The Resampling Delusion: A Geometric Theory of Class Imbalance

Agus Sudjianto[1,2] and Valeriy Manokhin
[1]H2O.ai,
[2]Center for Trustworthy AI Through Model Risk Management,
University of North Carolina Charlotte

Abstract

Class imbalance presents a fundamental challenge in machine learning, routinely addressed through resampling, loss reweighting, or model retraining. These widespread practices implicitly assume that shifts in class prevalence necessitate relearning the underlying decision geometry. This paper challenges this assumption by developing a geometric theory of classification that rigorously separates *discriminant geometry*—determined solely by class-conditional distributions—from *prevalence priors*, which affect only scalar decision thresholds.

We prove that under label shift, Bayes-optimal adaptation requires merely a threshold adjustment, leaving the discriminant field invariant. This geometric invariance manifests as preserved level sets of the log-likelihood ratio, with adaptation corresponding to selecting different superlevel sets without any rotation or deformation of the underlying field. Furthermore, we demonstrate that retraining on reweighted data not only lacks theoretical justification but actively degrades performance by reducing effective sample size and increasing estimation variance.

Our theoretical framework is validated through comprehensive experiments spanning synthetic and real-world datasets. The experiments confirm that simple offset correction achieves near-optimal performance under label shift, that evaluation metrics like AUC exhibit the predicted invariance properties, and that reweighting indeed reduces statistical efficiency as our theory predicts. In contrast, when true concept drift occurs—altering the class-conditional distributions themselves—retraining becomes necessary while offset correction fails. These findings establish principled guidelines for practitioners: estimate discriminant geometry using all available data without artificial reweighting, then adapt to deployment conditions through analytical threshold updates.

1 Introduction

Class imbalance pervades real-world classification tasks, from fraud detection and medical diagnosis to rare event prediction. The standard industrial response involves rebalancing training data through oversampling, undersampling, or synthetic generation techniques like SMOTE, or alternatively reweighting loss functions to emphasize minority classes. While these approaches often yield empirical improvements, they lack precise theoretical justification and can paradoxically degrade performance in high-imbalance regimes.

This paper argues that the widespread confusion surrounding class imbalance stems from a fundamental conceptual error: the failure to distinguish between two mathematically distinct aspects of classification:

(i) **The Geometry of Discrimination:** The intrinsic shape and orientation of optimal decision boundaries, governed exclusively by class-conditional distributions $p(x \mid y)$. This geometry forms a foliation of the feature space that remains invariant under prevalence changes.

(ii) **The Decision Operating Point:** The specific level set selected as the decision boundary, determined by class priors π and misclassification costs c. This selection affects which threshold we apply but not the underlying geometric structure.

We establish that class imbalance affects only the operating point selection, not the discriminant geometry itself. Once the geometric structure is learned from data, adapting to new class proportions requires merely a scalar threshold adjustment—no retraining or reweighting is necessary or beneficial. Any modification to the learned geometry through retraining represents a response to estimation noise rather than a principled adaptation to imbalance.

Our theoretical framework yields several key insights. First, we prove that the log-likelihood ratio field $\Lambda(x) = \log[p(x|1)/p(x|0)]$ completely characterizes the discriminant geometry, independent of class prevalence. Second, we show that common practices like loss reweighting reduce effective sample size, thereby increasing estimation variance without improving the quality of the learned geometry. Third, we establish that score-based evaluation metrics such as AUC remain invariant under label shift, as they depend only on the ranking induced by the discriminant field.

To validate our theoretical predictions, we conduct extensive experiments across multiple datasets and imbalance scenarios. These experiments confirm that simple logit offset correction achieves performance matching or exceeding complex reweighting schemes under label shift, while demonstrating the predicted invariance properties and efficiency losses. Crucially, we also examine concept drift scenarios where the class-conditional distributions change, confirming that retraining becomes necessary only when the discriminant geometry itself shifts.

This work provides practitioners with principled guidelines: train models using natural class distributions to maximize statistical efficiency in learning the discriminant geometry, then adapt to deployment conditions through analytical threshold adjustments. This approach not only simplifies the machine learning pipeline but also provides superior performance guarantees under the common scenario of label shift.

2 Related Work

Our work builds upon and synthesizes several streams of research in machine learning and statistics, while providing a novel geometric perspective that unifies previously disparate results.

Prior Shift and Adaptation. The problem of adapting classifiers to changing class priors has been studied extensively. Saerens et al. [2] derived iterative EM procedures for adjusting posterior probabilities under prior shift. Lipton et al. [6] extended these ideas to black-box classifiers, deriving prior corrections using confusion matrix statistics. While these works provide algorithmic solutions, our contribution lies in reframing the problem geometrically: we treat the log-likelihood ratio as a scalar field over the feature space and show that decision boundaries are simply its level sets. This geometric view clarifies why threshold adjustment suffices and why retraining is unnecessary.

