-box access and tailor to one model at a time. Where concept erasure edits model representations, we edit the data, guaranteeing forgetting for downstream models.

Domain adaptation, robustness, and coresets. Domain adaptation theory bounds target error by source error plus a divergence between domains (Ben-David et al., 2010). Our work flips and complements this paradigm: we intentionally *increase* divergence from an unwanted source p_1 while controlling proximity to a desired reference p_2. Distributionally-robust optimization (Namkoong & Duchi, 2016) protects against all shifts within a divergence ball, whereas we target the removal of one specific shift. Coreset and importance-sampling methods (Sener & Savarese, 2018) select representative subsets to approximate a distribution; we invert that idea to remove the most representative samples of the unwanted component, while preserving another.

B. Implications for Downstream Prediction

We now connect distributional unlearning to predictive performance in supervised learning. Consider a predictor $h : \mathcal{X} \to \Delta(\mathcal{Y})$, where $\mathcal{X}$ is the input space and $\Delta(\mathcal{Y})$ the probability simplex over label space $\mathcal{Y}$, trained on a distribution p over $\mathcal{X} \times \mathcal{Y}$ that satisfies (α, ε)-unlearning with respect to (p_1, p_2). We study how h performs under the true data-generating distributions p_1 and p_2.

Let $\ell(y, q) := -\log q(y)$, for $y \in \mathcal{Y}, q \in \Delta(\mathcal{Y})$, denote the log-loss. Define the expected loss under p as $\mathcal{L}(h; p) := \mathbb{E}_{(x,y)\sim p}[\ell(y, h(x))]$. Then, for any class of distributions $\mathcal{P}$, we have:

Proposition 2. *Let h minimize $\mathcal{L}(h; p)$, and let h_1, h_2 be optimal predictors under $p_1, p_2 \in \mathcal{P}$, respectively. If p satisfies (α, ε)-distributional unlearning with respect to (p_1, p_2), then:*

$$\mathcal{L}(h; p_1) - \mathcal{L}(h_1; p_1) \geq \alpha - \delta_1, \tag{3}$$

$$\mathcal{L}(h; p_2) - \mathcal{L}(h_2; p_2) \leq \varepsilon - \delta_2, \tag{4}$$

where $\delta_1 := \mathrm{KL}(p_{1,X} \| p_X), \delta_2 := \mathrm{KL}(p_{2,X} \| p_X)$ denote the marginal KL divergence over inputs.

These bounds show that distributional unlearning guarantees increased loss under the forgotten distribution and bounded degradation under the preserved one. In this sense, the (α, ε)-distributional unlearning framework provides meaningful control over downstream predictive behavior. Regarding the extra marginal KL term in the first inequality, which quantifies divergence on input distributions, the data-processing inequality gives $\mathrm{KL}(p_{1,X} \| p_X) \leq \mathrm{KL}(p_1 \| p)$, hence this extra term is always bounded by the same α we already control. A similar term appears in the second inequality, but can only improve the preservation bound.

C. Empirical Validation

We now evaluate distributional unlearning empirically across synthetic as well as real-world data. Our goals are threefold:

1. Compare the (deletion) sample-efficiency of random versus selective deletion strategies.
2. Assess how distributional unlearning shapes downstream predictive performance.
3. Validate the theoretical Pareto frontier of achievable (α, ε) values.

We consider four qualitatively different settings: synthetic Gaussians, Jigsaw toxic-comment moderation, SMS spam filtering, and image classification on CIFAR-10. In each case, we define a distribution p_1 to forget and a distribution p_2

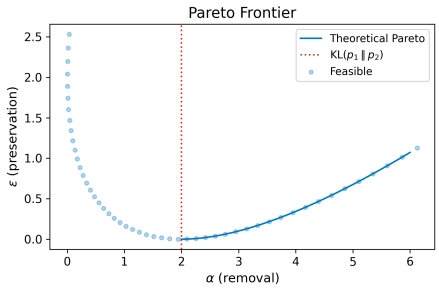


Figure 1. **Synthetic Gaussians.** The empirical frontier aligns with the theoretical prediction.

to preserve, apply several deletion methods to remove samples from p_1, and measure the resulting trade-off in statistical proximity and model performance. All results are averaged over multiple random seeds. We defer experimental details to Appendix E.

Results overview. We summarize our main findings in Table 2 for convenience, before detailing them in the next sections. Across Gaussians, toxic text, short-message spam, and natural images, the same selective removal recipe cuts the deletion budget by $15-72\,\%$ relative to random removal while reducing accuracy or recall on the forget distribution by at least half. The gains are largest in the low-divergence Gaussian regime ($72\,\%$), where theory predicts the greatest advantage for selective deletion. The same trends persist in high-diversity real data ($15-38\,\%$) confirming that distributional unlearning controls downstream predictive performance as predicted by Proposition 2. Crucially, we achieve these savings without harming utility on the retained distribution. Table 2 distills the central empirical message: selective removal delivers the desired distributional unlearning guarantees with substantially less data removal than naïve approaches consistently across domains and models.

C.1. Synthetic Gaussians: frontier verification and sample efficiency

We begin with a synthetic setting to evaluate distributional unlearning under controlled conditions. Let $p_1 = \mathcal{N}(0,1)$ and $p_2 = \mathcal{N}(\mu,1)$ for varying $\mu \in \{0.5, 2.5, 5\}$, which induces increasing divergence $\mathrm{KL}(p_1\|p_2)$. For each configuration, we draw $n_1 = n_2 = 1000$ samples from p_1 and p_2, respectively, and examine how different deletion strategies affect the post-edit distribution. We implement two mechanisms: *random removal*, which deletes a uniform subset of p_1 points, and *selective removal*, which prioritizes those samples with largest deviation from the empirical mean of p_2. After deleting f samples, we re-fit a Gaussian $\mathcal{N}(\hat{\mu},1)$ on the retained p_1 and p_2 samples via maximum likelihood, yielding a post-edit distribution p. We then compute forward KL divergences $\alpha = \mathrm{KL}(p_1\|p)$ and $\varepsilon = \mathrm{KL}(p_2\|p)$ to quantify forgetting and preservation, respectively.

Pareto frontier. Figure 1 confirms that the empirical (α, ε) trade-off closely matches the theoretical Pareto frontier derived in Proposition 1. To plot feasible empirical trade-offs, we set $p = \mathcal{N}(\mu,1)$ and vary $\mu \in \mathbb{R}$, while $p_1 = \mathcal{N}(0,1)$ and $p_2 = \mathcal{N}(2,1)$ so that $\mathrm{KL}(p_1\|p_2) = 2$. The latter quantity is the threshold predicted by the theory, and validated by Figure 1. Indeed, feasible trade-offs whose removal divergence α is below this threshold are pareto-suboptimal. They are dominated by the trade-off $(\alpha = \mathrm{KL}(p_1\|p_2), \varepsilon = 0)$, which can be achieved with the choice of distribution $p = p_2$.

Sample efficiency. We next compare the efficiency of the two removal strategies in the finite-sample case analyzed in Section 3. In Figure 2, we plot achieved α as a function of the number of p_1 samples removed. Selective removal reaches higher forgetting levels with fewer deletions than random removal, especially when $\mathrm{KL}(p_1\|p_2)$ is small ($\mu_2 = 0.5$, top left plot). For example, to reach 0.06 nats of removal divergence, i.e., half of that obtained by removing all samples, selective removal requires $5\times$ less samples than random removal. Analogous trends hold for preservation (bottom left plot): selective removal more effectively preserves the reference distribution p_2 throughout. The remaining plots show similar trends when increasing the divergence $\mathrm{KL}(p_1\|p_2)$ by increasing μ_2. As predicted in theory, selective removal offers the greatest savings when p_1 and p_2 are close and diminishes as the distributions diverge.

C.2. Jigsaw Toxic: forgetting a toxic sub-population

We now consider the Jigsaw toxic comments dataset (cjadams et al., 2017). Let p_1 be the set of comments that contain *any* of five high-frequency profanity keywords (see Appendix E). This sub-population accounts for $8.6\,\%$ of the training corpus. All remaining comments form the reference distribution p_2. For each deletion budget f corresponding to removal fractions $\{0, 0.05, \ldots, 1\}$, we remove f comments from p_1 ranked by four scoring rules computed based on TF-IDF embeddings, plus a random baseline:

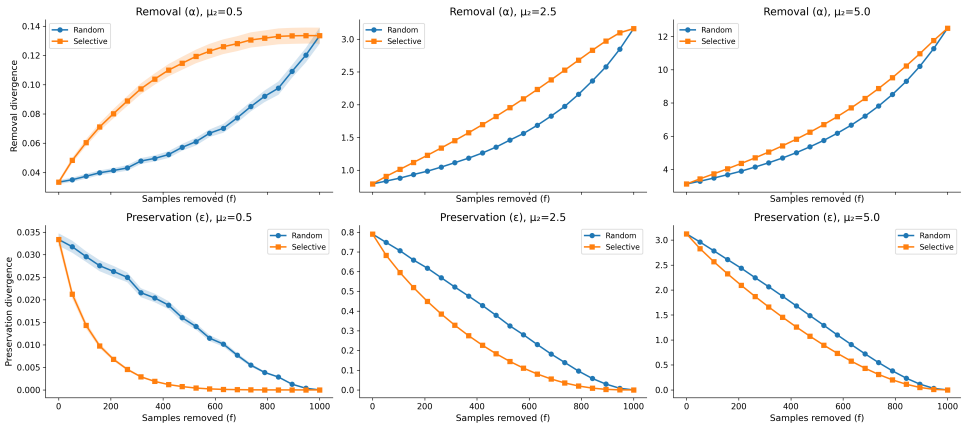
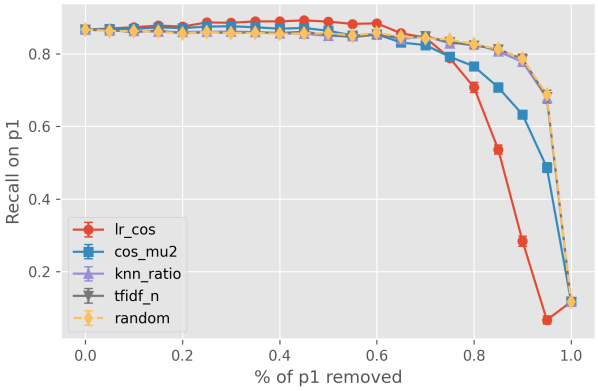
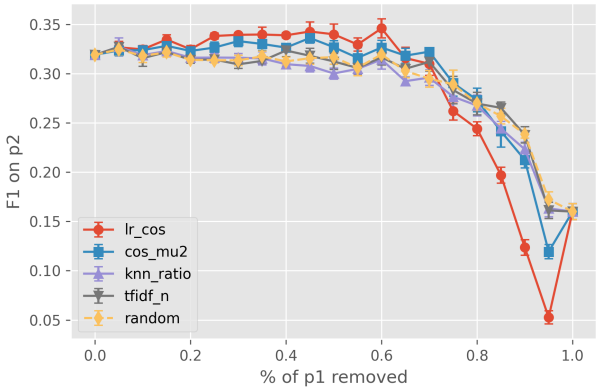


Figure 2. **Synthetic Gaussians.** Selective removal consistently requires fewer deletions, especially when $\mathrm{KL}(p_1 \| p_2)$ is small (left), for the same removal and preservation target as random removal. In high-divergence regimes (right), the gap between methods shrinks, as predicted by the theory.



(a) Recall on profane comments (p_1, forgotten).



