Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpolation-Based Oversampling beyond Class Reweighting: An Exact Endpoint-Affine Selection-Curvature Decomposition

Shengyu Chen

chenshengyu@tjcu.edu.cn

Tianjin University of Commerce

Shiyou Sui

Tianjin University of Commerce

Research Article

Keywords: imbalanced classification, interpolation-based oversampling, SMOTE, matched class weighting, selection-curvature decomposition, method screening

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10773202/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Interpolation-Based Oversampling beyond Class Reweighting: An Exact Endpoint-Affine Selection-Curvature Decomposition

Shengyu Chen · Shiyou Sui
Tianjin University of Commerce, Tianjin, China
Corresponding author: Shengyu Chen, chenshengyu@tjcu.edu.cn

Abstract

Interpolation-based oversampling changes both the total minority-class loss mass and the locations at which the loss is evaluated, so comparisons with an unweighted fit confound class reweighting with interpolation geometry. We construct a class-weighted control with exactly matched minority loss mass and objective normalization. Relative to this control, the residual gradient perturbation admits an exact endpoint-affine decomposition into an endpoint-selection term and an interpolation-curvature term. The decomposition yields explicit signatures for random oversampling, global interpolation, SMOTE, Borderline-SMOTE, and risk-weighted interpolation; for SMOTE, the selection term is governed by directed nearest-neighbour indegree imbalance. For smooth regularized empirical-risk minimization, the resulting signature determines the local optimizer response through the matched-control Hessian and yields a first-order, target-conditional validation score. Across 1,080 controlled comparisons on 12 imbalanced datasets using regularized logistic and probit regression, the full response attains a pooled median parameter-direction cosine of 0.997 and a median relative error of 0.119; learner-specific median relative errors are 0.115 and 0.122, respectively, while single-component ablations are materially less accurate under both learners. Across 216 learner-specific five-method held-out validation problems, the score recovers the oracle-best method in 87.5% of log-loss problems and 84.7% of Brier-score problems, with zero median selection regret. The framework isolates what interpolation contributes beyond matched class weighting and provides a mechanism-based approximation to candidate validation rankings without fitting every complete candidate model.

Keywords: imbalanced classification; interpolation-based oversampling; SMOTE; matched class weighting; selection-curvature decomposition; method screening

1. Introduction

Interpolation-based oversampling is usually evaluated by comparing a model fitted on an augmented dataset with one fitted on the original imbalanced sample. This comparison conflates two effects: adding minority observations increases total minority loss mass, whereas interpolation relocates that mass to synthetic feature locations. Class weighting can reproduce the first effect but not, in general, the second. Comparisons with the unweighted fit therefore do not isolate the geometric contribution of interpolation.

SMOTE and its variants remain widely used tools for imbalanced classification [1, 2]. Existing analyses study the distributional behaviour of synthetic samples [3-5], as well as concentration and excess-risk properties of synthetic oversampling [6]. Related work has also analyzed interpolation through regularization and directional derivatives [7, 8], high-dimensional logistic regression under data augmentation [9], and inverse-Hessian transmission of objective perturbations [10]. These studies do not identify, relative to an exactly matched class reweighting, how a specific oversampler separates into endpoint-selection and interpolation effects.

More broadly, imbalanced learning distinguishes data-level resampling from algorithm-level weighting and cost-sensitive strategies [11–13]. Recent work has examined how resampling and class weighting affect probability estimation in unbalanced classification [14], while synthetic oversampling continues to motivate new method design in current data-science research [15]. Our focus is different: rather than proposing another balancing rule or correcting posterior probabilities, we isolate the residual optimization effect of interpolation after exactly matching the class-weight contribution.

What does interpolation-based oversampling change beyond an exactly matched class reweighting?

We construct a matched class-weight control and isolate the residual geometry-only perturbation. For any differentiable positive-class loss and fixed interpolation generator, the residual gradient admits the exact endpoint-affine decomposition

$$\nabla \Delta_M(\theta) = B^M_{\mathrm{sel}}(\theta) + B^M_{\mathrm{curv}}(\theta).$$

The selection term records the effective endpoint masses induced by the generator; the curvature term records the deviation of the synthetic-point gradient from the barycentric interpolation of endpoint gradients. The identity is algebraic, applies to expected and finite realized generators, and requires no small-perturbation approximation.

This decomposition yields method-resolved signatures. Expected random oversampling coincides with matched class weighting; global interpolation with uniform endpoint marginals is a pure curvature intervention; and standard SMOTE generally contains both mechanisms because its directed neighbour graph redistributes endpoint mass. Borderline and risk-weighted variants introduce further base-point, neighbour-flow, or interpolation-depth asymmetry.

For a smooth regularized empirical-risk learner, the signature propagates through the matched-control Hessian:

$$\theta_M'(0) = -\tau H_0^{-1}[B_{\mathrm{sel}}^M(\theta_0) + B_{\mathrm{curv}}^M(\theta_0)].$$

The inverse-Hessian transmission is standard; the new object is the exact oversampling-specific forcing vector on which it acts. Projecting the first-order response onto a smooth held-out objective gives a target-conditional validation score that approximates candidate rankings without fitting every complete oversampled model.

The contributions are threefold. First, an exactly matched class-weight control isolates interpolation beyond scalar reweighting. Second, the endpoint-affine selection-curvature decomposition gives structural zero-effect conditions and explicit signatures for common oversamplers. Third, the signature yields optimizer and target responses that are tested by component ablation and target-conditional validation screening.

The scope is deliberately limited. We study label-preserving interpolation generators fixed during fitting and smooth objectives with a locally identifiable optimizer. We do not claim that synthetic observations create information, that interpolation necessarily improves predictive risk, or that the inverse-Hessian response is new. The aim is mechanism identification and pre-fit approximation of a specified validation ranking, not universal ranking, independent test-set selection, or minimax optimality.

2. Setup and matched-control identification

2.1 Data, loss, and interpolation generator

Let the majority and minority training samples be

$$D_0 = \{x_k^{(0)}\}_{k=1}^{n_0}, \qquad D_1 = \{x_i^{(1)}\}_{i=1}^{n}.$$

Let $\Theta \subset \mathbb{R}^p$ be open. For $y \in \{0,1\}$, let $\ell_y(\theta; x)$ be differentiable in θ, and let $\Omega(\theta)$ be a smooth regularizer. Define the positive-class gradient map

$$g_\theta(x) = \nabla_\theta \ell_1(\theta; x).$$

The decomposition below requires no logistic or linear-score form.

To allow methods that may use an expanded candidate pool, let $v_1, \dots, v_N$ denote fixed candidate endpoints conditional on the observed training data. Let $u = (u_1, \dots, u_N)$ be the reference endpoint weights, where $u_l \geq 0$ and $\sum_l u_l = 1$. For ordinary minority-only oversampling, $N = n$, $v_l = x_l^{(1)}$, and $u_l = 1/n$. If a method admits additional candidate endpoints, the original minority points retain weights $1/n$ and all added candidates receive reference weight zero.

An interpolation method M generates

$$Z = (1 - \Lambda)v_I + \Lambda v_J, \qquad 0 \leq \Lambda \leq 1,$$

where the conditional law of (I, J, Λ), given the real training sample $D = (D_0, D_1)$, is fixed before fitting and does not depend on θ. The endpoints may coincide, so random replication is included as the case $I = J$.