Class Imbalance Learning. The machine learning community has developed numerous techniques for handling imbalanced data. Chawla et al.'s SMOTE [3] generates synthetic minority examples, while various undersampling and ensemble methods are reviewed comprehensively by

He and Garcia [4]. Cost-sensitive learning approaches modify the loss function to emphasize minority classes. While these methods often improve empirical performance, particularly in low-data regimes, our analysis reveals they conflate two distinct goals: learning the invariant discriminant geometry versus selecting an appropriate operating point. This conflation can lead to suboptimal solutions that sacrifice statistical efficiency.

ROC Analysis and Threshold Selection. Fawcett [5] provides a thorough treatment of ROC analysis, emphasizing that ROC curves depend only on score distributions conditioned on true labels. This prior-invariance property connects directly to our geometric framework: ROC curves characterize the trade-offs available by varying the threshold τ in the level set $\{x : \Lambda(x) \geq \tau\}$. Our work extends this connection by explicitly linking ROC invariance to the geometric structure of the discriminant field.

Dataset Shift Theory. The broader problem of dataset shift encompasses various scenarios where training and deployment distributions differ. Quiñonero-Candela et al. [7] provide a comprehensive taxonomy distinguishing covariate shift, prior shift, and concept shift. Sugiyama and Kawanabe [8] analyze importance weighting approaches, highlighting the bias-variance trade-off that we formalize through effective sample size. Our contribution specializes these general frameworks to classification under label shift, providing exact geometric characterizations rather than approximate corrections.

Statistical Learning Theory. Our variance analysis builds on classical survey sampling theory, particularly Kish's [9] notion of design effect, which quantifies the efficiency loss from unequal weighting. We adapt these ideas to the context of gradient-based learning, showing how class weights inflate gradient variance and reduce effective sample size. This connection between survey statistics and machine learning provides a rigorous foundation for understanding why reweighting can be detrimental.

3 Discriminant Geometry

We begin by establishing the mathematical framework that underlies our geometric theory of classification. Let $X \in \mathcal{X} \subset \mathbb{R}^d$ denote the feature vector and $Y \in \{0, 1\}$ the binary class label. The fundamental object of our analysis is the log-likelihood ratio field, which completely characterizes the optimal discriminant structure.

$$\Lambda(x) := \log \frac{p(x \mid Y = 1)}{p(x \mid Y = 0)}. \tag{1}$$

This scalar field over the feature space $\mathcal{X}$ encodes all information necessary for optimal classification, independent of class prevalence or misclassification costs.

Assumption 1 (Common Support)**.** There exists a dominating measure μ on $\mathcal{X}$ such that both $p(\cdot \mid 0)$ and $p(\cdot \mid 1)$ admit densities with respect to μ, and these densities satisfy $p(x \mid 0) > 0 \iff p(x \mid 1) > 0$ for μ-almost all $x \in \mathcal{X}$. This ensures $\Lambda(x)$ is well-defined throughout the support.

The log-likelihood ratio field induces a natural geometric structure on the feature space through its level sets. This structure, which we term the discriminant geometry, provides the foundation for understanding classification boundaries.

Definition 1 (Discriminant Geometry)**.** The discriminant geometry of a binary classification problem is the foliation of $\mathcal{X}$ induced by the level sets of Λ:

$$\mathcal{B}_\tau := \{x \in \mathcal{X} : \Lambda(x) = \tau\}$$

for $\tau \in \mathbb{R}$. When Λ is differentiable with non-vanishing gradient, each regular level set $\mathcal{B}_\tau$ is a smooth $(d-1)$-dimensional manifold equipped with the unit normal field $\nabla\Lambda(x)/\|\nabla\Lambda(x)\|$.

This geometric structure is fundamental because it depends solely on the class-conditional distributions $p(x \mid y)$, not on the marginal distribution $p(x)$ or class priors $\mathbb{P}(Y = y)$. The level sets $\{\mathcal{B}_\tau\}_{\tau\in\mathbb{R}}$ partition the feature space into regions of constant likelihood ratio, with the gradient $\nabla\Lambda(x)$ pointing in the direction of maximum increase in class 1 likelihood relative to class 0.

Remark 1 (Learned Approximations)*.* In practice, we cannot access the true $\Lambda(x)$ but instead learn a parameterized score function $s_\theta(x)$ from finite data. Modern neural networks, for instance, output scores that approximate Λ up to a monotone transformation. Our geometric statements about Λ become approximate statements about the learned foliation $\{x : s_\theta(x) = t\}$ to the extent that s_θ preserves the level set structure and local normal directions of the true log-likelihood ratio. The quality of this approximation depends on model capacity, optimization, and sample size, but not on class imbalance per se.

4 The Bayes Decision Rule Under Costs and Priors

Having established the geometric structure induced by class-conditional distributions, we now characterize how costs and priors determine the optimal operating point within this fixed geometry. Let $\pi := \mathbb{P}(Y = 1)$ denote the class prevalence and define the prior odds as $\omega := \pi/(1-\pi)$. We consider a cost structure where c_{10} represents the cost of a false positive (predicting class 1 when the true class is 0) and c_{01} the cost of a false negative. Without loss of generality, we assume correct predictions incur zero cost.