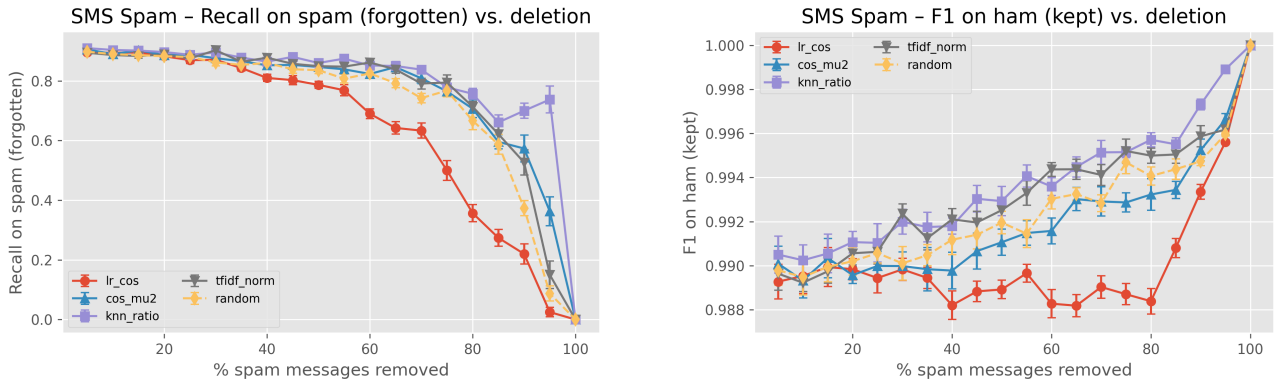
(b) Macro-*F1* on non-profane comments (p_2, kept).

Figure 3. **Jigsaw Toxic Comments.** Impact of removing profane comments on Jigsaw Toxic. *Left:* recall on the to-be-forgotten set p_1; *right:* macro-F_1 on the retained set p_2. Utility is almost unchanged up to 60% deletion; marked forgetting appears only around 80% deletion, with LR-COS showing the steepest drop. Error bars: ± 1 standard error over five randomness seeds.

1. **COS-MU2**: cosine distance to the p_2 mean only;
2. **LR-COS**: cosine margin between the comment and the mean vectors of p_2 vs. p_1;
3. **KNN-RATIO**: local k-NN density ratio (k=10);
4. **TFIDF-NORM**: comment ℓ_2 length; and
5. **RANDOM**: uniformly random baseline.

The first heuristic COS-MU2 above is a proxy for the selective removal method introduced in Section 3, where we use cosine distance between TF-IDF embeddings instead of Euclidean distance. The second heuristic LR-COS is a simple extension, and stands for a proxy of a likelihood ratio score, i.e., we remove samples most distinguishable from p_2 while being representive of p_1. After deletion we re-train a logistic–TF-IDF model on the kept data. We report *Recall*@p_1 (higher means less forgetting) and macro-F_1@p_2 (higher means stronger preservation).

Findings. We report our main findings on our Jigsaw task in Figure 3. Across all heuristics, Recall@ p_1 remains essentially flat around 0.88 until large budgets; a pronounced forgetting effect appears only at $f \geq 0.8|p_1|$, where LR-COS decreases recall by nearly 20 percentage points (Fig. 3a). Crucially, macro-F_1@ p_2 is almost unchanged up to $f = 0.6|p_1|$ (Fig. 3b) and declines gradually afterwards, indicating that removing up to 60 % of the profane comments has negligible



(a) Recall on spam (p_1, forgotten).

(b) Macro-*F1* on ham (p_2, kept).

Figure 4. **SMS Spam.** The likelihood-ratio score (LR-COS) pushes spam recall below 0.6 after deleting 70% of spam, whereas random deletion needs nearly 90% removal to reach the same point. Ham performance remains almost flat (<0.004 absolute change) until the final 100 % budget, affirming the tight preservation guarantee. Error bars: ± 1 standard error over ten seeds.

impact on overall utility. The sharp drop at full deletion confirms the theoretical pareto frontier prediction that excessive removal targets must harm performance on the retained distribution. This can be naturally explained on this toxic comment detection task, since p_1 the distribution of profane comments, carries task-relevant signal. By contrast, random removal requires removing most of p_1, the distribution of profane comments, to reach the same removal level of selective methods. Specifically, random removal requires removing 95% of p_1 to achieve around 0.7 recall, which means that LR-COS saves around 16% and COS-MU2 saves 11% in terms of deletion budget compared to random removel for the same Recall@ p_1 target. All other methods have a similar performance to random removal.

C.3. SMS Spam: a content-moderation unlearning task

We revisit the UCI SMS Spam Collection (Almeida et al., 2011), treating the *spam* class (p_1) as information to forget and the *ham* class (p_2) as information to preserve. Messages are vectorised with TF–IDF features. Deletion budgets again span 5% to 100% of the spam slice in 5-point increments. We compare the same five scoring rules as before (COS-MU2, LR-COS, KNN-RATIO, TFIDF-NORM, RANDOM) and average results over ten random seeds. Metrics reported are *recall* on p_1 and macro-*F1* on p_2.

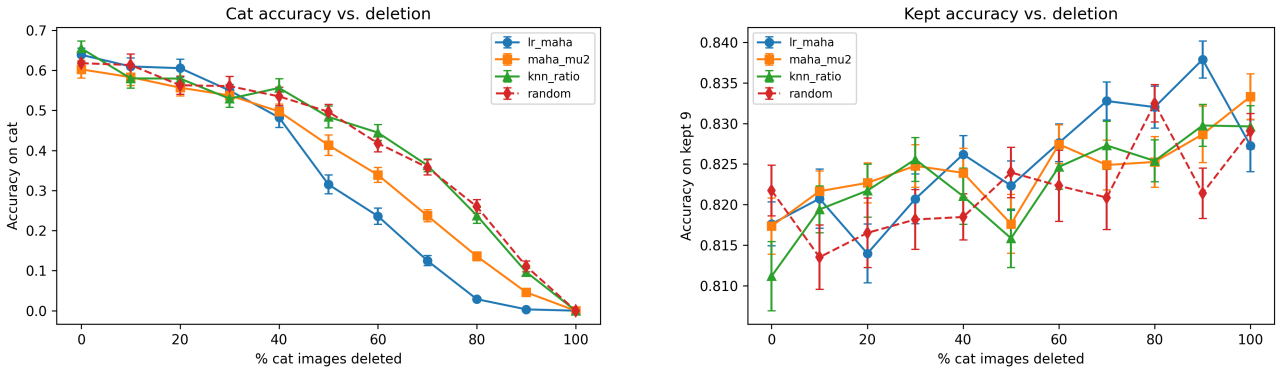
Findings. We report our main findings on our SMS Spam task in Figure 4. We observe that spam recall decays gradually until 75–80% deletion, after which all methods converge to zero as p_1 vanishes. LR-COS consistently dominates: it reaches a recall of 0.60 at the 70% deletion budget, whereas random deletion does not cross that threshold until 90% deletion. Throughout, ham macro-F_1 *increases* slightly (see Fig. 4b), an artefact of class-imbalance—removing spam reduces false positives in the ham slice—yet the difference across methods never exceeds 0.002. These results strengthen the evidence that selective deletion offers a $1.3-1.5\times$ sample-efficiency gain over random removal while preserving downstream utility almost perfectly.

C.4. CIFAR-10: privacy-motivated forgetting of an entire class

We treat the CIFAR-10 (Krizhevsky, 2009) cat category as a privacy-sensitive sub-population. Deleting a few individual images is insufficient; their aggregate statistics would still influence the model. Instead, we rank every cat image with three distance-based scores (LR-MAHA, MAHA-MU2, KNN-RATIO), and a random baseline. The former two methods are the direct equivalents of LR-COS and COS-MU2, by using the pretrained ResNet18 features to compute scores instead instead of the TF-IDF embeddings, and using the Mahalanobis distance[2] instead of cosine distance. We delete the top score-ranked samples for each deletion budget, re-train a CNN for ten epochs, and report accuracy on the cat test set and accuracy on the other nine classes test set.

Findings. Figure 5 shows that the cat footprint persists in terms of accuracy, which is problematic for privacy scenarios and motivates distributional unlearning, until $\geq 50\%$ data is removed (Fig. 5a), yet the model's utility on the other classes

[2]The Mahalanobis distance of vector x to probability distribution p, of mean μ and covariance Σ, is: $d(x,p) := \sqrt{(x-\mu)^\top \Sigma^{-1}(x-\mu)}$. We estimate μ and Σ empirically on the retained distribution p_2.



(a) Accuracy on cat images (p_1, forgotten).

(b) Accuracy on non-cat images (p_2, kept).

Figure 5. **CIFAR-10 images.** Removing cat images suppresses accuracy on that class (left) while leaving accuracy on the retained nine classes essentially unchanged (right, <0.03 variation). No substantial removal is observed until 50% deletion, before selective removal strategies LR-MAHA and MAHA-MU2 outperform random removal. Error bars: ± 1 standard error over thirty seeds.

remains stable (Fig. 5b) at around 82%. Selective removal strategies reach a given accuracy threshold, on the cat class, with substantially fewer deletions than random. Specifically, LR-MAHA halves the initial accuracy by deleting half of the cat images, which is $1.6\times$ more sample-efficient than random removal. Also, MAHA-MU2 reaches the same accuracy threshold with $1.3\times$ fewer deletions than random removal. These results underscore the value of distributional unlearning: strong class-level forgetting is achieved long before every single cat image is deleted, and with minimal collateral loss on the retained distribution.

D. Proofs

D.1. Pareto Frontier

Proposition 1 (Pareto Frontier)**.** *Let p_1, p_2 be two distributions in $\mathcal{P}$, the class of Gaussian distributions with shared covariance, with $\mathrm{KL}(p_1\|p_2) < \infty$. The Pareto frontier of (α, ε) values achievable by any $p \in \mathcal{P}$ is given by:*

$$\mathrm{PF}(p_1, p_2; \mathcal{P}) = \Big\{ \big(\alpha, \big(\sqrt{\alpha} - \sqrt{\mathrm{KL}(p_1\|p_2)}\big)^2\big) : \alpha \geq \mathrm{KL}(p_1\|p_2) \Big\}.$$

Proof. For simplicity, we consider univariate Gaussians with share variance. For d-dimensional Gaussians with covariance $\Sigma \in \mathbb{R}^{d\times d}$, the same result holdsafter replacing squared error by the Mahalanobis distance $\|x-\mu\|^2_{\Sigma^{-1}}$ (see Proposition 3).

Let $p = \mathcal{N}(\mu, \sigma^2) \in \mathcal{P}$. Since all distributions in $\mathcal{P}$ share the same variance, the KL divergence from p_i to p is:

$$\mathrm{KL}(p_i\|p) = \frac{(\mu_i - \mu)^2}{2\sigma^2}, \quad i = 1, 2.$$

Fix $\alpha \geq \mathrm{KL}(p_1\|p_2)$. We want to compute the minimal possible ε achievable under the constraint $\mathrm{KL}(p_1\|p) \geq \alpha$. Define this minimum as:

$$\varepsilon_\star(\alpha) := \min_{\mu\in\mathbb{R},\ (\mu-\mu_1)^2 \geq 2\sigma^2\alpha} \frac{(\mu_2 - \mu)^2}{2\sigma^2}.$$

This is a one-dimensional quadratic minimization problem subject to a quadratic inequality constraint. The feasible set is:

$$\mu \in (-\infty, \mu_1 - \sigma\sqrt{2\alpha}] \cup [\mu_1 + \sigma\sqrt{2\alpha}, \infty).$$

We minimize $(\mu_2 - \mu)^2$ over this set. This yields two cases:

- If $\mu_2 \in [\mu_1 - \sigma\sqrt{2\alpha}, \mu_1 + \sigma\sqrt{2\alpha}]$, then the closest feasible points are the endpoints. The minimizing value of μ is:

$$\mu = \mu_1 + \mathrm{sign}(\mu_2 - \mu_1) \cdot \sigma\sqrt{2\alpha},$$

and the resulting divergence is:

$$\varepsilon_\star(\alpha) = \frac{(\mu_2 - \mu_1 - \mathrm{sign}(\mu_2 - \mu_1) \cdot \sigma\sqrt{2\alpha})^2}{2\sigma^2}.$$

- If μ_2 already lies in the feasible set, i.e., $|\mu_2 - \mu_1| \geq \sigma\sqrt{2\alpha}$, then we can choose $\mu = \mu_2$, yielding $\varepsilon_\star(\alpha) = 0$.