Suppose that the procedure adds m positive observations. Define

$$W = n_0 + n + m, \qquad \tau = \frac{m}{W}.$$

If $r = m/n$ is the synthetic-to-minority ratio and $\mathrm{IR} = n_0/n$ is the training imbalance ratio, then

$$\tau = \frac{r}{\mathrm{IR} + 1 + r}.$$

Thus the nominal oversampling ratio r and the perturbation mass τ are not the same. In strongly imbalanced data, even $r = 1$ can represent a small perturbation of the complete empirical objective.

2.2 The exactly matched class-weight control

The expected oversampling objective is

$$\overline{R}_M(\theta) = \frac{1}{W}\left[\sum_{k=1}^{n_0} \ell_0\left(\theta; x_k^{(0)}\right) + \sum_{i=1}^{n} \ell_1\left(\theta; x_i^{(1)}\right) + m\,\mathbb{E}_M\{\ell_1(\theta; Z) \mid D\}\right] + \Omega(\theta).$$

The matched class-weight objective retains the real observations but allocates the same additional minority loss mass to the reference endpoint measure:

$$R_{\mathrm{cw}}(\theta) = \frac{1}{W}\left[\sum_{k=1}^{n_0} \ell_0\left(\theta; x_k^{(0)}\right) + \sum_{i=1}^{n} \ell_1\left(\theta; x_i^{(1)}\right) + m\sum_{l=1}^{N} u_l\, \ell_1(\theta; v_l)\right] + \Omega(\theta).$$

For ordinary minority-only oversampling, this is equivalent to assigning every real minority observation the weight $1 + m/n = 1 + r$ and every majority observation weight one, while using the same normalization W as the oversampled objective.

Define the geometry-only objective perturbation

$$\Delta_M(\theta) = \mathbb{E}_M\{\ell_1(\theta; Z) \mid D\} - \sum_{l=1}^{N} u_l\, \ell_1(\theta; v_l).$$

Theorem 1 (matched-control identity). For every $\theta \in \Theta$,

$$\boxed{\overline{R}_M(\theta) = R_{\mathrm{cw}}(\theta) + \tau\Delta_M(\theta).}$$

The identity is exact. It does not rely on a Taylor expansion, large-sample approximation, independence among generated observations, or a specific loss. The quantity Δ_M is precisely the residual effect that remains after matching the total minority loss mass.

A direct comparison between $\overline{R}_M$ and an unweighted original-data objective includes both class-mass and geometry effects. Theorem 1 removes the first effect by construction. This normalization is essential for interpreting the decomposition below: the resulting mechanism signature describes interpolation beyond class weighting, not oversampling in the undifferentiated sense.

3. Exact selection-curvature decomposition

3.1 Endpoint-affine reference

Let

$$w_{ij} = \Pr_M(I = i, J = j \mid D), \qquad \mu_{ij} = \mathbb{E}_M(\Lambda \mid I = i, J = j, D),$$

with the convention that terms with $w_{ij} = 0$ are omitted. The total barycentric mass assigned to endpoint v_l is

$$a_l^M = \sum_{j=1}^{N} w_{lj}\left(1 - \mu_{lj}\right) + \sum_{i=1}^{N} w_{il}\,\mu_{il}.$$

Because every generated point has barycentric coefficients summing to one,

$$\sum_{l=1}^{N} a_l^M = 1.$$

For a generated triple (I, J, Λ), the endpoint-affine reference to the gradient at Z is

$$(1 - \Lambda)g_\theta(v_I) + \Lambda g_\theta(v_J).$$

This endpoint-affine reference is the barycentric interpolation of the two endpoint gradient evaluations using the same coefficient that constructs the synthetic point. It reproduces the synthetic-point gradient whenever the gradient map is affine along the used segment.

Define the endpoint-selection term

$$B_{\mathrm{sel}}^M(\theta) = \sum_{l=1}^{N} (a_l^M - u_l)\, g_\theta(v_l),$$

and the interpolation-curvature term

$$B_{\mathrm{curv}}^M(\theta) = \mathbb{E}_M\left[g_\theta(Z) - (1 - \Lambda)g_\theta(v_I) - \Lambda g_\theta(v_J) \mid D\right].$$

Theorem 2 (exact selection-curvature decomposition). For every $\theta \in \Theta$,

$$\boxed{\nabla\Delta_M(\theta) = B_{\mathrm{sel}}^M(\theta) + B_{\mathrm{curv}}^M(\theta).}$$

The selection term records how the generator redistributes endpoint mass relative to the reference measure u. It depends on which observations serve as bases and neighbours and on the average interpolation depth assigned to each directed edge. The curvature term records the non-affinity of the learner's positive-class gradient along the generated edges. It therefore depends jointly on the synthetic geometry and the learner.

The decomposition should not be interpreted as an orthogonal decomposition. The two vectors may align, oppose, or cancel. Its value is structural: it separates the part representable by endpoint reweighting from the residual edge-interpolation effect under a fixed barycentric reference.

3.2 Structural characterization

For a fixed edge and coefficient, define the edgewise curvature remainder

$$\rho_\theta(i,j,\lambda) = g_\theta\left((1-\lambda)v_i + \lambda v_j\right) - (1-\lambda)g_\theta(v_i) - \lambda g_\theta(v_j).$$

The following proposition states the principal zero-effect and equivalence conditions.

Proposition 1 (selection, curvature, and matched-reweighting equivalence).

1. If $a_l^M = u_l$ for every l, then $B_{\mathrm{sel}}^M(\theta) = 0$ for every θ. Conversely, at a fixed θ, if the vectors
$$\{g_\theta(v_l) - g_\theta(v_N)\}_{l=1}^{N-1}$$
are linearly independent, then $B_{\mathrm{sel}}^M(\theta) = 0$ implies $a_l^M = u_l$ for every l.
2. If g_θ is affine on every interpolation segment used with positive generator probability, then $B_{\mathrm{curv}}^M(\theta) = 0$. More precisely, $\rho_\theta(i,j,\lambda) = 0$ if and only if the endpoint-affine relation holds at that generated triple. The expected curvature can also vanish through cancellation even when individual edge remainders are nonzero.
3. The expected oversampling and matched-control objectives are gradient-equivalent at θ if and only if
$$B_{\mathrm{sel}}^M(\theta) + B_{\mathrm{curv}}^M(\theta) = 0.$$
If this identity holds throughout a connected open set, then Δ_M is constant on that set. Hence the two objectives have the same minimizers there whenever those minimizers are unique.

The proposition clarifies that endpoint balance and gradient affinity are separate sufficient mechanisms for equivalence. They are not individually necessary because selection and curvature may cancel. The complete criterion is the vanishing of their vector sum.

3.3 Signatures of common oversamplers

The decomposition turns common oversampling rules into explicit gradient signatures.

3.3.1 Random oversampling

Let $I = J$ and let I be drawn from the reference weights u. Then $Z = v_I$, $a_l^M = u_l$, and every edgewise curvature remainder is zero. Therefore

$$B_{\mathrm{sel}}^{\mathrm{ROS}}(\theta) = B_{\mathrm{curv}}^{\mathrm{ROS}}(\theta) = 0.$$

Expected random oversampling is exactly equivalent to the matched class-weight objective. A finite replication sample can nevertheless have a nonzero empirical selection term because the realized replication counts fluctuate around u.