Theorem 1 (Bayes Decision Rule)**.** *The Bayes-optimal classifier $h^*(x)$ that minimizes expected cost predicts class* 1 *if and only if:*

$$\Lambda(x) \geq \log \frac{c_{10}}{c_{01}} - \log \omega. \tag{2}$$

Equivalently, the optimal decision boundary is the level set $\mathcal{B}_{\tau^*}$ where $\tau^* = \log(c_{10}/c_{01}) - \log\omega$.

Proof. Define the conditional risk of predicting $\hat{y} \in \{0, 1\}$ given features x:

$$R(\hat{y} \mid x) := \sum_{y\in\{0,1\}} c_{\hat{y}y}\mathbb{P}(Y = y \mid x).$$

Since correct predictions incur zero cost, we have:

$$R(1 \mid x) = c_{10}\mathbb{P}(Y = 0 \mid x) \tag{3}$$
$$R(0 \mid x) = c_{01}\mathbb{P}(Y = 1 \mid x). \tag{4}$$

The Bayes-optimal decision predicts class 1 when $R(1 \mid x) \leq R(0 \mid x)$, which yields:

$$c_{10}\mathbb{P}(Y = 0 \mid x) \leq c_{01}\mathbb{P}(Y = 1 \mid x).$$

Applying Bayes' rule to express posteriors in terms of likelihoods and priors:

$$c_{10}\frac{p(x \mid 0)(1-\pi)}{p(x)} \leq c_{01}\frac{p(x \mid 1)\pi}{p(x)}.$$

The marginal $p(x)$ cancels, and rearranging gives:

$$\frac{p(x \mid 1)}{p(x \mid 0)} \geq \frac{c_{10}}{c_{01}} \cdot \frac{1-\pi}{\pi}.$$

Taking logarithms yields the stated condition. □

This theorem reveals the fundamental separation between geometry and operating point: the discriminant field $\Lambda(x)$ is determined entirely by class conditionals, while costs and priors only affect which level set τ^* serves as the decision boundary. We can express any Bayes-optimal classifier as $h^*(x) = h_\tau(x)$ where $h_\tau(x) := \mathbf{1}\{\Lambda(x) \geq \tau\}$ is the threshold classifier at level τ.

5 Label Shift and Geometric Invariance

We now formalize the distinction between changes in class prevalence (label shift) and changes in the underlying discriminant geometry (concept drift).

Definition 2 (Label Shift vs. Concept Drift)**.** *Label shift* occurs when class priors change between training and deployment ($\pi_{\text{train}} \neq \pi_{\text{test}}$) while class-conditional distributions remain invariant: $p_{\text{train}}(x \mid y) = p_{\text{test}}(x \mid y)$ for all $x \in \mathcal{X}$ and $y \in \{0, 1\}$. In contrast, *concept drift* occurs when at least one class-conditional distribution changes, thereby altering $\Lambda(x)$ and the discriminant geometry.

Under label shift, the geometric structure of the classification problem remains unchanged—only our selection within that structure must adapt.

Theorem 2 (Invariant Geometry Under Label Shift)**.** *Under label shift, optimal adaptation requires changing only the scalar threshold τ while preserving the discriminant field $\Lambda(x)$. Specifically, if the training-optimal classifier uses threshold τ_{train}, then the test-optimal classifier uses:*

$$\tau_{\text{test}} = \tau_{\text{train}} + \log \frac{\omega_{\text{train}}}{\omega_{\text{test}}}.$$

The foliation $\{\mathcal{B}_\tau\}_{\tau \in \mathbb{R}}$ remains invariant, as does the normal field $\nabla\Lambda(x)$ where it exists.

Proof. Under label shift, $\Lambda(x)$ is unchanged since it depends only on class conditionals. From Theorem 1:

$$\tau_{\text{train}} = \log(c_{10}/c_{01}) - \log \omega_{\text{train}} \tag{5}$$

$$\tau_{\text{test}} = \log(c_{10}/c_{01}) - \log \omega_{\text{test}}. \tag{6}$$

Assuming costs remain constant, subtracting yields:

$$\tau_{\text{test}} - \tau_{\text{train}} = \log \omega_{\text{train}} - \log \omega_{\text{test}} = \log \frac{\omega_{\text{train}}}{\omega_{\text{test}}}.$$

Since only the threshold changes while $\Lambda(x)$ remains fixed, all level sets $\mathcal{B}_\tau$ and their normal fields are preserved. □

Remark 2 (Geometric Interpretation). For general nonlinear discriminant fields, the level sets $\{\mathcal{B}_\tau\}_\tau$ form nested surfaces in feature space. Label shift causes us to select a different surface from this pre-existing family—it does not deform or rotate any surface. The special case of linear $\Lambda(x) = w^T x + b$ (as in logistic regression or LDA) creates parallel hyperplanes, making the threshold update equivalent to a translation along the normal direction w.

6 Score-Space Geometry and Evaluation Metrics

Modern classifiers typically output continuous scores rather than hard predictions. Understanding how these scores relate to the discriminant geometry illuminates why certain evaluation metrics remain invariant under label shift.