Thus, for all $\alpha \geq 0$:

$$\varepsilon_\star(\alpha) = \frac{\left[\left(|\mu_2 - \mu_1| - \sigma\sqrt{2\alpha}\right)_+\right]^2}{2\sigma^2},$$

where $(x)_+ = \max\{x, 0\}$. Let $\Delta \coloneqq |\mu_2 - \mu_1|$, and recall that $\mathrm{KL}(p_1 \| p_2) = \frac{\Delta^2}{2\sigma^2}$. Then:

$$\Delta = \sigma\sqrt{2\mathrm{KL}(p_1 \| p_2)}.$$

Substituting into $\varepsilon_\star(\alpha)$:

$$\varepsilon_\star(\alpha) = \left(\sqrt{\alpha} - \sqrt{\mathrm{KL}(p_1 \| p_2)}\right)^2, \quad \text{for } \alpha \geq \mathrm{KL}(p_1 \| p_2).$$

Finally, note that any pair (α, ε) with $\alpha < \mathrm{KL}(p_1 \| p_2)$ satisfies $\varepsilon_\star(\alpha) = 0$, hence is dominated by $(\mathrm{KL}(p_1 \| p_2), 0)$. Therefore, the Pareto optimal points are exactly:

$$\left\{\left(\alpha, \left(\sqrt{\alpha} - \sqrt{\mathrm{KL}(p_1 \| p_2)}\right)^2\right) : \alpha \geq \mathrm{KL}(p_1 \| p_2)\right\},$$

as claimed. □

Next, we show that a qualitatively similar result holds more generally for any exponential-family member.

Proposition 3 (Pareto Frontier–Exponential Families)**.** *Let $(\mathcal{X}, \mu)$ be a measurable space and let*

$$\mathcal{P} = \left\{ p_\theta(x) = h(x)\exp(\theta^\top T(x) - A(\theta)) \,:\, \theta \in \Theta \subset \mathbb{R}^d \right\}$$

be a regular minimal exponential family (carrier $h > 0$, sufficient statistic $T\colon \mathcal{X} \to \mathbb{R}^d$, log-partition A). Fix $p_i(x) = p_{\theta_i}(x), \quad i = 1, 2$, and $\alpha \geq 0$. Define $v(\alpha) = \inf_{\theta\in\Theta}\left\{\mathrm{KL}(p_2 \| p_\theta) \mid \mathrm{KL}(p_1 \| p_\theta) \geq \alpha\right\}$, where $\mathrm{KL}(q \| p) = \int q \log(q/p)\, d\mu$. Then:

(i) The Pareto frontier for points in $\mathcal{P}$ is

$$\mathrm{PF}(p_1, p_2; \mathcal{P}) = \left\{\left(\alpha,\, v(\alpha)\right) : \alpha \geq \mathrm{KL}(p_1 \| p_2)\right\},$$

where $v(\alpha) = \mathrm{KL}(p_2 \| p_1) + \alpha + \frac{1}{\lambda^* - 1}(\theta_2 - \theta_1)^\top(\mathbb{E}_{p_2}[T] - \mathbb{E}_{p_1}[T])$, and λ^* is the unique scalar in $(0, 1)$ such that the distribution $p^* \in \mathcal{P}$ of mean $\mathbb{E}_{p^*}[T] = \frac{\lambda^* \mathbb{E}_{p_1}[T] - \mathbb{E}_{p_2}[T]}{\lambda^* - 1}$ satisfies $\mathrm{KL}(p_1 \| p^*) = \alpha$.

*(ii) **(Gaussian case)** If $\mathcal{P} = \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d\}$ with fixed $\Sigma \succ 0$, then for $\alpha \geq \mathrm{KL}(p_1 \| p_2)$:*

$$v(\alpha) = \left(\sqrt{\alpha} - \sqrt{\mathrm{KL}(p_1 \| p_2)}\right)^2.$$

Proof. In this proof, for a fixed $\alpha \geq 0$, we study the optimal value

$$v(\alpha) = \inf_{\theta\in\Theta}\left\{\mathrm{KL}(p_2 \| p_\theta) \mid \mathrm{KL}(p_1 \| p_\theta) \geq \alpha\right\}.$$

This enables characterzing the Pareto frontier of feasible (α, ε) trade-offs. Below, we focus on the non-trivial case $\alpha > \mathrm{KL}(p_1 \| p_2)$. Indeed, in the case $\alpha \leq \mathrm{KL}(p_1 \| p_2)$, the problem above's unconstrained minimizer p_2 is a feasible solution, so that $v(\alpha) = 0$ for all $\alpha \leq \mathrm{KL}(p_1 \| p_2)$.

Reparametrization and KKT. First, we recall the expression of the KL divergence in exponential families as a Bregman divergence. For any $\theta', \theta \in \Theta$ one has

$$\mathrm{KL}(p_{\theta'} \| p_\theta) = A(\theta) - A(\theta') - (\theta - \theta')^T \nabla A(\theta').$$

We let

$$f(\theta) = \mathrm{KL}(p_2 \| p_\theta), \quad g(\theta) = \mathrm{KL}(p_1 \| p_\theta).$$

Both are continuously differentiable on the open set Θ. The feasible set $\{g(\theta) \geq \alpha\}$ is nonconvex, so we check linear independence constraint qualification (LICQ, (Bertsekas, 1997)) at any minimizer θ^*. To do so, we recall (see (Wainwright et al., 2008)) that for exponential families $\nabla A(\theta) = \mathbb{E}_{p_\theta}[T]$ for any $\theta \in \Theta$, so that

$$\nabla g(\theta) = -\nabla A(\theta_1) + \nabla A(\theta) = -\mathbb{E}_{p_1}[T] + \mathbb{E}_{p_\theta}[T].$$

Minimality of the exponential family ensures $\mathbb{E}_{p_\theta}[T]$ is injective (Wainwright et al., 2008). This together with the fact that $p_1 \neq p_\theta$ for any feasible θ, since $\alpha > 0$, implies that $\nabla g(\theta^*) \neq 0$ and LICQ holds.

Next, we form the Lagrangian

$$\mathcal{L}(\theta, \lambda) = f(\theta) - \lambda\big(g(\theta) - \alpha\big), \quad \lambda \geq 0.$$

Since LICQ holds, KKT conditions are necessary (Bertsekas, 1997) for any minimizer θ^*. Hence, stationarity ($\nabla_\theta \mathcal{L} = 0$) gives

$$\nabla f(\theta^*) = \lambda \nabla g(\theta^*).$$

Given that we had derived $\nabla_\theta \mathrm{KL}(p_i \| p_\theta) = -\mathbb{E}_{p_i}[T] + \mathbb{E}_{p_\theta}[T]$, we obtain

$$-\mathbb{E}_{p_2}[T] + \mathbb{E}_{p^*}[T] = \lambda\,(-\mathbb{E}_{p_1}[T] + \mathbb{E}_{p^*}[T]),$$

hence $(1-\lambda)\,\mathbb{E}_{p^*}[T] = \mathbb{E}_{p_2}[T] - \lambda\,\mathbb{E}_{p_1}[T]$. Observe that we must have $\lambda \neq 1$, as otherwise $\mathbb{E}_{p_2}[T] = \mathbb{E}_{p_1}[T]$, but this contradicts the fact that $p_1 \neq p_1$ given that $\mathbb{E}_{p_\theta}[T]$ is injective for minimal exponential families (Wainwright et al., 2008). Therefore, we have the following:

$$\mathbb{E}_{p^*}[T] = \frac{\lambda\,\mathbb{E}_{p_1}[T] - \mathbb{E}_{p_2}[T]}{\lambda - 1}.$$

Now, we observe that the inequality constraint must be active. Otherwise, the minimizer lies in the interior of the feasible set and is thus a local minimum of the unconstrained problem. The latter admits p_2 as a unique minimizer, so the minimizer at hand must be p_2 but this is not feasible since we assume $\mathrm{KL}(p_1 \| p_2) < \alpha$. Therefore, the minimizer must lie at the boundary of the feasible set, and the inequality constraint is active. That is, we have $g(\theta^*) = \alpha$.

Optimal value. We are now ready to derive the expression of the optimal value:

$$v(\alpha) = f(\theta^\star) = \mathrm{KL}(p_2 \| p^*).$$

We use the following classical Pythagorean-type identity for Bregman divergences (see, e.g., Banerjee et al. (Banerjee et al., 2005)):

$$\mathrm{KL}(p_2 \| p^*) = \mathrm{KL}(p_2 \| p_1) + \mathrm{KL}(p_1 \| p^*) + (\theta_2 - \theta_1)^\top (\nabla A(\theta_1) - \nabla A(\theta^*)).$$

Now, consider the aforementioned KKT multiplier $\lambda > 0, \lambda \neq 1$ such that $\nabla A(\theta^*) = \mathbb{E}_{p^*}[T] = \frac{\lambda\,\mathbb{E}_{p_1}[T] - \mathbb{E}_{p_2}[T]}{\lambda - 1} = \frac{\lambda \nabla A(\theta_1) - \nabla A(\theta_2)}{\lambda - 1}$ as well as $\mathrm{KL}(p_1 \| p^*) = g(\theta^*) = \alpha$. We thus get

$$\begin{aligned} v(\alpha) = \mathrm{KL}(p_2 \| p^*) &= \mathrm{KL}(p_2 \| p_1) + \mathrm{KL}(p_1 \| p^*) + (\theta_2 - \theta_1)^\top (\nabla A(\theta_1) - \nabla A(\theta^*)) \\ &= \mathrm{KL}(p_2 \| p_1) + \mathrm{KL}(p_1 \| p^*) + \frac{1}{\lambda - 1}(\theta_2 - \theta_1)^\top (\nabla A(\theta_2) - \nabla A(\theta_1)) \\ &= \mathrm{KL}(p_2 \| p_1) + \alpha + \frac{1}{\lambda - 1}(\theta_2 - \theta_1)^\top (\nabla A(\theta_2) - \nabla A(\theta_1)). \end{aligned}$$

We now again use that $\nabla A(\theta) = \mathbb{E}_{p_\theta}[T]$ for all $\theta \in \Theta$. We then obtain:

$$v(\alpha) = \mathrm{KL}(p_2 \| p_1) + \alpha + \frac{1}{\lambda - 1}(\theta_2 - \theta_1)^\top (\mathbb{E}_{p_2}[T] - \mathbb{E}_{p_1}[T]). \tag{5}$$

Uniqueness of λ. We now show that there is a unique KKT multiplier λ for the optimal solution. We recall that $\lambda > 0$ is such that $\lambda \neq 1$ and:

$$\mathbb{E}_{p^*}[T] = \frac{\lambda\, \mathbb{E}_{p_1}[T] - \mathbb{E}_{p_2}[T]}{\lambda - 1} \qquad\qquad g(\theta^*) = \mathrm{KL}(p_1 \| p^*) = \alpha.$$

Above, the first equation uniquely defines the distribution p^*, by minimality of the family, and we now show that there is only a unique λ of interest such that the second equations above holds.