3.3.2 Global uniform interpolation

Suppose that I and J both have marginal distribution u and that Λ is independent of the endpoints. Then

$$a_l^M = \mathbb{E}(1-\Lambda)u_l + \mathbb{E}(\Lambda)u_l = u_l,$$

so

$$B_{\mathrm{sel}}^{\mathrm{UI}}(\theta) = 0.$$

Global uniform interpolation is therefore a pure expected-curvature intervention. Its optimizer effect is entirely due to the non-affinity of g_θ along globally selected interpolation edges.

3.3.3 Standard SMOTE

Standard SMOTE [1] selects each minority observation as a base point with probability $1/n$. Given base i, let J be uniform on its directed k-nearest-neighbour set $\mathcal{N}_k(i)$. Suppose that Λ is independent of the chosen edge and write $\bar{\lambda} = \mathbb{E}\Lambda$. Let d_l^- denote the directed indegree of node l, namely the number of base points i for which $l \in \mathcal{N}_k(i)$. Then

$$a_l^{\mathrm{SMOTE}} = \frac{1-\bar{\lambda}}{n} + \frac{\bar{\lambda}}{nk} d_l^-,$$

and hence

$$\boxed{B_{\mathrm{sel}}^{\mathrm{SMOTE}}(\theta) = \frac{\bar{\lambda}}{n} \sum_{l=1}^{n} \left(\frac{d_l^-}{k} - 1 \right) g_\theta\left(x_l^{(1)}\right).}$$

Thus SMOTE is not generally selection-neutral even though its base points are sampled uniformly. Observations that appear as neighbours unusually often receive excess endpoint mass. The selection term vanishes when the directed neighbour graph is indegree-regular, or when the resulting gradient contributions cancel.

3.3.4 Borderline-SMOTE-1

Borderline-SMOTE-1 [16] uses $q \in \mathbb{R}^n$ as the base-point distribution concentrated on dangerous minority observations, and let P be the row-stochastic minority-neighbour transition matrix. If Λ is independent of the chosen edge with mean $\bar{\lambda}$, then the effective endpoint mass vector is

$$a^{\mathrm{BLS}} = \left(1-\bar{\lambda}\right) q + \bar{\lambda} P^\top q.$$

Consequently,

$$B_{\mathrm{sel}}^{\mathrm{BLS}}(\theta) = \sum_{l=1}^{n} \left[\left(1-\bar{\lambda}\right) q_l + \bar{\lambda} (P^\top q)_l - \frac{1}{n} \right] g_\theta\left(x_l^{(1)}\right).$$

This formula separates concentration on dangerous base points from the neighbour-flow effect.

3.3.5 Risk-weighted interpolation

For a general base distribution q, row-stochastic neighbour transition matrix P, and edge-specific mean coefficient μ_{ij},

$$a_l^M = \sum_j q_l \, P_{lj} \left(1 - \mu_{lj}\right) + \sum_i q_i \, P_{il} \mu_{il}.$$

The formula distinguishes changes in base selection, neighbour flow, and interpolation depth. Two risk-weighted methods can therefore have the same generated sample count but different selection signatures.

Table 1 summarizes the expected mechanism structure.

Method	Expected selection term	Expected curvature term	Structural source
Random oversampling	zero	zero	exact matched reweighting
Global uniform interpolation	zero	generally nonzero	gradient non-affinity along global edges
SMOTE	generally nonzero	generally nonzero	directed k-NN indegree and local edge curvature
Borderline-SMOTE-1	typically nonzero	generally nonzero	dangerous-base concentration and local edges
Risk-weighted interpolation	generally nonzero	generally nonzero	nonuniform base, neighbour, and depth weights

Table 1 Expected selection-curvature structure of common oversamplers

3.4 Curvature as an edgewise second derivative

The curvature terminology can be made exact. For a used edge, let

$$d = v_j - v_i, \qquad z_t = v_i + td, \qquad h_\theta(t) = g_\theta(z_t).$$

If h_θ is twice continuously differentiable, then

$$\rho_\theta(i,j,\lambda) = -\int_0^1 K_\lambda(t) h_\theta''(t)\, dt,$$

where

$$K_\lambda(t) = \begin{cases} (1-\lambda)t, & 0 \le t \le \lambda, \\ \lambda(1-t), & \lambda < t \le 1. \end{cases}$$

Since K_λ is nonnegative and

$$\int_0^1 K_\lambda(t)\,dt = \frac{\lambda(1-\lambda)}{2},$$

we obtain the edgewise bound

$$\| \rho_\theta(i,j,\lambda) \| \le \frac{\lambda(1-\lambda)}{2} \max_{0\le t\le 1} \| h_\theta''(t) \|.$$

This representation shows that curvature is not a property of the synthetic cloud alone. It is the second directional derivative of the learner's gradient map along the generated edges.

For the logistic specialization, augment x by an intercept coordinate, $\tilde{x} = (1, x^\top)^\top$, and let

$$\ell_y(\theta; x) = \log\left(1 + e^{\theta^\top \tilde{x}}\right) - y\theta^\top \tilde{x}, \qquad \sigma(s) = \frac{1}{1+e^{-s}}.$$

Then

$$g_\theta(x) = (\sigma(\theta^\top \tilde{x}) - 1)\tilde{x}.$$

Let $\tilde{d} = (0, d^\top)^\top$. Along z_t, the second edge derivative is

$$h_\theta''(t) = \sigma''(\theta^\top \tilde{z}_t)\left(\theta^\top \tilde{d}\right)^2 \tilde{z}_t + 2\sigma'(\theta^\top \tilde{z}_t)\left(\theta^\top \tilde{d}\right)\tilde{d}.$$

Using the bounds $|\sigma'| \le 1/4$ and $|\sigma''| \le 1/\left(6\sqrt{3}\right)$ gives the explicit bound in Appendix A.6.

4. Optimizer and target response

4.1 Objective path

Define the matched-control-to-oversampling path

$$R_{M,t}(\theta) = R_{\mathrm{cw}}(\theta) + t\tau\Delta_M(\theta), \qquad 0 \le t \le 1,$$

and let

$$\theta_M(t) \in \arg\min_{\theta\in\Theta} R_{M,t}(\theta).$$

Thus $\theta_M(0) = \theta_0$ is the matched-control optimizer and $\theta_M(1) = \overline{\theta}_M$ is the expected-oversampling optimizer.

Assume that, conditional on the training data:

1. $R_{M,t}$ is twice continuously differentiable in a neighbourhood of the fitted path;
2. the selected minimizer is locally unique and $t \mapsto \theta_M(t)$ is continuous;
3. the path Hessian

$$H_{M,t} = \nabla^2 R_{M,t}\left(\theta_M(t)\right)$$

is nonsingular for $0 \le t \le 1$.

Global strong convexity is sufficient but not necessary. The condition is a local identifiability assumption for the fitted empirical-risk path.