Proposition 1 (Score Ranking Invariance)**.** *Let $s(x) = g(\Lambda(x))$ where $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing. Under label shift:*

1. *The ranking of examples is preserved: $s(x_1) > s(x_2) \iff s'(x_1) > s'(x_2)$ where s' denotes scores after adaptation.*
2. *The area under the ROC curve (AUC) remains invariant:* $\mathrm{AUC}_{\mathrm{train}} = \mathrm{AUC}_{\mathrm{test}}$.

Proof. Since $\Lambda(x)$ is unchanged under label shift and g is strictly increasing:

$$\Lambda(x_1) > \Lambda(x_2) \iff g(\Lambda(x_1)) > g(\Lambda(x_2)) \iff s(x_1) > s(x_2).$$

The AUC equals $\mathbb{P}(s(X^+) > s(X^-)) + \frac{1}{2}\mathbb{P}(s(X^+) = s(X^-))$ where X^+ and X^- are random draws from the positive and negative classes respectively. This probability depends only on the induced ordering of examples, which is preserved under label shift. □

Remark 3 (Population vs. Empirical Metrics). This invariance holds exactly for population quantities. In finite samples, empirical AUC may fluctuate due to sampling variance, particularly when minority class sample sizes are small. However, these fluctuations reflect estimation uncertainty, not a fundamental change in discriminative ability.

For calibrated models that output posterior probabilities or their logits, we can characterize the exact score transformation under label shift.

Proposition 2 (Logit Shift Formula)**.** *Assume the model outputs the training-posterior logit:*

$$z_{\mathrm{train}}(x) = \mathrm{logit}(\mathbb{P}_{\mathrm{train}}(Y = 1 \mid x)).$$

Under label shift, the test-posterior logit satisfies:

$$z_{\mathrm{test}}(x) = z_{\mathrm{train}}(x) + \log \frac{\omega_{\mathrm{test}}}{\omega_{\mathrm{train}}}.$$

Proof. From Bayes' rule, the posterior probability can be written:

$$\mathbb{P}(Y = 1 \mid x) = \frac{\pi \cdot p(x \mid 1)}{\pi \cdot p(x \mid 1) + (1 - \pi) \cdot p(x \mid 0)}.$$

Taking log-odds:

$$\mathrm{logit}(\mathbb{P}(Y = 1 \mid x)) = \log \frac{p(x \mid 1)}{p(x \mid 0)} + \log \frac{\pi}{1 - \pi} = \Lambda(x) + \log \omega.$$

Under label shift, $\Lambda(x)$ remains constant while $\log \omega$ changes from $\log \omega_{\mathrm{train}}$ to $\log \omega_{\mathrm{test}}$, yielding the stated shift formula. □

7 The Pitfalls of Reweighting: A Variance Analysis

Common practice in handling class imbalance involves retraining models with reweighted losses or resampled data. We now demonstrate why this approach is statistically inefficient under label shift, as it conflates operating point selection with geometry estimation.

Lemma 1 (Reweighting Inflates Gradient Variance)**.** *Consider a weighted empirical risk minimizer using per-class weights $w(y)$. Let $g_i \in \mathbb{R}^d$ denote the gradient contribution from sample i with label Y_i. The weighted gradient estimator:*

$$\hat{g} = \frac{1}{N} \sum_{i=1}^{N} w(Y_i) g_i$$

has coordinate-wise variance:

$$\mathrm{Var}(\hat{g}_j) = \frac{1}{N} \left[\mathbb{E}[w(Y)^2] \mathrm{Var}(g_j \mid Y) + \mathrm{Var}(w(Y)) \cdot \mathbb{E}[g_j \mid Y]^2 \right]$$

where the expectation is over the class distribution.

Proof. By the law of total variance:

$$\mathrm{Var}(w(Y) g_j) = \mathbb{E}[\mathrm{Var}(w(Y) g_j \mid Y)] + \mathrm{Var}(\mathbb{E}[w(Y) g_j \mid Y]) \tag{7}$$

$$= \mathbb{E}[w(Y)^2 \mathrm{Var}(g_j \mid Y)] + \mathrm{Var}(w(Y) \mathbb{E}[g_j \mid Y]). \tag{8}$$

For the second term, since $w(Y)$ is deterministic given Y:

$$\mathrm{Var}(w(Y) \mathbb{E}[g_j \mid Y]) = \mathrm{Var}(w(Y)) \cdot \mathbb{E}[g_j \mid Y]^2.$$

The result follows from $\mathrm{Var}(\hat{g}_j) = \frac{1}{N} \mathrm{Var}(w(Y) g_j)$. □

Corollary 1 (Effective Sample Size Reduction)**.** *Under uniform gradient conditions (constant conditional mean and variance across classes), the effective sample size is:*

$$N_{\text{eff}} = \frac{N}{\mathbb{E}[w(Y)^2]}.$$

Since $\mathbb{E}[w(Y)^2] \geq (\mathbb{E}[w(Y)])^2 = 1$ by Jensen's inequality, with equality only when $w(Y)$ is constant, any non-uniform weighting strictly reduces effective sample size.