Let $\theta^*(\lambda)$ be the unique (by minimality) parameter such that $\mathbb{E}_{p^*}[T] = \frac{\lambda\, \mathbb{E}_{p_1}[T] - \mathbb{E}_{p_2}[T]}{\lambda - 1}$. We define

$$H(\lambda) \coloneqq \mathrm{KL}(p_1 \| p^*) = A(\theta^*) - A(\theta_1) - (\theta^* - \theta_1)^\top \nabla A(\theta_1)$$

Taking the derivative above, we get

$$\frac{dH}{d\lambda}(\lambda) = (\nabla A(\theta^*) - \nabla A(\theta_1))^\top \frac{d\theta^*}{d\lambda}(\lambda).$$

We again use that

$$\nabla A(\theta^*) = \mathbb{E}_{p^*}[T] = \frac{\lambda\, \mathbb{E}_{p_1}[T] - \mathbb{E}_{p_2}[T]}{\lambda - 1} = \frac{\lambda \nabla A(\theta_1) - \nabla A(\theta_2)}{\lambda - 1}.$$

First, replacing in the previous derivative equation, we obtain:

$$\frac{dH}{d\lambda}(\lambda) = (\nabla A(\theta^*) - \nabla A(\theta_1))^\top \frac{d\theta^*}{d\lambda}(\lambda) = \frac{1}{\lambda - 1}(\nabla A(\theta_1) - \nabla A(\theta_2))^\top \frac{d\theta^*}{d\lambda}(\lambda).$$

Second, taking the derivative with respect to λ in the expression of $\nabla A(\theta^*)$ yields:

$$\nabla^2 A(\theta^*) \cdot \frac{d\theta^*}{d\lambda}(\lambda) = \frac{1}{(\lambda - 1)^2}(\nabla A(\theta_1) - \nabla A(\theta_2)).$$

We observe that the Fisher information matrix $\nabla^2 A(\theta^*)$ is positive definite since the family is regular. Multiplying by the inverse of the latter then yields:

$$\frac{d\theta^*}{d\lambda}(\lambda) = \frac{1}{(\lambda - 1)^2} \nabla^2 A(\theta^*)^{-1} (\nabla A(\theta_1) - \nabla A(\theta_2)).$$

Plugging the above in the latest expression of the derivative of H yields:

$$\begin{aligned}\frac{dH}{d\lambda}(\lambda) &= \frac{1}{\lambda - 1}(\nabla A(\theta_1) - \nabla A(\theta_2))^\top \frac{d\theta^*}{d\lambda}(\lambda) \\ &= \frac{1}{(\lambda - 1)^3}(\nabla A(\theta_1) - \nabla A(\theta_2))^\top \nabla^2 A(\theta^*)^{-1} (\nabla A(\theta_1) - \nabla A(\theta_2)).\end{aligned}$$

Since the matrix $\nabla^2 A(\theta^*)$ is positive definite by regularity of the family, the corresponding quadratic form above is positive (recall $\theta_1 \neq \theta_2$), and the sign of the derivative is that of $\lambda - 1$. Therefore, H is decreasing on $(0, 1)$ and increasing on $(1, +\infty)$. It is straighforward to check that $H(0^+) = \mathrm{KL}(p_1 \| p_2)$, $H(1^-) = H(1^+) = +\infty$, and $H(+\infty) = 0$. Since $\mathrm{KL}(p_1 \| p_2) < \alpha$ by assumption, there exists a unique $\lambda^* \in (0, 1)$ such that $H(\lambda^*) = \alpha$ and a unique $\lambda_1^* > 1$ such that $H(\lambda_1^*) = \alpha$.

We now discard λ_1^* thanks to the expression of the optimal value expression (5). Indeed, the second term in (5) is positive for λ_1^* since $\lambda_1^* > 1$ and $(\theta_2 - \theta_1)^\top (\mathbb{E}_{p_2}[T] - \mathbb{E}_{p_1}[T]) = (\theta_2 - \theta_1)^\top (\nabla A(\theta_2) - \nabla A(\theta_1)) > 0$ by strict convexity of A and the fact that $p_1 \neq p_2$. On the other hand, this same second term is negative for λ_* since $\lambda_* < 1$. Therefore, the optimal value is smaller for the choice of λ^*, so that:

$$v(\alpha) = \mathrm{KL}(p_2 \| p_1) + \alpha + \frac{1}{\lambda^* - 1}(\theta_2 - \theta_1)^\top (\mathbb{E}_{p_2}[T] - \mathbb{E}_{p_1}[T]), \tag{6}$$

where λ^* is the unique scalar in $(0, 1)$ such that:

$$\mathbb{E}_{p^*}[T] = \frac{\lambda^*\, \mathbb{E}_{p_1}[T] - \mathbb{E}_{p_2}[T]}{\lambda^* - 1} \qquad\qquad g(\theta^*) = \mathrm{KL}(p_1 \| p^*) = \alpha.$$

Finally, we note that $v(\alpha)$ is non-decreasing by definition; increasing α shrinks the feasible set. Also, we recall that $v(\alpha) = 0$ for all $\alpha < \mathrm{KL}(p_1\|p_2)$, i.e., all trade-offs (α, ε) with $\alpha < \mathrm{KL}(p_1\|p_2), \varepsilon > 0$ are dominated by $(\mathrm{KL}(p_1\|p_2), 0)$. Therefore, we conclude that the pareto frontier is given by:

$$\mathrm{PF}(p_1, p_2; \mathcal{P}) = \Big\{ \big(\alpha,\, v(\alpha)\big) : \alpha \geq \mathrm{KL}(p_1\|p_2) \Big\}.$$

Gaussian case. For $p_\mu = \mathcal{N}(\mu, \Sigma)$ one has $T(x) = x$, $E_{p_\mu}[T] = \mu$, and

$$\mathrm{KL}(p_i\|p_\mu) = \tfrac{1}{2}(\mu_i - \mu)^\top \Sigma^{-1}(\mu_i - \mu).$$

By the conditions on λ^* we have

$$\mu^* = \frac{\lambda^*\mu_1 - \mu_2}{\lambda^* - 1}, \quad \mathrm{KL}(p_1\|p_{\mu^*}) = \alpha.$$

Using the KL divergence expression for Gaussians $\mathcal{N}(\mu, \Sigma)$ (recall $\mathrm{KL}(\mathcal{N}(\mu_1, \Sigma), \mathcal{N}(\mu_2, \Sigma)) = \frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)$), we get

$$\mu_1 - \mu^* = \frac{\mu_2 - \mu_1}{\lambda^* - 1}, \quad \alpha = \frac{\mathrm{KL}(p_1\|p_2)}{(\lambda^* - 1)^2}.$$

Solving for $\lambda^* \in (0, 1)$ yields $\lambda^* = 1 - \sqrt{\mathrm{KL}(p_1\|p_2)/\alpha}$ and

$$\mu^* = \mu_1 + \sqrt{\frac{\alpha}{\mathrm{KL}(p_1\|p_2)}}(\mu_2 - \mu_1).$$

Thus, direct computations yield

$$v(\alpha) = \tfrac{1}{2}\Big(\|\mu_2 - \mu_1\|_{\Sigma^{-1}} - \sqrt{2\alpha}\Big)^2 = \big(\sqrt{\mathrm{KL}(p_1\|p_2)} - \sqrt{\alpha}\big)^2,$$

with $v(\alpha) = 0$ if $\alpha \leq \mathrm{KL}(p_1\|p_2)$. This concludes the proof. □

Discussion. In Proposition 3 we show that, in any regular exponential family, the trade-off between removal (α) and preservation (ε) can be quantified. This yields a removal-preservation trade-off curve that faithfully reproduces the shared-covariance Gaussian Pareto frontier—namely the familiar $(\sqrt{\alpha} - \sqrt{D})^2$ parabola—while in other families it gives an explicit but generally non-algebraic trade-off curve.

D.2. Predictive Performance

Proposition 2. *Let h minimize $\mathcal{L}(h; p)$, and let h_1, h_2 be optimal predictors under $p_1, p_2 \in \mathcal{P}$, respectively. If p satisfies (α, ε)-distributional unlearning with respect to (p_1, p_2), then:*

$$\mathcal{L}(h; p_1) - \mathcal{L}(h_1; p_1) \geq \alpha - \delta_1, \tag{3}$$
$$\mathcal{L}(h; p_2) - \mathcal{L}(h_2; p_2) \leq \varepsilon - \delta_2, \tag{4}$$

where $\delta_1 := \mathrm{KL}(p_{1,X}\|p_X), \delta_2 := \mathrm{KL}(p_{2,X}\|p_X)$ denote the marginal KL divergence over inputs.

Proof. Let $h(y \mid x) := h(x)(y)$ denote the conditional distribution defined by the hypothesis h, and suppose h minimizes the expected log-loss under p. Since $\ell(y, q) = -\log q(y)$ is a strictly proper scoring rule, the unique minimizer of $\mathcal{L}(h; p)$ is the true conditional distribution $h(x) = p(\cdot \mid x)$, where $p(x, y) = p^X(x)p(y \mid x)$.

We begin by analyzing the expected log-loss of this hypothesis under an arbitrary distribution q over $\mathcal{X} \times \mathcal{Y}$:

$$\mathcal{L}(h; q) = \mathbb{E}_{(x,y)\sim q}[-\log h(y \mid x)] = \mathbb{E}_{x\sim q^X}\mathbb{E}_{y\sim q(\cdot|x)}[-\log p(y \mid x)],$$

where q^X denotes the marginal distribution of x under q, and $q(\cdot \mid x)$ the corresponding conditional.

Now recall the standard identity for any two conditional distributions $q(\cdot \mid x)$ and $p(\cdot \mid x)$:

$$\mathbb{E}_{y\sim q(\cdot|x)}[-\log p(y \mid x)] = \mathrm{KL}(q(y \mid x) \,\|\, p(y \mid x)) + H(q(y \mid x)),$$

where $H(q(y \mid x)) = \mathbb{E}_{y\sim q(\cdot|x)}[-\log q(y \mid x)]$ is the Shannon entropy of the label distribution under q for fixed x.

Taking the expectation over $x \sim q^X$, we get:

$$\begin{aligned}\mathcal{L}(h; q) &= \mathbb{E}_{x\sim q^X}\left[\mathrm{KL}(q(y \mid x) \parallel p(y \mid x)) + H(q(y \mid x))\right] \\ &= \mathbb{E}_{x\sim q^X}\mathrm{KL}(q(y \mid x) \parallel p(y \mid x)) + \mathbb{E}_{x\sim q^X} H(q(y \mid x)).\end{aligned}$$

The second term is the expected entropy, which corresponds to the Bayes-optimal risk under q:

$$\mathcal{L}(h_q^*; q) := \inf_{h'} \mathcal{L}(h'; q) = \mathbb{E}_{x\sim q^X} H(q(y \mid x)). \tag{7}$$

Next, we relate the expected conditional KL term to the total KL divergence between the joint distributions. Using the chain rule for KL divergence, we have:

$$\mathrm{KL}(q \parallel p) = \mathrm{KL}(q^X \parallel p^X) + \mathbb{E}_{x\sim q^X}\mathrm{KL}(q(y \mid x) \parallel p(y \mid x)). \tag{8}$$

This decomposition holds generally for joint distributions with conditional factorizations. Solving for the conditional KL term, we obtain:

$$\mathbb{E}_{x\sim q^X}\mathrm{KL}(q(y \mid x) \parallel p(y \mid x)) = \mathrm{KL}(q \parallel p) - \mathrm{KL}(q^X \parallel p^X). \tag{9}$$

Substituting the above into the expression for $\mathcal{L}(h; q)$ and using (7), we get:

$$\mathcal{L}(h; q) = \mathrm{KL}(q \parallel p) - \mathrm{KL}(q^X \parallel p^X) + \mathcal{L}(h_q^*; q). \tag{10}$$

We now apply this to $q = p_1$ and $q = p_2$, noting that p satisfies (α, ε)-distributional unlearning, i.e., $\mathrm{KL}(p_1 \parallel p) \geq \alpha$ and $\mathrm{KL}(p_2 \parallel p) \leq \varepsilon$.