Theorem 3 (exact optimizer-path response). Under the preceding assumptions, the fitted path is continuously differentiable and

$$\boxed{\theta_M'(t) = -\tau H_{M,t}^{-1}\left[B_{\mathrm{sel}}^M\left(\theta_M(t)\right) + B_{\mathrm{curv}}^M\left(\theta_M(t)\right)\right].}$$

In particular,

$$\boxed{\theta_M'(0) = -\tau H_0^{-1} B_M(\theta_0), \qquad B_M = B_{\mathrm{sel}}^M + B_{\mathrm{curv}}^M.}$$

The method signature B_M is the generator-specific forcing vector. The inverse Hessian is the learner-specific transmission operator. A large signature need not produce a large parameter displacement if it lies in a stiff Hessian direction, while a modest signature can be amplified in a weakly curved direction.

4.2 Full-path and first-order optimizer control

Integrating the path equation gives an exact representation of the complete expected-oversampling displacement.

Corollary 3.1 (exact full-path representation and global displacement bound). Under the assumptions of Theorem 3,

$$\bar{\theta}_M - \theta_0 = -\tau \int_0^1 H_{M,t}^{-1}\, B_M\left(\theta_M(t)\right) dt.$$

If, in addition, every path Hessian has smallest eigenvalue at least $\mu > 0$ and the method-signature norm is uniformly bounded by a finite constant G along the fitted path, then

$$\|\bar{\theta}_M - \theta_0\| \le \frac{\tau}{\mu}\int_0^1 \|B_M(\theta_M(t))\|\, dt \le \frac{\tau}{\mu} G.$$

Under uniform Hessian coercivity, the resulting norm bound is a finite-perturbation certificate: a uniformly small signature cannot move the optimizer far from the matched-control solution.

For local prediction, suppose the path derivatives are uniformly bounded as specified below.

$$H_{M,t} \succcurlyeq \mu I, \qquad \| B_M(\theta_M(t)) \| \le G,$$
$$\| \nabla B_M(\theta_M(t)) \| \le A, \qquad \| \nabla^3 R_{M,t}(\theta_M(t)) \| \le T.$$

Corollary 3.2 (first-order optimizer approximation). Under these conditions,

$$\boxed{\overline{\theta}_M - \theta_0 = -\tau H_0^{-1} B_M(\theta_0) + R_{\theta,M},}$$

where

$$\boxed{\| R_{\theta,M} \| \le \tau^2 \left(\frac{AG}{\mu^2} + \frac{TG^2}{2\mu^3} \right).}$$

Thus the perturbation mass, rather than the nominal synthetic-to-minority ratio, is the natural local scale; the quadratic remainder explains why direction accuracy can remain high after magnitude accuracy deteriorates.

4.3 Target-conditional response and screening score

Let V be a fixed continuously differentiable population or held-out validation objective, with smaller values preferred. Conditional on the training and validation data, V is deterministic.

Corollary 3.3 (exact target-functional path response). Under the assumptions of Theorem 3,

$$V(\bar{\theta}_M) - V(\theta_0) = -\tau \int_0^1 \nabla V(\theta_M(t))^\top H_{M,t}^{-1} B_M(\theta_M(t))\, dt.$$

This identity projects the Hessian-transmitted method signature onto the specified target along the fitted path.

Freezing this response at the matched-control solution gives the computable first-order score below; smoothness supplies a quadratic remainder bound.

$$S_M(V) := -\tau\, \nabla V(\theta_0)^\top H_0^{-1} B_M(\theta_0).$$

Proposition 2 (first-order validation response and ordering certificate). Under the preceding assumptions,

$$V(\bar{\theta}_M) - V(\theta_0) = S_M(V) + R_{V,M}.$$

with the deterministic remainder bound

$$|R_{V,M}| \le \tau^2 \left[\|\nabla V(\theta_0)\| C_{\theta,M} + \frac{L_V G_M^2}{2\mu_M^2} \right] =: \varepsilon_{V,M}, \qquad C_{\theta,M} := \frac{A_M G_M}{\mu_M^2} + \frac{T_M G_M^2}{2\mu_M^3}.$$

The score requires one matched-control fit, its Hessian, the candidate signature, and the validation gradient; it does not require fitting the complete oversampled model. The explicit uncertainty radius is used only when a certified ordering is required.

For two generators sharing the same reference measure and perturbation mass, separated score intervals give a sufficient pairwise ordering certificate; comparison with zero gives the corresponding certificate relative to matched class weighting.

$$S_{M_1}(V) + \varepsilon_{V,M_1} < S_{M_2}(V) - \varepsilon_{V,M_2} \quad \Rightarrow \quad V(\bar{\theta}_{M_1}) < V(\bar{\theta}_{M_2}).$$

Likewise, comparison with the matched control is certified by

$$S_M(V) + \varepsilon_{V,M} < 0 \quad \Rightarrow \quad V(\bar{\theta}_M) < V(\theta_0), \qquad S_M(V) - \varepsilon_{V,M} > 0 \quad \Rightarrow \quad V(\bar{\theta}_M) > V(\theta_0).$$

Overlapping intervals leave the ordering unresolved. The score is target-conditional: a method can be preferred for one validation objective and rejected for another.

5. Empirical validation

5.1 Experimental design

The empirical study uses two smooth convex learners: L^2-regularized logistic regression and L^2-regularized probit regression. For each learner, the realized selection-curvature signature, matched-control Hessian, parameter displacement, and held-out target displacement are computed under the same matched-control normalization. The interpolation generators, cross-validation folds, and generator seeds are shared across learners, so the comparison changes the learner while holding the realized synthetic geometry fixed.

The experiment uses 12 public imbalanced binary datasets: Pulsar-Dataset-HTRU2, arrhythmia, default-of-credit-card-clients, dis, haberman, jm1, mammography, mfeat-fourier, oil, satimage, wisconsin, and yeast_me2. They span 306–30,000 observations, 3–261 effective features, and training imbalance ratios of 1.86–64.03.

Three-fold stratified cross-validation is used. The supplied analysis files are numeric and contain no missing entries; a fold-fitted median-imputation stage is retained as an inactive safeguard, globally constant predictors are removed, and the remaining predictors are standardized using training-fold statistics only. Every validation fold retains its original class distribution, and oversampling is applied only within the training fold. Every oversampled fit is compared with a class-weight control having the same total minority loss mass and the same normalization. All cross-validation and generator randomness is deterministically derived from master seed 20260901.

Logistic regression uses C = 1, the liblinear solver, tolerance 10^{-8}, at most 3,000 iterations, and intercept scaling 1,000. Probit regression uses the smooth loss $\ell(y,s) = -\log \Phi((2y-1)s)$, the normalized L^2 penalty corresponding to C = 1 with the same intercept scaling, and L-BFGS-B optimization with gradient tolerance 10^{-8}. Five generators are evaluated:

- random oversampling;
- global uniform interpolation;
- risk-weighted interpolation, with base-point and minority-neighbour probabilities proportional to $\rho_i + 0.05$, where ρ_i is the fraction of majority observations among the 10 nearest training neighbours;
- standard SMOTE with k = 5;
- Borderline-SMOTE-1 with k = 5 and the 10-neighbour danger criterion.

The synthetic-to-minority ratios are $r \in \{0.1, 0.5, 1\}$. For each dataset-fold-ratio-method combination, the same realized synthetic sample is evaluated under both learners. The resulting design contains

$$12 \times 3 \times 5 \times 3 \times 2 = 1{,}080$$

successful learner-dataset-fold-method-ratio comparisons, with no run errors.