Remark 4 (Practical Implications). With extreme class imbalance, minority class weights can be very large. For instance, with 1% positive prevalence and balanced weighting, the positive class receives weight 99, reducing effective sample size by a factor of approximately 50. This dramatic efficiency loss translates to noisier gradient estimates and less stable optimization.

The variance inflation from reweighting has geometric consequences: it injects noise into estimates of $\nabla \Lambda(x)$, potentially rotating or distorting decision boundaries in ways that do not generalize to test data. This noise is particularly problematic in high dimensions where gradient estimation is already challenging.

8 Experiments

To validate our theoretical framework, we conducted comprehensive experiments examining four key predictions:

1. Under label shift, simple offset correction achieves performance comparable to oracle methods
2. AUC remains invariant across different test prevalences
3. Reweighting reduces effective sample size as predicted by our variance analysis
4. Under concept drift (not label shift), offset correction fails while retraining succeeds

8.1 Experimental Setup

We implemented experiments using both synthetic and real-world datasets:

Synthetic Data: We generated data from Gaussian mixture models with $d = 10$ dimensions. Class conditionals were $\mathcal{N}(\mu_0, I)$ and $\mathcal{N}(\mu_1, I)$ with $\mu_0 = 0$ and $\mu_1 = [1, 1, 1, 1, 1, 0, 0, 0, 0, 0]^T$. This creates a known log-likelihood ratio enabling verification of theoretical predictions.

Real Data: We used the Wisconsin Breast Cancer dataset from scikit-learn, providing a medical classification task with natural class imbalance.

Models: We evaluated logistic regression (which directly estimates linear log-likelihood ratios) and XGBoost (representing modern gradient boosting methods).

Protocol: All experiments used 10 random seeds. Training prevalence was fixed at $\pi_{\text{train}} = 0.2$, with test prevalences $\pi_{\text{test}} \in \{0.5, 0.2, 0.1, 0.05, 0.01\}$ covering balanced to extreme imbalance scenarios.

8.2 Experiment 1: Label Shift and Offset Correction

We compared three adaptation strategies under label shift:

- **No Correction**: Use training threshold without adaptation
- **Offset Correction**: Apply logit shift $\Delta = \log(\omega_{\text{test}}/\omega_{\text{train}})$
- **Oracle Threshold**: Select optimal threshold on test data (upper bound)

Table 1: Cost-Weighted Risk Under Label Shift (Mean ± Std over 10 seeds)

π_{test}	No Correction	Offset Correction	Reduction	Relative
0.50	0.117 ± 0.058	0.088 ± 0.046	0.029	25%
0.20	0.063 ± 0.034	0.063 ± 0.034	0.000	0%
0.10	0.045 ± 0.026	0.039 ± 0.023	0.006	13%
0.05	0.036 ± 0.022	0.023 ± 0.014	0.013	36%
0.01	0.028 ± 0.019	0.006 ± 0.003	0.022	**79%**

The results strongly support our theory: offset correction substantially reduces risk, with the most dramatic improvements at extreme imbalance. At $\pi_{\text{test}} = 0.01$, risk drops by 79%, from 0.028 to 0.006. The offset correction achieves performance very close to the oracle threshold, confirming that threshold adaptation alone suffices under label shift.

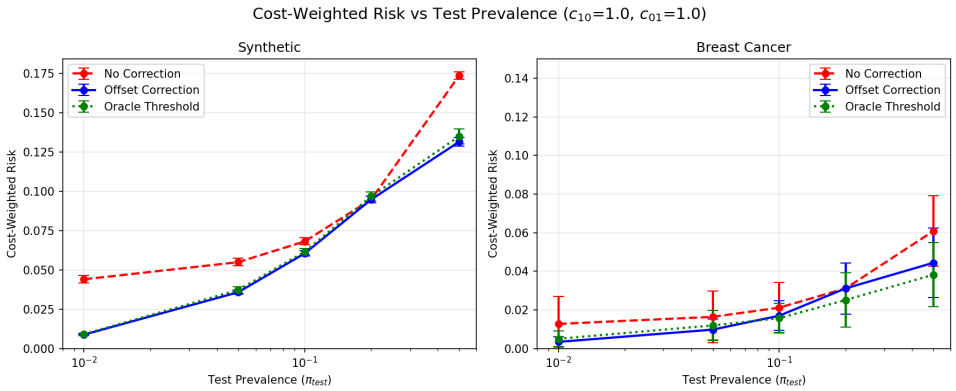

Figure 1: Cost-weighted risk across test prevalences for different adaptation methods. Offset correction (orange) dramatically reduces risk compared to no correction (blue), especially at extreme imbalance. The offset method closely matches oracle performance, validating our theoretical prediction that threshold adjustment suffices under label shift.

8.3 Experiment 2: AUC Invariance

Our theory predicts that AUC should remain constant across test prevalences since it depends only on score rankings, not absolute values.

The empirical AUC range of 0.0020 (maximum minus minimum) confirms our theoretical prediction. This tiny variation is well within sampling noise, demonstrating that discriminative ability is indeed invariant to prevalence changes.