For p_1, we define $\delta_1 := \mathrm{KL}(p_1^X \parallel p^X)$ and compute:

$$\mathcal{L}(h; p_1) - \mathcal{L}(h_1; p_1) = \mathrm{KL}(p_1 \parallel p) - \mathrm{KL}(p_1^X \parallel p^X) = \mathrm{KL}(p_1 \parallel p) - \delta_1 \geq \alpha - \delta_1.$$

For p_2, define $\delta_2 := \mathrm{KL}(p_2^X \parallel p^X)$ and similarly compute:

$$\mathcal{L}(h; p_2) - \mathcal{L}(h_2; p_2) = \mathrm{KL}(p_2 \parallel p) - \mathrm{KL}(p_2^X \parallel p^X) = \mathrm{KL}(p_2 \parallel p) - \delta_2 \leq \varepsilon - \delta_2.$$

This completes the proof. □

D.3. Random Removal

Lemma 3 (Finite-sample concentration)**.** *Let $\hat{\mu}$ be the empirical mean of n samples drawn from $\mathcal{N}(\mu, \sigma^2)$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$|\hat{\mu} - \mu| \leq \sigma\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Proof. This follows directly from Hoeffding's inequality for sub-Gaussian variables. □

Theorem 1 (Random Removal)**.** *Let $p_1, p_2 \in \mathcal{P}$ and $\delta \in (0, 1)$. We observe n_1 samples from p_1 and n_2 samples from p_2, and randomly remove f samples from p_1 before fitting. Then with probability at least $1 - \delta$, the resulting MLE distribution satisfies (α, ε)-distributional unlearning with:*

$$\begin{aligned}\alpha &\geq \big(\frac{1}{2} - 3\big(\tfrac{n_1-f}{n_2}\big)^2\big)\mathrm{KL}(p_1\|p_2) - \frac{3\ln(4/\delta)}{2n_2}\big(1 + \tfrac{n_1-f}{n_2}\big), \\ \varepsilon &\leq 3\big(\tfrac{n_1-f}{n_2}\big)^2\mathrm{KL}(p_1\|p_2) + \frac{3\ln(4/\delta)}{n_2}\big(1 + \tfrac{n_1-f}{n_2}\big).\end{aligned}$$

Proof. We recall that $p_1 = \mathcal{N}(\mu_1, \sigma^2), p_2 = \mathcal{N}(\mu_2, \sigma^2) \in \mathcal{P}$ and $p := \mathcal{N}(\mu, \sigma^2) \in \mathcal{P}$ are univariate Gaussian distributions. We are given n_1 i.i.d. samples $x_1^{(1)}, \dots, x_1^{(n_1)}$ from p_1 and n_2 i.i.d. samples $x_2^{(1)}, \dots, x_2^{(n_2)}$ from p_2.

Upon removing $f \leq n_1$ randomly chosen samples $x_1^{(1)}, \dots, x_1^{(n_1-f)}$ from the target distribution p_1, we set the center μ of the unlearned distribution p to be:

$$\mu = \frac{(n_1 - f)\hat{\mu}_1 + n_2\hat{\mu}_2}{n_1 - f + n_2}, \tag{11}$$

where $\hat{\mu}_1 := \frac{1}{n_1-f}\sum_{i=1}^{n_1-f} x_1^{(i)}$ and $\hat{\mu}_2 := \frac{1}{n_2}\sum_{i=1}^{n_2} x_2^{(i)}$. We also observe that a standard Hoeffding bound (Lemma 3) yields that:

$$|\hat{\mu}_1 - \mu_1| \leq \sigma\sqrt{\frac{2\ln(4/\delta)}{f}}, \quad |\hat{\mu}_2 - \mu_2| \leq \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}, \tag{12}$$

each with probability $1 - \frac{\delta}{2}$, so that both hold with probability $1 - \delta$ thanks to a union bound. We also recall that

$$\mathrm{KL}(p_1 \parallel p) = \frac{(\mu_1 - \mu)^2}{2\sigma^2}, \quad \mathrm{KL}(p_2 \parallel p) = \frac{(\mu_2 - \mu)^2}{2\sigma^2}. \tag{13}$$

Preservation bound. First, we upper bound the KL divergence of p_2 from p. To do so, we first use the triangle inequality to get

$$\begin{aligned}
|\mu - \mu_2| &= \left|\frac{(n_1 - f)\hat{\mu}_1 + n_2\hat{\mu}_2}{n_1 - f + n_2} - \mu_2\right| = \left|\frac{n_1 - f}{n_1 - f + n_2}(\hat{\mu}_1 - \mu_2) + \frac{n_2}{n_1 - f + n_2}(\hat{\mu}_2 - \mu_2)\right| \\
&= \left|\frac{n_1 - f}{n_1 - f + n_2}(\mu_1 - \mu_2) + \frac{n_1 - f}{n_1 - f + n_2}(\hat{\mu}_1 - \mu_1) + \frac{n_2}{n_1 - f + n_2}(\hat{\mu}_2 - \mu_2)\right| \\
&\leq \frac{n_1 - f}{n_1 - f + n_2}|\mu_1 - \mu_2| + \frac{n_1 - f}{n_1 - f + n_2}|\hat{\mu}_1 - \mu_1| + \frac{n_2}{n_1 - f + n_2}|\hat{\mu}_2 - \mu_2|.
\end{aligned}$$

Therefore, using (12) we have with probability $1 - \delta$:

$$|\mu - \mu_2| \leq \frac{n_1 - f}{n_1 - f + n_2}|\mu_1 - \mu_2| + \frac{n_1 - f}{n_1 - f + n_2}\sigma\sqrt{\frac{2\ln(4/\delta)}{f}} + \frac{n_2}{n_1 - f + n_2}\sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}.$$

Taking squares, using Jensen's inequality, and simplifying further since $f \geq 0$, yields:

$$|\mu - \mu_2|^2 \leq 3\left(\frac{n_1 - f}{n_2}\right)^2|\mu_1 - \mu_2|^2 + 3\left(\frac{n_1 - f}{n_2}\right)^2\sigma^2\frac{2\ln(4/\delta)}{n_1 - f} + \sigma^2\frac{6\ln(4/\delta)}{n_2}. \tag{14}$$

Dividing both sides by $2\sigma^2$ and then using (30) yields with probability $1 - \delta$:

$$\mathrm{KL}(p_2 \parallel p) \leq 3\left(\frac{n_1 - f}{n_2}\right)^2\mathrm{KL}(p_1 \parallel p_2) + \frac{3\ln(4/\delta)}{n_2}\left(1 + \frac{n_1 - f}{n_2}\right). \tag{15}$$

Removal bound. Second, we lower bound the KL divergence of p_1 from p. To do so, we use Jensen's inequality and (14) to obtain that, with probability $1 - \delta$, we have

$$\begin{aligned}
|\mu_1 - \mu_2|^2 &= |\mu_1 - \mu + \mu - \mu_2|^2 \leq 2|\mu_1 - \mu|^2 + 2|\mu - \mu_2|^2 \\
&\leq 2|\mu_1 - \mu|^2 + 6\left(\frac{n_1 - f}{n_2}\right)^2|\mu_1 - \mu_2|^2 + \frac{6\sigma^2\ln(4/\delta)}{n_2}\left(1 + \frac{n_1 - f}{n_2}\right)
\end{aligned}$$

Rearranging terms and dividing by $4\sigma^2$ along with (30) yields that with probability $1 - \delta$ we have

$$\begin{aligned}
\mathrm{KL}(p_1 \parallel p) = \frac{|\mu_1 - \mu|^2}{2\sigma^2} &\geq \frac{|\mu_1 - \mu_2|^2}{4\sigma^2} - 3\left(\frac{n_1 - f}{n_2}\right)^2\frac{|\mu_1 - \mu_2|^2}{2\sigma^2} - \frac{3\ln(4/\delta)}{2n_2}\left(1 + \frac{n_1 - f}{n_2}\right) \\
&= \left(\frac{1}{2} - 3\left(\frac{n_1 - f}{n_2}\right)^2\right)\mathrm{KL}(p_1 \parallel p_2) - \frac{3\ln(4/\delta)}{2n_2}\left(1 + \frac{n_1 - f}{n_2}\right).
\end{aligned}$$

Conclusion. With probability $1-\delta$, we have (α, ε)-distributional unlearning with

$$\alpha \geq \left(\frac{1}{2} - 3\left(\frac{n_1 - f}{n_2}\right)^2\right) \mathrm{KL}(p_1 \parallel p_2) - \frac{3\ln(4/\delta)}{2n_2}\left(1 + \frac{n_1 - f}{n_2}\right),$$
$$\varepsilon \leq 3\left(\frac{n_1 - f}{n_2}\right)^2 \mathrm{KL}(p_1 \parallel p_2) + \frac{3\ln(4/\delta)}{n_2}\left(1 + \frac{n_1 - f}{n_2}\right).$$

□

D.4. Selective Removal

Lemma 4 (Dvoretzky–Kiefer–Wolfowitz Inequality)**.** *Let $x_1, x_2, \ldots, x_n$ be independent and identically distributed random variables with cumulative distribution function F. Define the empirical distribution function by*

$$\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \leq t\}.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$\sup_{t\in\mathbb{R}} \left|\widehat{F}(t) - F(t)\right| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Lemma 5. *Let $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma > 0$. Consider n_1 i.i.d. samples $x_1^{(1)}, \ldots, x_1^{(n_1)}$ from $\mathcal{N}(\mu_1, \sigma^2)$ and n_2 i.i.d. samples $x_2^{(1)}, \ldots, x_2^{(n_2)}$ from $\mathcal{N}(\mu_2, \sigma^2)$. We define $\hat{\mu}_2$ the average of the samples from $\mathcal{N}(\mu_2, \sigma^2)$, $\hat{\mu}_1$ the average of the $n_1 - f \leq n_1$ closest samples from $x_1^{(1)}, \ldots, x_1^{(n_1)}$ to $\hat{\mu}_2$. We define $F : t \in \mathbb{R} \mapsto \Phi(\frac{t-|\mu_1-\mu_2|}{\sigma}) - \Phi(\frac{-t-|\mu_1-\mu_2|}{\sigma})$, where Φ is the standard normal CDF.*

For any $\delta \in (0,1)$, we have with probability $1-\delta$,

$$|\hat{\mu}_1 - \mu_2| \leq F^{-1}\left(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(2/\delta)}{2n_1}}\right). \tag{16}$$

Proof. Recall from Equation (27) that $\hat{\mu}_1$ is the average of the $n_1 - f$ samples, out of n_1 i.i.d. from p_1, with the closest distance to $\hat{\mu}_2$, the empirical mean of n_2 samples from p_2.