For each realized generator, the finite-sample counterparts $\hat{B}_{\mathrm{sel}}$, $\hat{B}_{\mathrm{curv}}$, and $\hat{B}$ are computed at the matched-control optimum $\hat{\theta}_0$. The predicted parameter displacement is

$$\Delta\hat{\theta}_{\mathrm{pred}} = -\tau \hat{H}_0^{-1} \hat{B}(\hat{\theta}_0),$$

and the observed displacement is

$$\Delta\hat{\theta}_{\mathrm{obs}} = \hat{\theta}_M - \hat{\theta}_0.$$

To test whether either mechanism alone can reproduce the fitted response, we evaluate three predictions at the same matched-control optimum:

$$\delta_{\mathrm{sel}} = -\tau H_0^{-1} B_{\mathrm{sel}}, \qquad \delta_{\mathrm{curv}} = -\tau H_0^{-1} B_{\mathrm{curv}}, \qquad \delta_{\mathrm{full}} = -\tau H_0^{-1}(B_{\mathrm{sel}} + B_{\mathrm{curv}}).$$

Because the response map is linear in the forcing vector, the full prediction is the sum of the selection-only and curvature-only responses. The ablation therefore tests whether either component alone is sufficient to reproduce the fitted displacement, including cases in which the two components reinforce or partially cancel.

The principal diagnostics are the direction cosine

$$\frac{\langle \Delta\hat{\theta}_{\mathrm{pred}}, \Delta\hat{\theta}_{\mathrm{obs}} \rangle}{\| \Delta\hat{\theta}_{\mathrm{pred}} \| \| \Delta\hat{\theta}_{\mathrm{obs}} \|}$$

and the relative error

$$\frac{\| \Delta\hat{\theta}_{\mathrm{pred}} - \Delta\hat{\theta}_{\mathrm{obs}} \|}{\| \Delta\hat{\theta}_{\mathrm{obs}} \|}.$$

Held-out score displacement is used as a secondary optimizer-level diagnostic. Section 5.4 separately treats the held-out fold as a validation sample for the smooth target-functional score of Section 4.3. The screening experiment is learner-

specific and evaluates approximation of the corresponding full-fit validation ranking; it does not make an independent test-set selection claim.

5.2 Optimizer-response accuracy

The finite-generator identity holds to numerical precision. We therefore focus on the substantive question: how accurately the first-order response predicts the displacement of the fitted oversampled model.

$$\max \| \hat{B} - \hat{B}_{\mathrm{sel}} - \hat{B}_{\mathrm{curv}} \| = 4.907 \times 10^{-15}.$$

The maximum gradient norm at a fitted matched-control optimum is 2.273×10^{-5} for logistic regression and 6.973×10^{-7} for probit regression. These checks confirm that the objective normalization, finite-generator bookkeeping, and numerical optimization are mutually consistent for both learners.

Across 540 comparisons per learner, logistic regression attains a median parameter-direction cosine of 0.997, first quartile 0.990, and median relative error 0.115; probit regression attains 0.997, 0.990, and 0.122, respectively. In the more local regime $r \leq 0.5$, the median direction cosines are 0.998 for both learners and the median relative errors are 0.096 for logistic and 0.097 for probit. The median held-out score-direction cosines are 0.997 and 0.996. Fig. 1 shows direction accuracy against synthetic loss mass for both learners, and Table 2 summarizes accuracy by learner and oversampling ratio.

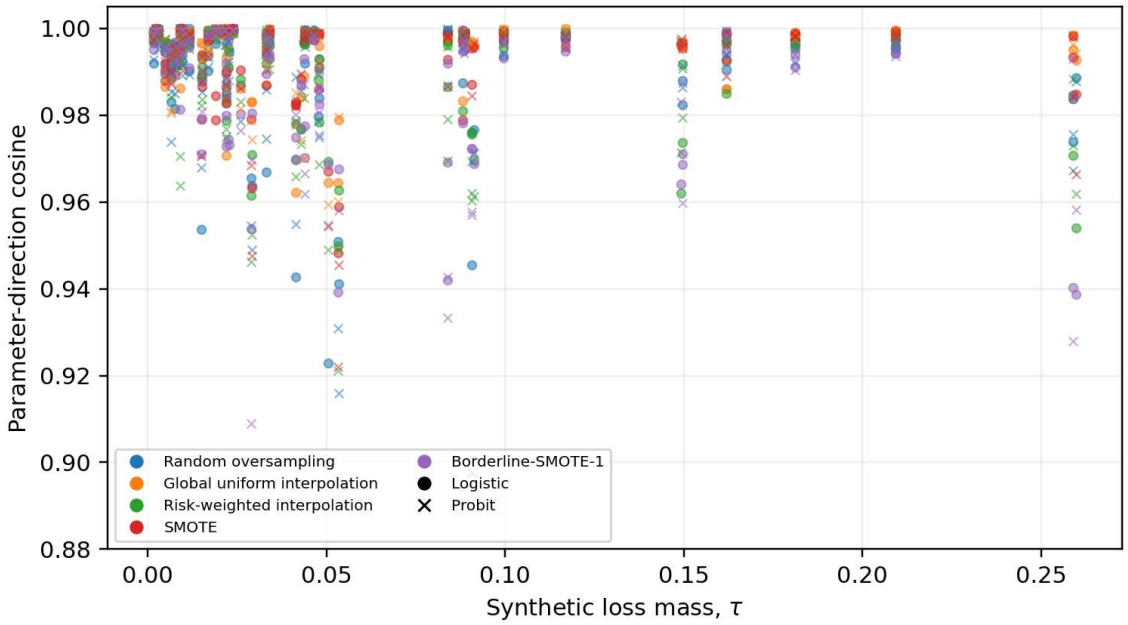


Fig. 1 First-order parameter direction accuracy as a function of synthetic loss mass for logistic and probit regression

Learner	r	Median τ	Median cos.	Q1 cos.	Median rel. err.
Logistic	0.1	0.009	0.999	0.996	0.062
Logistic	0.5	0.045	0.996	0.988	0.134
Logistic	1.0	0.086	0.994	0.979	0.167
Probit	0.1	0.009	0.999	0.997	0.058
Probit	0.5	0.045	0.997	0.989	0.130
Probit	1.0	0.086	0.994	0.979	0.185

Table 2 First-order parameter accuracy by learner and oversampling ratio

5.3 Component ablation and method signatures

The ablation compares the observed optimizer displacement with selection-only, curvature-only, and full first-order predictions for each learner. For logistic regression, the full signature attains median parameter-direction cosine 0.997 and median relative error 0.115, compared with 0.987 and 0.208 for selection alone; curvature alone gives cosine 0.214 over the 432 nonzero-curvature comparisons and relative error 0.992. For probit regression, the corresponding full values are 0.997 and 0.122, compared with 0.983 and 0.245 for selection alone and 0.183 and 0.994 for curvature alone. Fig. 2 and Table 3 summarize the cross-learner ablation results.