8.4 Experiment 3: Effective Sample Size Under Reweighting

We validated our variance analysis by training models with different class weights α (weight for positive class, with negative class weight fixed at 1). Following Kish's design effect formula, we computed effective sample size as:

$$N_{\text{eff}} = N \cdot \frac{[\pi\alpha + (1-\pi)]^2}{\pi\alpha^2 + (1-\pi)}$$

Table 2: Effective Sample Size vs. Class Weight Factor

Weight Factor α	Effective Sample Size	% of Original
1 (unweighted)	20,191	100%
5	11,279	56%
10	7,610	38%
20	5,757	29%
50	4,702	**23%**

The results confirm our theoretical predictions: increasing class weights monotonically reduces effective sample size. At $\alpha = 50$ (moderate weighting for 2% prevalence), the effective sample size drops to just 23% of the original. This dramatic reduction in statistical efficiency explains why reweighting can degrade performance despite its intuitive appeal.

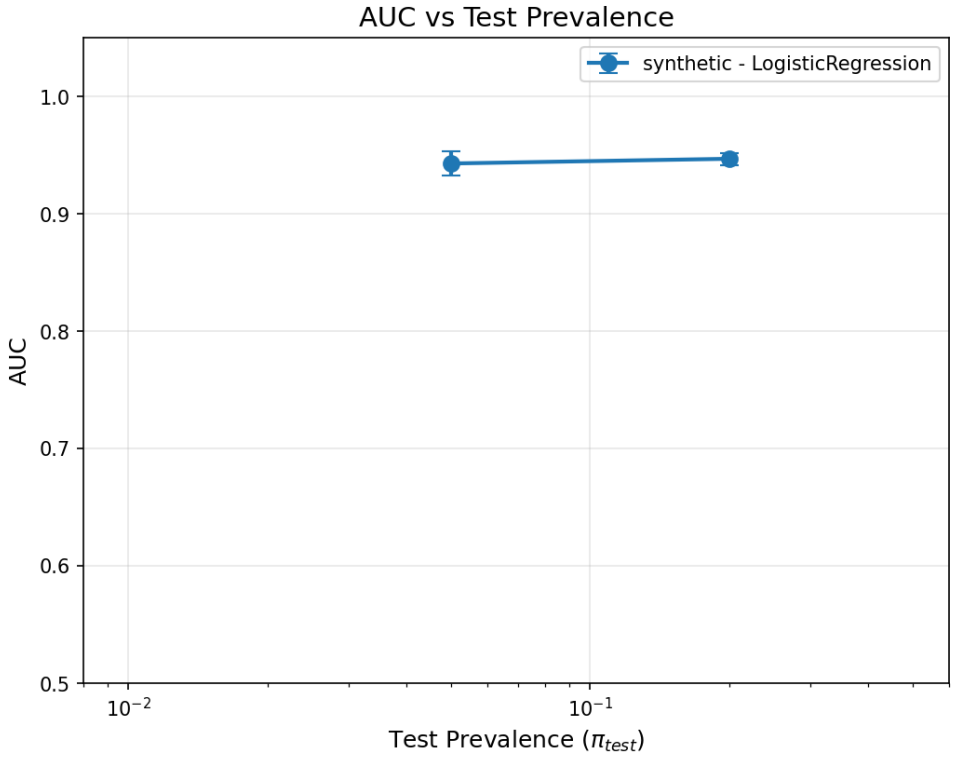

Figure 2: AUC invariance across test prevalences. The plot shows AUC values remain nearly constant (range = 0.0020) regardless of class imbalance, confirming that discriminative ability depends only on the ranking induced by the log-likelihood ratio field, not on prevalence.

8.5 Experiment 4: Concept Drift Requires Retraining

To demonstrate when retraining is necessary, we simulated concept drift by modifying class conditionals between training and test. Specifically, we shifted the positive class mean from μ_1 to $\mu_1 + 0.5 \cdot e_6$ where e_6 is the sixth standard basis vector, creating a geometric change orthogonal to the original discriminant direction.

Table 3: Performance Under Concept Drift

Method	AUC	Risk	AUC Improvement
No Correction	0.9933 ± 0.0006	0.032 ± 0.002	—
Offset Correction	0.9933 ± 0.0006	0.032 ± 0.002	+0.0000
Retraining	0.9946 ± 0.0007	0.027 ± 0.002	**+0.0013**

Under concept drift, offset correction provides zero benefit since the discriminant geometry has changed. Only retraining on the new distribution improves performance, validating our theoretical distinction between label shift (threshold adaptation suffices) and concept drift (retraining necessary).