Denote by $\hat{\tau}_f := |x_1^{(n_1-f:n_1)} - \hat{\mu}_2|$ the $(n_1 - f)$-th largest distance of $\hat{\mu}_2$ to p_1 samples. It is then immediate from the triangle inequality that

$$|\hat{\mu}_1 - \mu_2| = |\frac{1}{n_1 - f}\sum_{i=1}^{n_1-f} x_1^{(i:n_1)} - \mu_2| \leq \frac{1}{n_1 - f}\sum_{i=1}^{n_1-f} |x_1^{(i:n_1)} - \mu_2| \leq \hat{\tau}_f. \tag{17}$$

Besides, denoting by $\widehat{F}_1$ the empirical CDF of the empirical distribution over $\left\{|x_1^{(i)} - \mu_2| \,:\, i \in [n_1]\right\}$, we have for all $t \in \mathbb{R}$:

$$\widehat{F}_1(t) = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{\left\{|x_1^{(i)} - \mu_2| \leq t\right\}}. \tag{18}$$

Yet, we recall that with probability $1 - \frac{\delta}{2}$, we have

$$|\hat{\mu}_2 - \mu_2| \leq \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}. \tag{19}$$

Therefore, the triangle inequality gives for $i \in [n_1]$, $|x_1^{(i)} - \mu_2| \leq |x_1^{(i)} - \hat{\mu}_2| + |\hat{\mu}_2 - \mu| \leq |x_1^{(i)} - \hat{\mu}_2| + \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}$, and we deduce

$$\widehat{F}_1(t) = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{\left\{|x_1^{(i)} - \mu_2| \leq t\right\}} \leq \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{\left\{|x_1^{(i)} - \hat{\mu}_2| \leq t - \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}\right\}}. \tag{20}$$

In particular, by definition of $\hat{\tau}_f$, we have

$$\widehat{F}_1(\hat{\tau}_f + \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}) \leq \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{\left\{|x_1^{(i)} - \hat{\mu}_2| \leq \hat{\tau}_f\right\}} = \frac{n_1 - f}{n_1} = 1 - \frac{f}{n_1}. \tag{21}$$

Now, observe that $|x_1^{(i)} - \mu_2|$ follows a folded normal distribution of location $\mu_1 - \mu_2$ and scale σ^2, since $x_1^{(i)}$ follows $p_1 = \mathcal{N}(\mu_1, \sigma^2)$. Denote by F_1 its CDF. Thanks to the Dvoretzky–Kiefer–Wolfowitz inequality (Lemma 4), we have with probability $1 - \frac{\delta}{2}$ that for all $t \in \mathbb{R}$,

$$|\widehat{F}_1(t) - F_1(t)| \leq \sqrt{\frac{\ln(4/\delta)}{2n_1}}. \tag{22}$$

Plugging the above in the previous inequality, and using a union bound, we get with probability $1 - \delta$,

$$F_1(\hat{\tau}_f + \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}) \leq \widehat{F}_1(\hat{\tau}_f + \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}) + \sqrt{\frac{\ln(4/\delta)}{2n_1}} \leq 1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}. \tag{23}$$

By taking the inverse F_1^{-1} of the CDF F_1 and rearranging terms, we obtain with probability $1 - \delta$ that

$$\hat{\tau}_f \leq F_1^{-1}\left(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\right) - \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}. \tag{24}$$

Finally, going back to (17), we obtain with probability $1 - \delta$ that

$$|\hat{\mu}_1 - \mu_2| \leq \hat{\tau}_f \leq F_1^{-1}\left(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\right) - \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}. \tag{25}$$

□

Lemma 6. *Let $\mu_1, \mu_2 \in \mathbb{R}$ and suppose that $x \sim \mathcal{N}(\mu_1, \sigma^2)$. Define the random variable $z := |x - \mu_2|$, with cumulative distribution function*

$$F_\sigma(t) := \mathbb{P}[z \leq t] = \Phi\Big(\frac{t - |\mu_1 - \mu_2|}{\sigma}\Big) - \Phi\Big(\frac{-t - |\mu_1 - \mu_2|}{\sigma}\Big), \quad t \geq 0,$$

where Φ denotes the standard normal CDF. Then, for any $p \in (0, 1)$ the inverse CDF satisfies

$$F_\sigma^{-1}(p) = \sigma\, g^{-1}\Big(p; \frac{|\mu_1 - \mu_2|^2}{2\sigma^2}\Big),$$

where the function $g(u; \kappa)$ is defined by $g(u; \kappa) := \Phi(u - \sqrt{2\kappa}) + \Phi(u + \sqrt{2\kappa}) - 1$, and $g^{-1}(p; \kappa)$ denotes the inverse function in u satisfying $g(u; \kappa) = p$. In particular, when $\mu_1 = \mu_2$ (so that $\kappa = 0$) we have $g(u; 0) = 2\Phi(u) - 1$ and thus $F_{0,1}^{-1}(p) = \Phi^{-1}\Big(\frac{p+1}{2}\Big)$.

Proof. Since $x \sim \mathcal{N}(\mu_1, \sigma^2)$, we have that $z = |x - \mu_2|$ has CDF

$$F_\sigma(t) = \Phi\Big(\frac{t - |\mu_1 - \mu_2|}{\sigma}\Big) - \Phi\Big(\frac{-t - |\mu_1 - \mu_2|}{\sigma}\Big), \quad t \geq 0.$$

Introduce the change of variable $u = \frac{t}{\sigma}$ so that $t = \sigma\, u$. Then,

$$F_\sigma(\sigma\, u) = \Phi\Big(u - \frac{|\mu_1 - \mu_2|}{\sigma}\Big) - \Phi\Big(-u - \frac{|\mu_1 - \mu_2|}{\sigma}\Big).$$

Using the symmetry $\Phi(-x) = 1 - \Phi(x)$, this becomes

$$F_\sigma(\sigma\, u) = \Phi\Big(u - \frac{|\mu_1 - \mu_2|}{\sigma}\Big) + \Phi\Big(u + \frac{|\mu_1 - \mu_2|}{\sigma}\Big) - 1.$$

Defining $\kappa = \frac{|\mu_1 - \mu_2|^2}{2\sigma^2}$ and setting

$$g(u;\kappa) \coloneqq \Phi(u - \sqrt{2\kappa}) + \Phi(u + \sqrt{2\kappa}) - 1,$$

we have $F_\sigma(\sigma\, u) = g(u;\kappa)$. Thus, if u^* is the unique solution of $g(u^*;\kappa) = p$, then

$$F_\sigma(\sigma\, u^*) = p,$$

so that

$$F_\sigma^{-1}(p) = \sigma\, u^* = \sigma\, g^{-1}(p;\kappa).$$

In the special case $\mu_1 = \mu_2$ (so that $\kappa = 0$), we obtain $g(u;0) = 2\Phi(u) - 1$, whose inverse is given by $u = \Phi^{-1}((p+1)/2)$. Hence, $F_1^{-1}(p) = \Phi^{-1}((p+1)/2)$, as required. □

Theorem 2 (Selective Removal)**.** *Let $p_1, p_2 \in \mathcal{P}$ and $\delta \in (0,1)$. Let f samples from p_1 be removed according to Selective Removal. Then with probability at least $1-\delta$, the resulting estimate satisfies (α,ε)-distributional unlearning with:*

$$\alpha \geq \frac{1}{2}\mathrm{KL}(p_1\|p_2) - \frac{1}{2}\left(\tfrac{n_1-f}{n_2}\right)^2 \Gamma(f, n_1, \delta, \mathrm{KL}(p_1\|p)) - \tfrac{\ln(4/\delta)}{n_2},$$
$$\varepsilon \leq \left(\tfrac{n_1-f}{n_2}\right)^2 \Gamma(f, n_1, \delta, \mathrm{KL}(p_1\|p)) + \tfrac{2\ln(4/\delta)}{n_2},$$

where $\Gamma(f, n_1, \delta, \mathrm{KL}(p_1\|p)) \coloneqq g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1\|p_2)\Big)^2$ with $g(u;\kappa) \coloneqq \Phi(u - \sqrt{2\kappa}) + \Phi(u + \sqrt{2\kappa}) - 1$, for $u, \kappa > 0$, and Φ is the standard normal CDF.

Proof. We recall that $p_1 = \mathcal{N}(\mu_1, \sigma^2), p_2 = \mathcal{N}(\mu_2, \sigma^2) \in \mathcal{P}$ and $p \coloneqq \mathcal{N}(\mu, \sigma^2) \in \mathcal{P}$ are univariate Gaussian distributions. We are given n_1 i.i.d. samples $x_1^{(1)}, \ldots, x_1^{(n_1)}$ from p_1 and n_2 i.i.d. samples $x_2^{(1)}, \ldots, x_2^{(n_2)}$ from p_2.

The distance-based selection removes $f \leq n_1$ selected samples from the target distribution p_1 with the f largest distances to $\hat{\mu}_2 \coloneqq \frac{1}{n_2}\sum_{i=1}^{n_2} x_2^{(i)}$ the empirical estimator of the mean of p_2. That is, denoting by $x_1^{(1:n_1)}, \ldots, x_1^{(n_1:n_1)}$ the original n_1 samples from p_1 reordered by increasing distance to $\hat{\mu}_2$:

$$|x_1^{(1:n_1)} - \hat{\mu}_2| \leq \ldots \leq |x_1^{(n_1:n_1)} - \hat{\mu}_2|, \tag{26}$$

with ties broken arbitrarily, then distance-based selection retains only $x_1^{(1:n_1)}, \ldots, x_1^{(n_1-f:n1)}$ to obtain

$$\hat{\mu}_1 \coloneqq \tfrac{1}{n_1-f}\sum_{i=1}^{n_1-f} x_1^{(i:n_1)}. \tag{27}$$

Subsequently, we set the center μ of the unlearned distribution p to be:

$$\mu = \frac{(n_1-f)\hat{\mu}_1 + n_2\hat{\mu}_2}{n_1 - f + n_2}, \tag{28}$$

where $\hat{\mu}_1 = \frac{1}{n_1-f}\sum_{i=1}^{n_1-f} x_1^{(i:n_1)}$ and $\hat{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_2^{(i)}$. We also observe that a standard Hoeffding bound (Lemma 3) yields that:

$$|\hat{\mu}_2 - \mu_2| \leq \sigma\sqrt{\frac{2\ln(4/\delta)}{n_2}}, \tag{29}$$

with probability $1 - \frac{\delta}{2}$. We also recall that

$$\mathrm{KL}(p_1 \parallel p) = \frac{(\mu_1 - \mu)^2}{2\sigma^2}, \quad \mathrm{KL}(p_2 \parallel p) = \frac{(\mu_2 - \mu)^2}{2\sigma^2}. \tag{30}$$

Preservation bound. First, we upper bound the KL divergence of p_2 from p. To do so, we first use the triangle inequality to get

$$|\mu - \mu_2| = \left| \frac{(n_1 - f)\hat{\mu}_1 + n_2\hat{\mu}_2}{n_1 - f + n_2} - \mu_2 \right| = \left| \frac{n_1 - f}{n_1 - f + n_2}(\hat{\mu}_1 - \mu_2) + \frac{n_2}{n_1 - f + n_2}(\hat{\mu}_2 - \mu_2) \right|$$
$$\leq \frac{n_1 - f}{n_1 - f + n_2} |\hat{\mu}_1 - \mu_2| + \frac{n_2}{n_1 - f + n_2} |\hat{\mu}_2 - \mu_2| .$$

Therefore, using (29) we have with probability $1 - \frac{\delta}{2}$:

$$|\mu - \mu_2| \leq \frac{n_1 - f}{n_1 - f + n_2} |\hat{\mu}_1 - \mu_2| + \frac{n_2}{n_1 - f + n_2} \sigma \sqrt{\frac{2\ln(4/\delta)}{n_2}}.$$

Moreover, we know from Lemma 5 that with probabilty $1 - \frac{\delta}{2}$

$$|\hat{\mu}_1 - \mu_2| \leq F^{-1}\left(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\right).$$

Using the above in the previous inequality with a union bound, yields that with probability $1 - \delta$

$$|\mu - \mu_2| \leq \frac{n_1 - f}{n_1 - f + n_2} F^{-1}\left(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\right) + \frac{n_2}{n_1 - f + n_2} \sigma \sqrt{\frac{2\ln(4/\delta)}{n_2}}.$$

We can further simplify the above using Lemma 6, which implies that for all $p > 0$

$$F^{-1}(p) = \sigma\, g^{-1}\Big(p; \frac{|\mu_1 - \mu_2|^2}{2\sigma^2}\Big),$$

where the function g is defined by $g(u; \kappa) := \Phi(u - \sqrt{2\kappa}) + \Phi(u + \sqrt{2\kappa}) - 1$, for $u, \kappa > 0$. Plugging this in the previous bound yields

$$|\mu - \mu_2| \leq \frac{n_1 - f}{n_1 - f + n_2} \sigma\, g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \frac{|\mu_1 - \mu_2|^2}{2\sigma^2}\Big) + \frac{n_2}{n_1 - f + n_2} \sigma \sqrt{\frac{2\ln(4/\delta)}{n_2}}. \tag{31}$$

Dividing both sides by $\sigma\sqrt{2}$ and then using (30) yields with probability $1 - \delta$:

$$\sqrt{\mathrm{KL}(p_2 \parallel p)} \leq \frac{n_1 - f}{(n_1 - f + n_2)\sqrt{2}} g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1 \parallel p_2)\Big) + \frac{\sqrt{n_2 \log(4/\delta)}}{n_1 - f + n_2}. \tag{32}$$