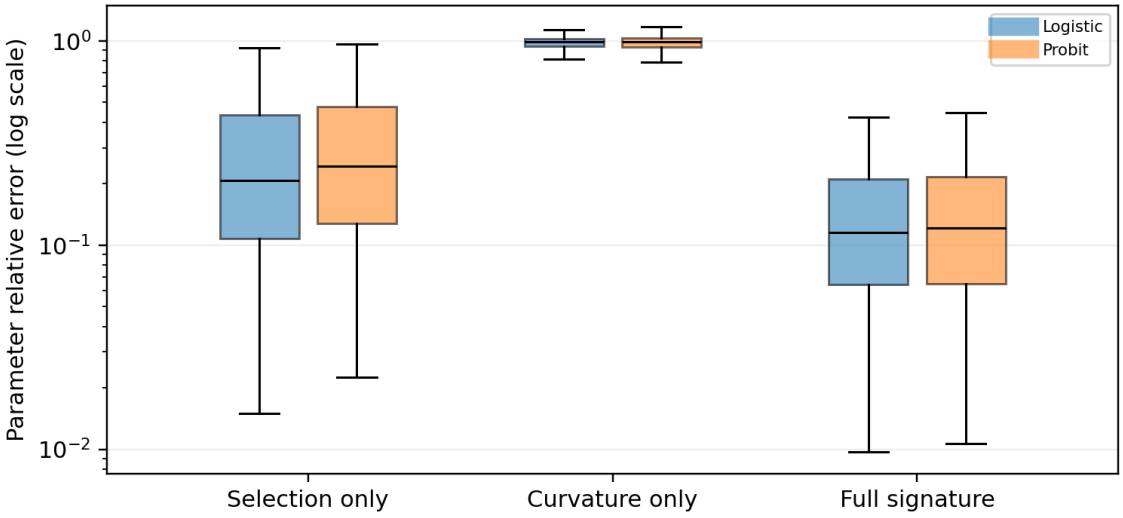


Fig. 2 Parameter-relative-error distributions for the selection-only, curvature-only, and full predictions under both learners

Learner	Prediction	Param. cos.	Param. rel. err.	Score cos.	Score rel. RMSE
Logistic	Selection	0.987	0.208	0.987	0.238
Logistic	Curvature	0.214	0.992	0.214	0.995
Logistic	Full	0.997	0.115	0.997	0.136
Probit	Selection	0.983	0.245	0.982	0.282
Probit	Curvature	0.183	0.994	0.211	0.994
Probit	Full	0.997	0.122	0.996	0.143

Table 3 Overall component-ablation diagnostics by learner

Random oversampling is the internal control: its curvature component is identically zero, so the full and selection-only predictions coincide. Excluding the 108 structural random-oversampling ties within each learner, the full signature has lower relative error than selection alone in 93.3% of logistic and 96.3% of probit comparisons, than curvature alone in 97.5% and 98.4%, and than both single-component predictions in 91.2% and 94.7%. The median improvement over the better single-component predictor is 0.075 for logistic regression and 0.119 for probit regression.

The held-out score analysis shows the same pattern. For logistic regression, the median score-direction cosine rises from 0.987 for selection alone and 0.214 for curvature alone to 0.997 for the full signature, while median relative RMSE falls from 0.238 and 0.995 to 0.136. For probit regression, the corresponding cosines are 0.982, 0.211, and 0.996, and the relative RMSE values are 0.282, 0.994, and 0.143. Thus the gain from combining the two mechanisms is reproduced under a distinct smooth loss.

The empirical method signatures are also structurally stable across learners. Borderline-SMOTE, risk-weighted interpolation, and SMOTE are selection dominated under the implemented generators; random oversampling has identically zero curvature; and global uniform interpolation has zero expected selection but nonzero realized selection from finite endpoint-count fluctuations. Curvature remains nonzero for genuine interpolation methods and changes with the learner, as predicted by Section 3. Fig. 3 and Table 4 summarize the learner-specific signatures.

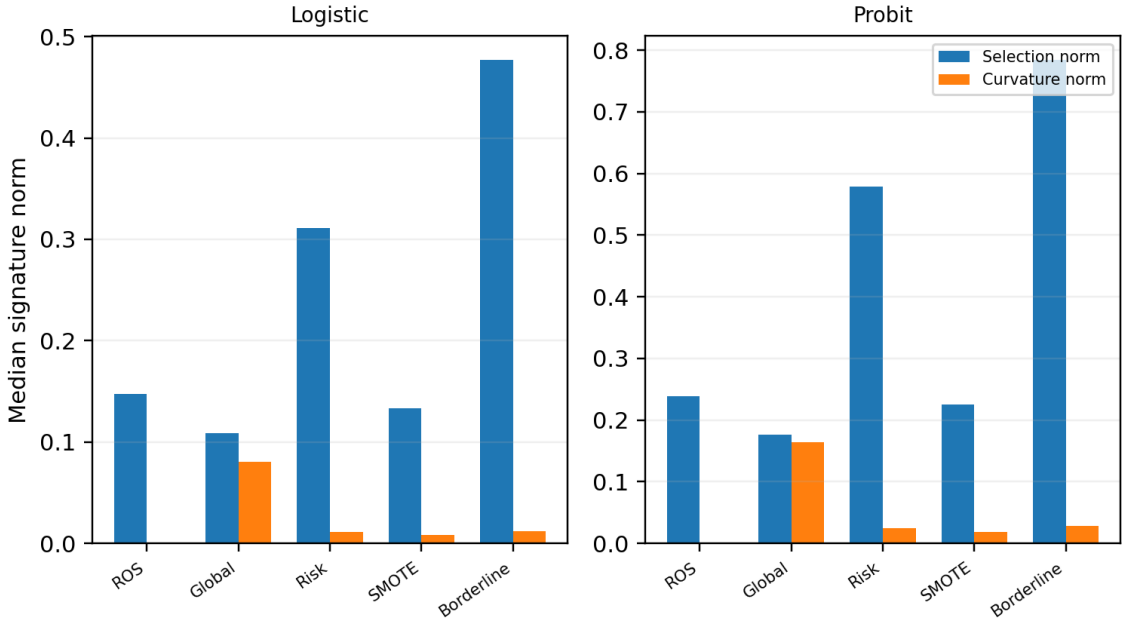


Fig. 3 Method-specific selection and curvature signatures for logistic and probit regression

Logistic

Method	B_s	B_c	Sel. frac.	Cancel.
ROS	0.147	0.000	1.000	1.000
Global	0.109	0.081	0.594	0.759
Risk-wtd.	0.311	0.011	0.964	0.962
SMOTE	0.133	0.008	0.910	0.942
Borderline	0.477	0.012	0.971	0.971

Probit

Method	B_s	B_c	Sel. frac.	Cancel.
ROS	0.239	0.000	1.000	1.000
Global	0.176	0.164	0.578	0.772
Risk-wtd.	0.579	0.024	0.952	0.946
SMOTE	0.226	0.018	0.903	0.934
Borderline	0.784	0.029	0.969	0.969

Table 4 Median empirical selection-curvature signature diagnostics by learner

5.4 Target-conditional method screening

Each learner-dataset-fold-ratio combination defines a five-method screening problem, giving 108 problems per learner and 216 in total. The matched-control model is fitted once for each learner and group, candidate signatures and scores are computed at the shared learner-specific control solution, and complete candidate fits are used only to establish the oracle validation ranking.