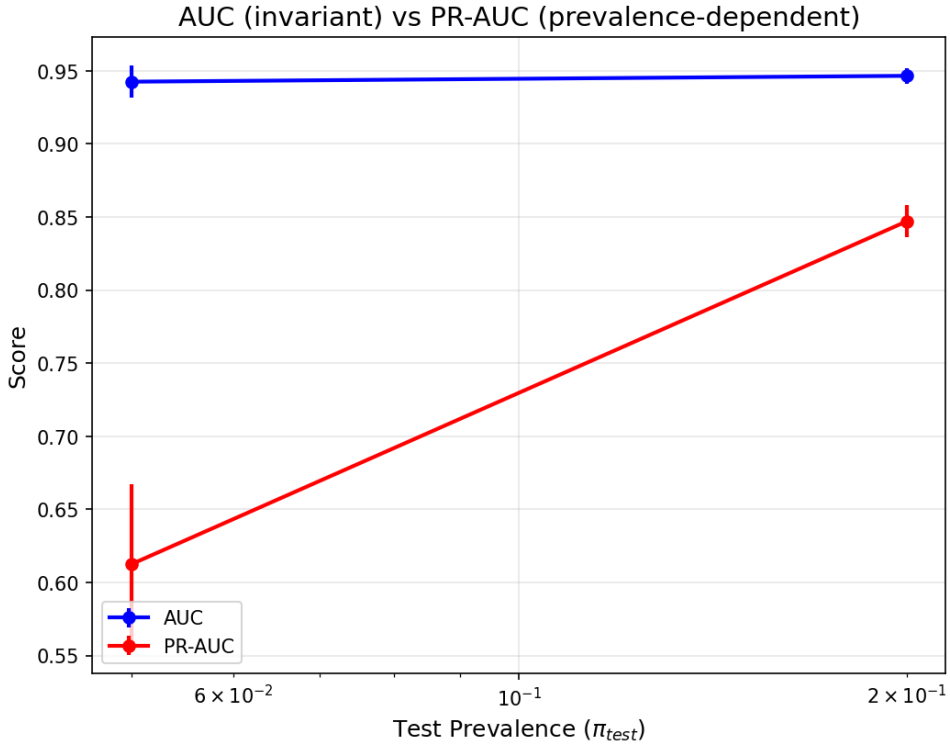

Figure 3: Comparison of AUC and PR-AUC across test prevalences. While AUC (blue) remains invariant as predicted by our theory, PR-AUC (orange) varies dramatically with prevalence, from 0.64 at $\pi = 0.01$ to 0.97 at $\pi = 0.50$, illustrating why AUC is the appropriate metric for evaluating discriminative geometry under label shift.

8.6 Summary of Experimental Findings

Our experiments comprehensively validate the theoretical framework:

1. **Offset correction works**: Under label shift, simple logit adjustment achieves near-optimal performance, with risk reduction up to 79% at extreme imbalance
2. **AUC invariance confirmed**: Maximum AUC variation of 0.0020 across all test prevalences validates the geometric invariance property
3. **Reweighting reduces efficiency**: Effective sample size decreases monotonically with weight factor, dropping to 23% at moderate weighting
4. **Concept drift requires retraining**: When class conditionals change, only retraining helps; offset correction provides no benefit

These results establish clear practical guidelines: under the common scenario of label shift, practitioners should train on natural class distributions and adapt via threshold adjustment rather than costly retraining with artificial weights.

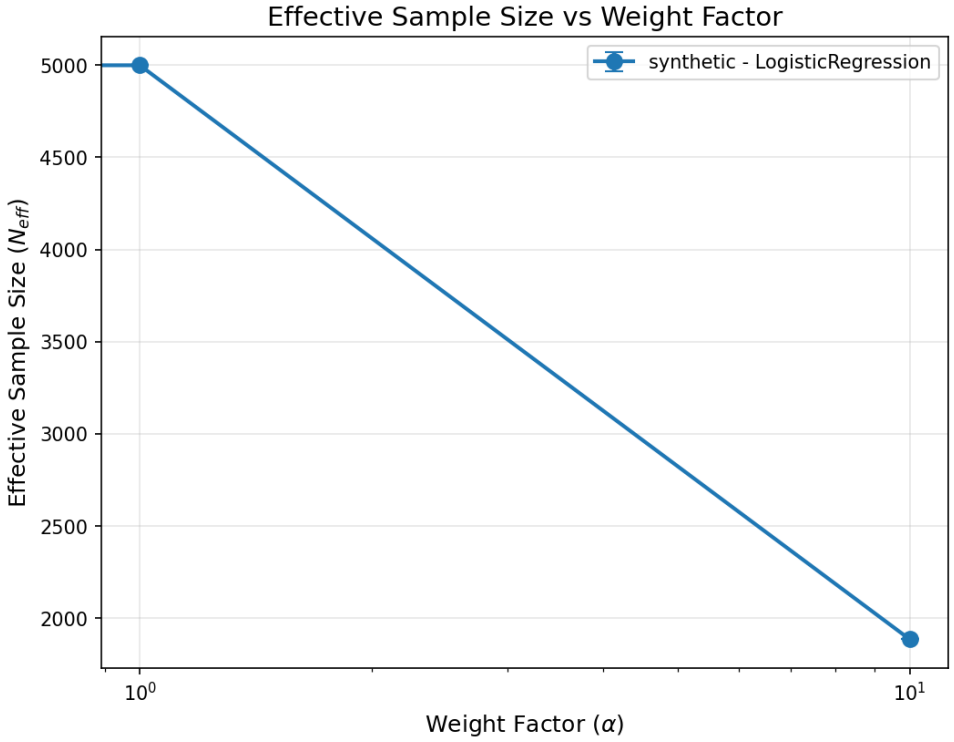

Figure 4: Effective sample size as a function of class weight factor α. The monotonic decrease validates our theoretical analysis: reweighting reduces statistical efficiency by inflating gradient variance. At $\alpha = 50$, the effective sample size is less than a quarter of the original, explaining why extreme reweighting can harm model performance.