The above directly implies, by taking squares and using Jensen's inequality and that $f \geq 0$, that with probability $1 - \delta$:

$$\mathrm{KL}(p_2 \parallel p) \leq \left(\frac{n_1 - f}{n_2}\right)^2 g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1 \parallel p_2)\Big)^2 + \frac{2\ln(4/\delta)}{n_2}. \tag{33}$$

Removal bound. Second, we lower bound the KL divergence of p_1 from p. To do so, we use Jensen's inequality and (31) to obtain that, with probability $1 - \delta$, we have

$$|\mu_1 - \mu_2|^2 = |\mu_1 - \mu + \mu - \mu_2|^2 \leq 2|\mu_1 - \mu|^2 + 2|\mu - \mu_2|^2$$
$$\leq 2|\mu_1 - \mu|^2 + 2\left(\frac{n_1 - f}{n_1 - f + n_2}\right)^2 \sigma^2\, g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \frac{|\mu_1 - \mu_2|^2}{2\sigma^2}\Big)^2$$
$$+ \left(\frac{n_2}{n_1 - f + n_2}\right)^2 \sigma^2 \frac{4\ln(4/\delta)}{n_2}.$$

Rearranging terms, using that $f \geq 0$, and dividing by $4\sigma^2$ along with (30) yields that with probability $1-\delta$ we have

$$\begin{aligned}\mathrm{KL}(p_1 \parallel p) = \frac{|\mu_1-\mu|^2}{2\sigma^2} &\geq \frac{|\mu_1-\mu_2|^2}{4\sigma^2} - \frac{1}{2}\left(\frac{n_1-f}{n_2}\right)^2 g^{-1}\Big(1-\frac{f}{n_1}+\sqrt{\frac{\ln(4/\delta)}{2n_1}}; \frac{|\mu_1-\mu_2|^2}{2\sigma^2}\Big)^2 - \frac{\ln(4/\delta)}{n_2} \\ &= \frac{1}{2}\mathrm{KL}(p_1 \parallel p_2) - \frac{1}{2}\left(\frac{n_1-f}{n_2}\right)^2 g^{-1}\Big(1-\frac{f}{n_1}+\sqrt{\frac{\ln(4/\delta)}{2n_1}}; \frac{|\mu_1-\mu_2|^2}{2\sigma^2}\Big)^2 - \frac{\ln(4/\delta)}{n_2}.\end{aligned}$$

Now, using (30) we get

$$\mathrm{KL}(p_1 \parallel p) \geq \frac{1}{2}\mathrm{KL}(p_1 \parallel p_2) - \frac{1}{2}\left(\frac{n_1-f}{n_2}\right)^2 g^{-1}\Big(1-\frac{f}{n_1}+\sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1 \parallel p_2)\Big)^2 - \frac{\ln(4/\delta)}{n_2}$$

Conclusion. With probability $1-\delta$, we have (α, ε)-distributional unlearning with

$$\begin{aligned}\alpha &\geq \frac{1}{2}\mathrm{KL}(p_1 \parallel p_2) - \frac{1}{2}\left(\frac{n_1-f}{n_2}\right)^2 g^{-1}\Big(1-\frac{f}{n_1}+\sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1 \parallel p_2)\Big)^2 - \frac{\ln(4/\delta)}{n_2}, \\ \varepsilon &\leq \left(\frac{n_1-f}{n_2}\right)^2 g^{-1}\Big(1-\frac{f}{n_1}+\sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1 \parallel p_2)\Big)^2 + \frac{2\ln(4/\delta)}{n_2}.\end{aligned}$$

□

D.5. Simplified Sample Complexity for Selective Removal

In this section, we can simplify the result of Theorem 2 on selective removal by simplifying cumbersome terms. This leads to Corollary 8, which then yields to the sample complexity results in Table 1.

We first prove an upper bound on the inverse CDF of a folded Normal for small quantiles.

Lemma 7. *Let $\mu_1, \mu_2 \in \mathbb{R}, \sigma > 0$, and $x \sim \mathcal{N}(\mu_1, \sigma^2)$. Define the random variable $z = |x-\mu_2|$, which follows a folded normal distribution whose cumulative distribution function (CDF) is given by*

$$F(t) \coloneqq \mathbb{P}[z \leq t] = \Phi\Big(\frac{t-|\mu_1-\mu_2|}{\sigma}\Big) - \Phi\Big(\frac{-t-|\mu_1-\mu_2|}{\sigma}\Big), \quad t \geq 0, \tag{34}$$

where Φ denotes the standard normal CDF. Then, for any $p > 0$ such that $F^{-1}(p) \leq |\mu_1-\mu_2|$, it holds that:

$$F^{-1}(p) \leq \frac{\sigma\, p}{\varphi\Big(\frac{|\mu_1-\mu_2|}{\sigma}\Big)},$$

where $\varphi(x) \coloneqq \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big), x \in \mathbb{R}$, is the standard normal density.

Proof. Since F, defined in (34), is continuously differentiable and strictly increasing on $[0, |\mu_1-\mu_2|]$ (with $F(0)=0$) and by the assumption $F^{-1}(p) \leq |\mu_1-\mu_2|$, the Mean Value Theorem guarantees that there exists some $\xi \in [0, F^{-1}(p)]$ such that

$$F\Big(F^{-1}(p)\Big) = F(0) + F'(\xi)\Big(F^{-1}(p) - 0\Big).$$

Since $F(0) = 0$ and F is strictly increasing, one may directly write, via the Mean Value Theorem, that there exists $\xi \in [0, F^{-1}(p)]$ with

$$F\Big(F^{-1}(p)\Big) = F'(\xi)\, F^{-1}(p).$$

By definition of the inverse CDF, $F\Big(F^{-1}(p)\Big) = p$; hence,

$$p = F'(\xi)\, F^{-1}(p).$$

It remains to lower-bound $F'(\xi)$ for $\xi \in [0, |\mu_1 - \mu_2|]$. We recall that for $t \geq 0$,

$$F(t) = \Phi\Big(\frac{t - |\mu_1 - \mu_2|}{\sigma}\Big) - \Phi\Big(\frac{-t - |\mu_1 - \mu_2|}{\sigma}\Big).$$

Taking the derivative with respect to t yields

$$\begin{aligned} F'(t) &= \frac{1}{\sigma}\,\varphi\Big(\frac{t - |\mu_1 - \mu_2|}{\sigma}\Big) + \frac{1}{\sigma}\,\varphi\Big(\frac{-t - |\mu_1 - \mu_2|}{\sigma}\Big) \\ &= \frac{1}{\sigma}\,\varphi\Big(\frac{t - |\mu_1 - \mu_2|}{\sigma}\Big) + \frac{1}{\sigma}\,\varphi\Big(\frac{t + |\mu_1 - \mu_2|}{\sigma}\Big), \end{aligned}$$

where we used that the standard normal density ϕ is symmetric.

For $t \in [0, |\mu_1 - \mu_2|]$, observe that $t - |\mu_1 - \mu_2| \leq 0$. Because the standard normal density is symmetric and nonincreasing on $[0, \infty)$, we have

$$\varphi\Big(\frac{t - |\mu_1 - \mu_2|}{\sigma}\Big) = \varphi\Big(\frac{|\mu_1 - \mu_2| - t}{\sigma}\Big) \geq \varphi\Big(\frac{|\mu_1 - \mu_2|}{\sigma}\Big).$$

Also, the second term $\varphi\Big(\frac{t+|\mu_1-\mu_2|}{\sigma}\Big)$ is nonnegative. Hence, for all $t \in [0, |\mu_1 - \mu_2|]$ we have

$$F'(t) \geq \frac{1}{\sigma}\,\varphi\Big(\frac{|\mu_1 - \mu_2|}{\sigma}\Big).$$

In particular, at $t = \xi$ we obtain

$$F'(\xi) \geq \frac{1}{\sigma}\,\varphi\Big(\frac{|\mu_1 - \mu_2|}{\sigma}\Big).$$

Substituting this lower bound into the equation $p = F'(\xi)\,F^{-1}(p)$ yields

$$p \geq \frac{F^{-1}(p)}{\sigma}\,\varphi\Big(\frac{|\mu_1 - \mu_2|}{\sigma}\Big).$$

Rearranging the inequality gives the desired upper bound:

$$F^{-1}(p) \leq \frac{\sigma\, p}{\varphi\Big(\frac{|\mu_1-\mu_2|}{\sigma}\Big)}.$$

This completes the proof. □

We are now ready to prove the corollary below. Note that retaining dependences on α, ε only in this corollary leads to the result of Table 1.

Corollary 8. *Let $p_1, p_2 \in \mathcal{P}$ and $\delta \in (0,1)$. Let f samples from p_1 be removed according to our distance-based scoring rule or at random before MLE. Then in each case, with probability at least $1-\delta$, the resulting estimate satisfies (α, ε)-distributional unlearning if:*

1. *Random removal: $n_2 \geq \frac{12\ln(4/\delta)}{\min\{\varepsilon,\alpha\}}$ and $\mathrm{KL}(p_1 \parallel p_2) \geq 8\alpha$, and*

$$f \geq n_1 - n_2\sqrt{\frac{2\mathrm{KL}(p_1 \parallel p_2) - \alpha}{12\mathrm{KL}(p_1 \parallel p_2)}}, \qquad f \geq n_1 - n_2 \min\left\{1, \sqrt{\frac{\varepsilon}{6\mathrm{KL}(p_1 \parallel p_2)}}\right\}.$$

2. *Selective removal: $n_2 \geq 2\ln(4/\delta)\max\left\{\frac{1}{\varepsilon}, \frac{1}{\sqrt{\varepsilon}}, \frac{1}{\alpha}, \sqrt{\mathrm{KL}(p_1 \parallel p_2) - 4\alpha}\right\}$, $\mathrm{KL}(p_1 \parallel p_2) \geq 4\alpha$, and $f \geq n_1\left(\frac{3}{2} + \sqrt{\frac{\ln(4/\delta)}{2n_1}} - \Phi(2\sqrt{2\mathrm{KL}(p_1 \parallel p_2)})\right)$, and*

$$f \geq n_1 - \sqrt{n_1 n_2}\left(\frac{\varepsilon}{16\pi}\right)^{1/4}\exp(-\mathrm{KL}(p_1 \parallel p_2)),$$

$$f \geq n_1 - \sqrt{n_1 n_2}\left(\frac{\mathrm{KL}(p_1 \parallel p_2) - 4\alpha}{8\pi}\right)^{1/4}\exp(-\mathrm{KL}(p_1 \parallel p_2)).$$

Proof. We treat random and selective removal separately below.