Two smooth held-out validation objectives are used: binary log-loss and Brier score. Let $p\theta(x)$ denote the fitted positive-class probability, with $p\theta(x) = \sigma(\theta^T x)$ for logistic regression and $p\theta(x) = \Phi(\theta^T x)$ for probit regression. With validation sample size n_v, the objectives are

$$V_{\log}(\theta) = -\frac{1}{n_v}\sum_{i=1}^{n_v}\{y_i \log p_\theta(x_i) + (1-y_i)\log[1-p_\theta(x_i)]\}, \qquad V_{\mathrm{Br}}(\theta) = \frac{1}{n_v}\sum_{i=1}^{n_v}[p_\theta(x_i) - y_i]^2.$$

For either objective V, the screened method minimizes the first-order score over the five candidates. Its selection regret is

$$\mathrm{Regret}_V = V(\bar{\theta}_{\widehat{M}}) - \min_{M\in\mathcal{M}} V(\bar{\theta}_M), \qquad \widehat{M} \in \arg\min_{M\in\mathcal{M}} S_M(V).$$

For each learner and objective V, the screened method minimizes the first-order score over the five candidates. We report top-one and top-two recovery, rank correlation, and selection regret. A target-free baseline selects the smallest total-signature norm.

For logistic regression, the log-loss score selects the oracle-best method in 87.0% of the 108 groups and places it in the top two in 93.5%; the corresponding Brier rates are 83.3% and 94.4%. For probit regression, the rates are 88.0% and 96.3% for log-loss and 86.1% and 96.3% for Brier score. Pooled across 216 learner-specific problems, top-one recovery

is 87.5% for log-loss and 84.7% for Brier score. Median Spearman rank correlation is 1.0 for both objectives and both learners; mean correlations are 0.900 and 0.890 for logistic regression and 0.921 and 0.904 for probit regression. Median selection regret is zero in all four learner-target combinations. The 90th-percentile absolute regret is 8.97×10^{-5} and 4.81×10^{-5} for logistic log-loss and Brier score, and 8.02×10^{-5} and 1.09×10^{-5} for probit; these equal 4.54%, 4.90%, 3.32%, and 3.49% of the corresponding within-group objective ranges. Fig. 4 and Table 5 summarize the screening results.

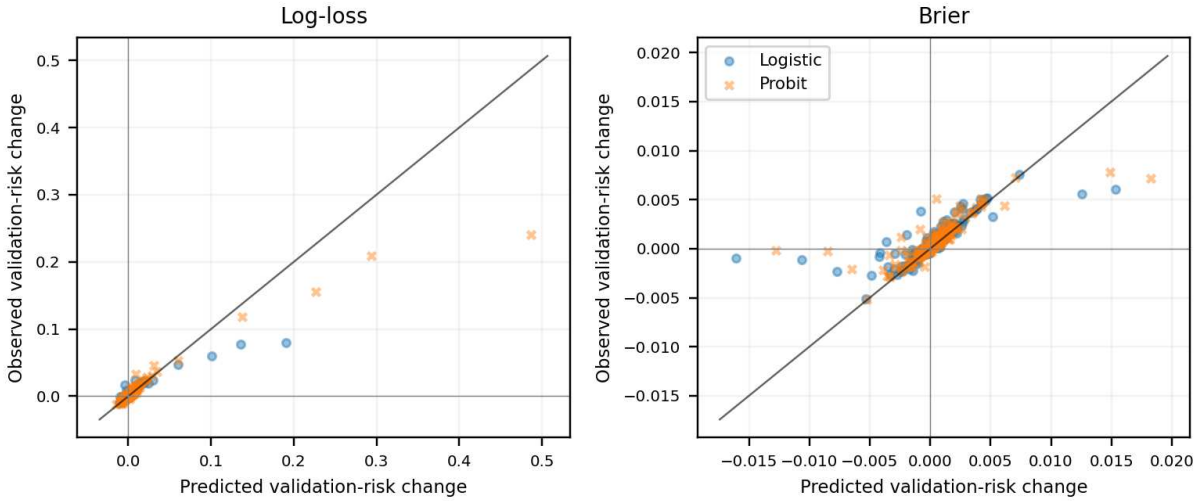


Fig. 4 Predicted and observed held-out validation-risk changes for all 1,080 learner-candidate fits

Learner	Target	Top-1	Top-2	Med. ρ	Mean ρ	P90 regret	P90/range
Logistic	Log-loss	87.0%	93.5%	1.000	0.900	8.97×10^{-5}	4.54%
Logistic	Brier	83.3%	94.4%	1.000	0.890	4.81×10^{-5}	4.90%
Probit	Log-loss	88.0%	96.3%	1.000	0.921	8.02×10^{-5}	3.32%
Probit	Brier	86.1%	96.3%	1.000	0.904	1.09×10^{-5}	3.49%

Table 5 Five-method screening performance by learner over 216 learner-dataset-fold-ratio problems

The target projection is essential under both learners. Selecting the smallest total-signature norm attains top-one accuracy of 31.5% and 23.1% for logistic log-loss and Brier score, and 29.6% and 19.4% for probit. Selecting the smallest predicted parameter movement gives 20.4% and 18.5% for logistic and 24.1% and 21.3% for probit. All are far below the target-projected screening score, showing that the ranking information comes from the signed validation-gradient projection rather than perturbation magnitude alone.

The screening workflow is also cheaper than fitting all five complete candidate models. Excluding the shared control fit, the measured median full-fit-to-screening time ratio is 2.06 for logistic regression and 5.95 for probit regression in the present implementation. Timing details and their hardware-dependent interpretation are given in Appendix B.1.

The experiment evaluates recovery of a ranking on the specified learner-specific validation objective, not post-selection generalization on an independent test set.

6. Discussion

The matched-control perspective changes the interpretation of interpolation-based oversampling. After the additional minority loss mass is exactly reproduced by class weighting, the remaining perturbation is attributable to endpoint-mass redistribution and gradient non-affinity along interpolation segments. The empirical results support this mechanism view under two distinct smooth losses: the full selection-curvature response is substantially more accurate than either component alone for both logistic and probit regression, while the method-level patterns agree with the structural formulas in Section 3.3. The close cross-learner agreement also indicates that the observed first-order accuracy is not an artifact of the logistic loss alone. These signatures still describe how a specified generator acts on a specified learner; they do not imply a universal ranking of oversamplers.

The target projection adds a method-screening interpretation. A signature has no intrinsic preference direction until it is transmitted through the learner Hessian and projected onto a specified smooth validation objective. The resulting first-order score approximates the validation ranking produced by complete candidate fits for both learners, with top-one

recovery near 88% for log-loss and 83–86% for Brier score and zero median selection regret. This claim is conditional on the observed training and validation data and does not establish post-selection generalization to an independent test set.

The analysis has several limitations. The generator law is fixed during fitting, and the objective must be sufficiently smooth with a locally identifiable optimizer. The experiments now cover two smooth convex learners with different link losses but fixed hyperparameters and a linear predictor structure; broader nonlinear smooth architectures, adaptive generators, and nested validation-test selection remain open. Approximation error also grows with synthetic loss mass, and finite generator realizations introduce Monte Carlo endpoint fluctuations around their expected signatures. For practical evaluation, an interpolation oversampler should therefore be compared with an exactly matched class-weight control, with the realized selection-curvature signature reported before any claim of predictive superiority.