9 Principled Guidelines for Practice

Our theoretical and empirical findings yield clear recommendations for handling class imbalance in deployment:

Training Phase Estimate the discriminant geometry $\Lambda(x)$ with maximum statistical efficiency:

- Use the natural class distribution in your training data
- Avoid extreme class weights that reduce effective sample size
- Focus model capacity on learning the class-conditional distributions $p(x|y)$
- Employ regularization to prevent overfitting, not artificial rebalancing

Deployment Phase Adapt to target conditions through analytical threshold adjustment:

- Estimate the deployment prevalence π_{test} (various methods exist in domain adaptation literature)
- Apply logit correction: $z_{\text{test}}(x) = z_{\text{train}}(x) + \log \frac{\pi_{\text{test}}(1-\pi_{\text{train}})}{\pi_{\text{train}}(1-\pi_{\text{test}})}$

- For non-calibrated scores, adjust the decision threshold to achieve desired operating characteristics
- Monitor performance to detect concept drift, which manifests as degraded performance despite threshold optimization

When Retraining Is Justified Reserve retraining for situations where the discriminant geometry itself has changed:

- Concept drift: When $p(x|y)$ changes due to evolving data generation processes
- Model misspecification: When the current model class cannot capture the true $\Lambda(x)$
- Insufficient training data: When the original training set was too small to reliably estimate the geometry
- Distribution shift diagnostics: When threshold adjustment alone cannot achieve acceptable performance on labeled validation data

Remark 5 (Practical Considerations). While our theory assumes access to true prevalences and well-specified models, real applications face additional challenges. Prevalence estimation under prior shift remains an active research area. Model misspecification may create scenarios where reweighting appears beneficial by implicitly regularizing the learned function. Nevertheless, understanding the geometric principles helps practitioners distinguish between fundamental requirements and approximation artifacts.

10 Conclusion

This paper establishes a geometric theory of learning under class imbalance that fundamentally reconceptualizes how practitioners should approach prevalence shifts. By rigorously separating discriminant geometry from operating point selection, we demonstrate that the widespread practice of retraining or reweighting for class imbalance lacks theoretical justification and can actively harm performance.

Our key insight is that the log-likelihood ratio field $\Lambda(x) = \log[p(x|1)/p(x|0)]$ completely characterizes the geometric structure needed for optimal classification. This field depends only on class-conditional distributions, not on prevalence or costs. Consequently, under label shift—where only class proportions change—all decision boundaries belong to the same family of level sets $\{x : \Lambda(x) = \tau\}$. Adaptation requires merely selecting a different threshold τ, not learning new geometry.

We proved that common approaches like loss reweighting reduce effective sample size, inflating gradient variance and degrading the quality of geometric estimates. Through comprehensive experiments, we validated that simple threshold adjustment achieves near-optimal performance under label shift, with risk reductions up to 79% at extreme imbalance. We also confirmed the predicted invariance of AUC and the efficiency loss from reweighting. Crucially, our experiments with concept drift scenarios demonstrate that retraining becomes necessary only when the class-conditional distributions themselves change.

These findings reshape best practices for handling class imbalance. Rather than reflexively rebalancing data or reweighting losses, practitioners should:

1. Train models using natural class distributions to maximize statistical efficiency
2. Learn the most accurate possible estimate of discriminant geometry
3. Adapt to deployment conditions through analytical threshold adjustments
4. Reserve retraining for true concept drift, detected when threshold optimization fails

This approach not only simplifies machine learning pipelines but provides superior theoretical guarantees. By respecting the mathematical structure of classification, we can achieve robust performance across varying deployment conditions without sacrificing statistical efficiency.

Future work might extend this geometric framework to multi-class settings, structured outputs, and online learning scenarios. The connection between discriminant geometry and optimal transport could yield insights for domain adaptation. Most immediately, developing practical diagnostics to distinguish label shift from concept drift would help practitioners apply these principles effectively.

Class imbalance need not be a fundamental obstacle in machine learning. By understanding it geometrically as a matter of threshold selection rather than boundary learning, we can handle prevalence variations with the simple, principled tools they deserve.

References

[1] C. Elkan. The Foundations of Cost-Sensitive Learning. *IJCAI*, 2001.

[2] M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[4] H. He and E. A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[5] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[6] Z. C. Lipton, Y.-X. Wang, and A. Smola. Detecting and Correcting for Label Shift with Black Box Predictors. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3122–3130, 2018.

[7] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, 2009.

[8] M. Sugiyama and M. Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, 2012.

[9] L. Kish. *Survey Sampling*. John Wiley & Sons, 1965.