Random Removal. Consider removing f samples using the random removal mechanism, before maximum likelihood estimation. From Theorem 1, with probability $1-\delta$, we achieve (α, ε)-unlearning with:

$$\alpha \geq \left(\frac{1}{2} - 3\left(\frac{n_1 - f}{n_2}\right)^2\right) \mathrm{KL}(p_1 \parallel p_2) - \frac{3\ln(4/\delta)}{2n_2}\left(1 + \frac{n_1 - f}{n_2}\right) \qquad \text{(removal)},$$

$$\varepsilon \leq 3\left(\frac{n_1 - f}{n_2}\right)^2 \mathrm{KL}(p_1 \parallel p_2) + \frac{3\ln(4/\delta)}{n_2}\left(1 + \frac{n_1 - f}{n_2}\right) \qquad \text{(preservation)}.$$

Therefore, assuming that $n_2 \geq \frac{12\ln(4/\delta)}{\min\{\varepsilon,\alpha\}}$ and $\mathrm{KL}(p_1 \parallel p_2) \geq 8\alpha$, direct calculations show that it is sufficient to set:

$$f \geq n_1 - n_2\sqrt{\frac{2\mathrm{KL}(p_1 \parallel p_2) - \alpha}{12\mathrm{KL}(p_1 \parallel p_2)}} \qquad \text{(removal)},$$

$$f \geq n_1 - n_2 \min\left\{1, \sqrt{\frac{\varepsilon}{6\mathrm{KL}(p_1 \parallel p_2)}}\right\} \qquad \text{(preservation)}.$$

Selective Removal. For selective removal, Theorem 2 shows that, with probability $1-\delta$, we achieve (α, ε)-unlearning with:

$$\alpha \geq \frac{1}{2}\mathrm{KL}(p_1 \parallel p_2) - \frac{1}{2}\left(\frac{n_1 - f}{n_2}\right)^2 g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1 \parallel p_2)\Big)^2 - \frac{\ln(4/\delta)}{n_2} \qquad \text{(removal)},$$

$$\varepsilon \leq \left(\frac{n_1 - f}{n_2}\right)^2 g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1 \parallel p_2)\Big)^2 + \frac{2\ln(4/\delta)}{n_2} \qquad \text{(preservation)}.$$

We can simplify these bounds using Lemma 7. Indeed, thanks to the latter and using the same notation and a simple change of variable, we have for all $p, \kappa > 0$ such that $p \leq F(|\mu_1 - \mu_2|)$

$$g^{-1}\Big(p; \kappa\Big) \leq \frac{p}{\phi(\sqrt{2\kappa})} = p\sqrt{2\pi}\exp(\kappa). \tag{35}$$

Now, we plug in $p = 1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}$. Assuming $f \geq n_1\left(-\Phi(2\sqrt{2\mathrm{KL}(p_1 \parallel p_2)}) + \frac{3}{2} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\right)$, which directly implies that $p \leq F(|\mu_1 - \mu_2|)$ as required, we then obtain

$$g^{-1}\Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}; \mathrm{KL}(p_1 \parallel p_2)\Big) \leq \Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\Big)\sqrt{2\pi}\exp\left(\mathrm{KL}(p_1 \parallel p_2)\right).$$

Plugging the above back in the first bounds due to Theorem 2, we obtain:

$$\alpha \geq \frac{1}{2}\mathrm{KL}(p_1 \parallel p_2) - \frac{1}{2}\left(\frac{n_1 - f}{n_2}\right)^2 \Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\Big)^2 2\pi\exp\left(2\mathrm{KL}(p_1 \parallel p_2)\right) - \frac{\ln(4/\delta)}{n_2},$$

$$\varepsilon \leq \left(\frac{n_1 - f}{n_2}\right)^2 \Big(1 - \frac{f}{n_1} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\Big)^2 2\pi\exp\left(2\mathrm{KL}(p_1 \parallel p_2)\right) + \frac{2\ln(4/\delta)}{n_2}.$$

Therefore, assuming that $n_2 \geq 2\ln(4/\delta)\max\left\{\frac{1}{\varepsilon}, \frac{1}{\sqrt{\varepsilon}}, \frac{1}{\alpha}, \sqrt{\mathrm{KL}(p_1 \parallel p_2) - 4\alpha}\right\}$ and $\mathrm{KL}(p_1 \parallel p_2) \geq 4\alpha$, and recalling we had assumed $f \geq n_1\left(-\Phi(2\sqrt{2\mathrm{KL}(p_1 \parallel p_2)}) + \frac{3}{2} + \sqrt{\frac{\ln(4/\delta)}{2n_1}}\right)$, direct calculations show that it is sufficient to set:

$$f \geq n_1 - \sqrt{n_1 n_2}\left(\frac{\varepsilon}{16\pi}\right)^{1/4}\exp(-\mathrm{KL}(p_1 \parallel p_2)) \qquad \text{(removal)},$$

$$f \geq n_1 - \sqrt{n_1 n_2}\left(\frac{\mathrm{KL}(p_1 \parallel p_2) - 4\alpha}{8\pi}\right)^{1/4}\exp(-\mathrm{KL}(p_1 \parallel p_2)) \qquad \text{(preservation)}.$$

□

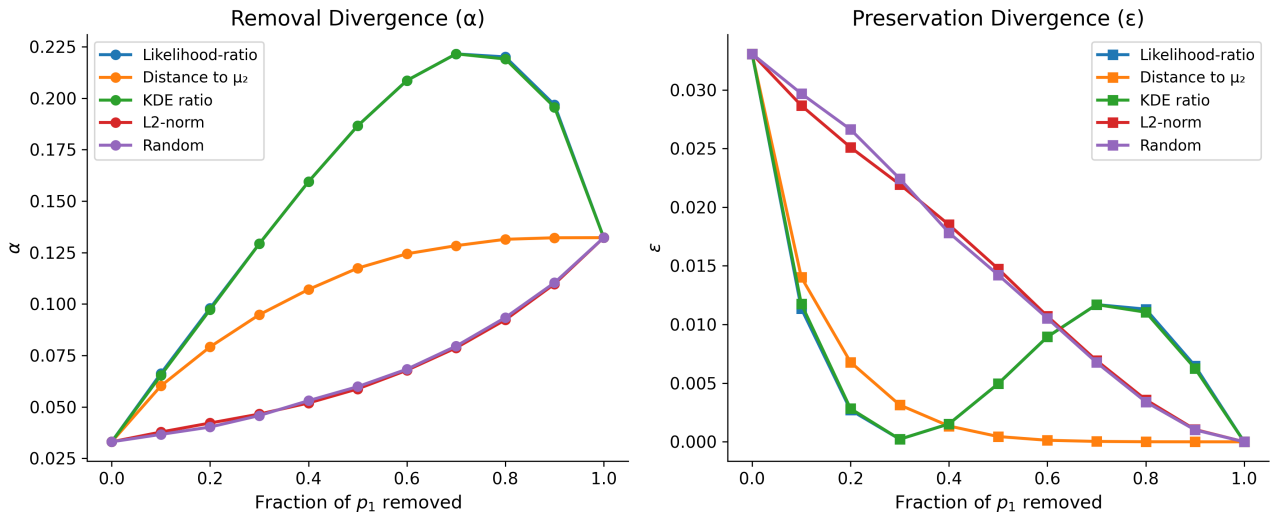


Figure 6. **Synthetic Gaussians.** Comparison of all strategies considered in the real-world datasets, using the low-divergence scenario of synthetic Gaussians (left plot, Fig. 2). Likelihood-ratio surpasses the distance-based strategy for lower deletion budgets. However, for larger budgets, it sacrifices utility for excessive removal. This is likely because it deletes samples representative of p_1 but close to p_2 since p_1 and p_2 are similar here.

E. Experimental Details

E.1. Heuristic Deletion Strategies

Across all datasets, we evaluate five scoring strategies for ranking samples in the forget distribution p_1 for removal. These scores approximate statistical dissimilarity from the retained distribution p_2 and correspond to different operational interpretations of divergence:

- **LR-COS / LR-MAHA** (likelihood-ratio inspired):
  A proxy for the log-likelihood ratio of x under p_2 versus p_1:
  $$s(x) = d(x, \mu_2) - d(x, \mu_1)$$
  where $d(\cdot, \mu)$ denotes cosine distance in TF–IDF space (text) or Mahalanobis distance in ResNet feature space (images). Points are scored high when they are far from p_2 and close to p_1.

- **COS-MU2 / MAHA-MU2** (dissimilarity to p_2):
  Measures the distance from each $x \in p_1$ to the empirical mean of p_2:
  $$s(x) = d(x, \mu_2)$$
  This approximates the contribution of each point to the KL divergence $\mathrm{KL}(p_2 \| \hat{p})$ when p is modeled as a Gaussian.

- **KNN-RATIO** (local density ratio):
  Estimates the ratio of k-NN densities:
  $$s(x) = \frac{\widehat{p_1}(x)}{\widehat{p_2}(x)}$$
  where $\widehat{p_i}(x) = \exp(-\|x - \mathrm{NN}_k^{(i)}(x)\|^2/\sigma^2)$ is a local Gaussian kernel density using $k = 10$ nearest neighbors. This captures how typical x is under p_1 versus p_2.

- **TFIDF-NORM / L2-NORM**:
  Uses the ℓ_2 norm of the raw input (TF–IDF or real-valued) as a proxy for informativeness or deviation from the origin:
  $$s(x) = \|x\|_2$$

- **RANDOM**:
  Samples points uniformly at random from p_1 as a baseline.

E.2. Synthetic Gaussians

We draw $n_1 = n_2 = 1{,}000$ samples from $p_1 = \mathcal{N}(0, 1)$ and $p_2 = \mathcal{N}(\mu_2, 1)$ for $\mu_2 \in \{0.5, 2.5, 5.0\}$, with 20 seeds. After computing scores using each strategy, we remove the top-f fraction of p_1 points, fit a Gaussian $\mathcal{N}(\hat{\mu}, 1)$ to the retained data, and compute:

$$\alpha = \mathrm{KL}(p_1 \| \hat{p}), \qquad \varepsilon = \mathrm{KL}(p_2 \| \hat{p})$$

These metrics match the forward-KL objectives of removal and preservation. No predictive model is trained; results reflect pure distributional divergence. For completeness, in Figure 6, we plot a comparison of all strategies considered in the real-world datasets, using the low-divergence scenario of synthetic Gaussians (left plot, Fig. 2). Likelihood-ratio surpasses the distance-based strategy for lower deletion budgets. However, for larger budgets, it sacrifices utility for excessive removal. This is likely because it deletes samples representative of p_1 but close to p_2 since p_1 and p_2 are similar here. This indicates that no removal strategy strictly dominates all others across all divergence (between p_1 and p_2) scenarios.

E.3. Jigsaw Toxic Comments

We use the Jigsaw Toxic Comment Classification dataset, with 140K examples filtered to length 5–200 tokens. We define p_1 as all training comments containing any of the keywords: "f*ck", "s*it", "d*mn", "b*tch", "a*s", and p_2 as the remaining comments. For each of 5 random seeds, we:

1. Stratified-split p_1, p_2 into 70/30 train/val.
2. Compute TF–IDF embeddings (40K max features, 1–2 grams, sublinear TF, min_df=5).
3. Score and remove f of p_1 training points using each heuristic.
4. Downsample p_2 to 5× the remaining p_1 size.
5. Train an ℓ_2-regularized logistic regression on the edited data.
6. Evaluate *Recall@*p_1 and macro *F1@*p_2 on the validation sets.

E.4. SMS Spam Collection

We use the UCI SMS Spam dataset (5574 examples, 13.4% spam). We apply:

1. TF–IDF vectorization (20K features, 1–2 grams, stopword removal).
2. Scoring of spam (p_1) messages using each heuristic.
3. Removal of top-f fraction of spam for each strategy.
4. Retrain a logistic regression classifier.
5. Evaluate *Recall@spam* and *F1@ham* on a held-out 20% test split.

We run 10 seeds and report mean $\pm$ standard error.

E.5. CIFAR-10 Class Removal

We treat the "cat" class as p_1 and the other 9 classes as p_2. We use the standard CIFAR-10 split (50K train, 10K test), and proceed as follows:

1. Train a 3-block CNN (Conv–BN–ReLU ×2 + MaxPool, widths 32–64–128, global avg pool + linear head) for 10 epochs on the full training set.
2. Extract features for all training images using the penultimate layer.
3. Compute Mahalanobis distance scores for each cat image (p_1) using:

$$s_{\text{maha}}(x) = \sqrt{(x - \mu)^\top \Sigma^{-1} (x - \mu)}$$

   where μ and Σ are estimated from p_2.

4. Delete the top-f fraction of cat images under each scoring method.
5. Retrain the same CNN architecture on the edited training set.
6. Evaluate:
$$\text{Accuracy}_{\text{cat}}, \qquad \text{Accuracy}_{\text{non-cat}}$$
on the test set. Results are averaged over 30 random seeds.

E.6. Computing Environment

All experiments were run on a HPE DL380 Gen10 equipped with two Intel(R) Xeon(R) Platinum 8358P CPUs running at 2.60GHz, 128 GB of RAM, a 740 GB SSD, and two NVIDIA A10 GPUs. Training for vision experiments was implemented in PyTorch, while text-based experiments used Scikit-learn. All experiments were conducted using a single GPU.