7. Conclusion

Interpolation-based oversampling differs from exactly matched class weighting through two identifiable mechanisms: redistribution of effective endpoint mass and non-affinity of the learner gradient along interpolation segments. Their sum gives the exact gradient perturbation relative to the matched control and, after Hessian transmission, determines the local optimizer response. Across regularized logistic and probit regression, the resulting signatures yield closely replicated first-order prediction accuracy, materially outperform single-component responses, and provide a target-conditional approximation to candidate validation rankings. The framework is therefore a mechanism-analysis tool for isolating what interpolation contributes beyond scalar reweighting, rather than a universal guarantee that oversampling improves prediction.

Appendix A. Proofs and technical details

A.1 Proof of Theorem 1

Subtract the matched-control objective from the expected oversampling objective. The majority term, the retained minority term, and the regularizer cancel, leaving

$$\overline{R}_M(\theta) - R_{\mathrm{cw}}(\theta) = \frac{m}{W}\left[\mathbb{E}_M\{\ell_1(\theta; Z) \mid D\} - \sum_{l=1}^{N} u_l\, \ell_1(\theta; v_l)\right].$$

Since $m/W = \tau$, the right-hand side is $\tau \Delta_M(\theta)$.

A.2 Proof of Theorem 2

Differentiate Δ_M with respect to θ and use conditional differentiation under the finite or dominated generator expectation:

$$\nabla \Delta_M(\theta) = \mathbb{E}_M\{g_\theta(Z) \mid D\} - \sum_{l=1}^{N} u_l\, g_\theta(v_l).$$

Add and subtract the endpoint-affine reference:

$$\mathbb{E}_M\big[(1-\Lambda) g_\theta(v_I) + \Lambda g_\theta(v_J) \mid D\big].$$

The difference between the generated gradient and this reference is $B^M_{\mathrm{curv}}(\theta)$. For the remaining term,

$$\mathbb{E}_M\big[(1-\Lambda) g_\theta(v_I) + \Lambda g_\theta(v_J) \mid D\big] = \sum_{i,j} w_{ij}\left[(1-\mu_{ij}) g_\theta(v_i) + \mu_{ij} g_\theta(v_j)\right].$$

Collecting all coefficients multiplying the same endpoint gradient gives

$$\sum_{i,j} w_{ij}\left[(1-\mu_{ij}) g_\theta(v_i) + \mu_{ij} g_\theta(v_j)\right] = \sum_{l=1}^{N} a_l^M\, g_\theta(v_l).$$

Subtracting $\sum_l u_l\, g_\theta(v_l)$ gives $B^M_{\mathrm{sel}}(\theta)$.

A.3 Proof of Proposition 1

If $a_l^M = u_l$ for every l, the selection formula immediately gives $B^M_{\mathrm{sel}} = 0$.

For the converse under the stated independence condition, note that $\sum_l (a_l^M - u_l) = 0$. Hence

$$B_{\mathrm{sel}}^{M}(\theta) = \sum_{l=1}^{N-1} (a_l^M - u_l)\,[g_\theta(v_l) - g_\theta(v_N)].$$

Linear independence forces $a_l^M - u_l = 0$ for $l = 1, \ldots, N-1$, and the zero-sum condition then gives $a_N^M = u_N$.

If g_θ is affine on a used segment, barycentric interpolation gives

$$g_\theta\left((1-\lambda)v_i + \lambda v_j\right) = (1-\lambda)g_\theta(v_i) + \lambda g_\theta(v_j),$$

so the edge remainder is zero. The converse at a fixed triple is merely the same equality read in reverse. Averaging the edge remainders yields B_{curv}^{M}; therefore nonzero edge terms can cancel in expectation.

Finally, Theorem 1 and Theorem 2 give

$$\nabla\overline{R}_M(\theta) - \nabla R_{\mathrm{cw}}(\theta) = \tau[B_{\mathrm{sel}}^{M}(\theta) + B_{\mathrm{curv}}^{M}(\theta)].$$

Because $\tau > 0$ when synthetic observations are added, gradient equivalence is equivalent to the vanishing of the total signature. If the total signature vanishes throughout a connected open set, then $\nabla\Delta_M = 0$ there, so Δ_M is constant. Adding a constant does not change the minimizer.

A.4 Proof of Theorem 3, Corollaries 3.1-3.3, and Proposition 2

The first-order condition along the fitted path is

$$F\big(t, \theta_M(t)\big) := \nabla R_{\mathrm{cw}}\big(\theta_M(t)\big) + t\tau\nabla\Delta_M\big(\theta_M(t)\big) = 0.$$

The derivative of F with respect to θ is $H_{M,t}$, which is nonsingular by assumption. The implicit-function theorem therefore yields a continuously differentiable local path. Differentiating the first-order condition gives

$$H_{M,t}\theta_M{}'(t) + \tau\nabla\Delta_M\big(\theta_M(t)\big) = 0.$$

Theorem 2 gives the displayed formula in Theorem 3.

Integrating the identity in Theorem 3 from t = 0 to t = 1 gives

$$\bar{\theta}_M - \theta_0 = -\tau\int_0^1 H_{M,t}^{-1}\, B_M\big(\theta_M(t)\big)\, dt.$$

Under the uniform coercivity assumption, the inverse-Hessian operator norm is at most 1/μ. Taking norms in the integral identity and using the uniform signature bound G yields

$$\|\bar{\theta}_M - \theta_0\| \le \frac{\tau}{\mu}\int_0^1 \|B_M(\theta_M(t))\|\, dt \le \frac{\tau}{\mu}G.$$

This proves Corollary 3.1.

For Corollary 3.2, first observe that

$$\|\,\theta_M{}'(t)\,\| \le \tau\,\|\,H_{M,t}^{-1}\,\|\|\,B_M\big(\theta_M(t)\big)\,\| \le \frac{\tau G}{\mu}.$$

Differentiate the path equation once more. Since

$$\frac{d}{dt}H_{M,t} = \tau\nabla B_M\big(\theta_M(t)\big) + \nabla^3 R_{M,t}\big(\theta_M(t)\big)[\theta_M{}'(t)],$$

we obtain

$$H_{M,t}\theta_M{}''(t) + 2\tau\nabla B_M\big(\theta_M(t)\big)\theta_M{}'(t) + \nabla^3 R_{M,t}\big(\theta_M(t)\big)[\theta_M{}'(t), \theta_M{}'(t)] = 0.$$

Consequently,

$$\|\,\theta_M{}''(t)\,\| \le \tau^2\left(\frac{2AG}{\mu^2} + \frac{TG^2}{\mu^3}\right).$$

Taylor's theorem between $t = 0$ and $t = 1$ gives

$$\theta_M(1) - \theta_M(0) = \theta_M{}'(0) + \int_0^1 (1-t)\,\theta_M{}''(t)\, dt.$$

The integral has norm at most one half of the preceding uniform bound, which proves Corollary 3.2.

For Corollary 3.3, the chain rule and Theorem 3 give

$$V(\bar{\theta}_M) - V(\theta_0) = -\tau\int_0^1 \nabla V(\theta_M(t))^{\top} H_{M,t}^{-1} B_M\big(\theta_M(t)\big)\, dt.$$

after integrating from t = 0 to t = 1.

For Proposition 2, Taylor's theorem applied to V between the control optimizer and the complete expected-oversampling solution gives, for some point on the connecting segment,

$$V(\bar{\theta}_M) - V(\theta_0) = \nabla V(\theta_0)^\top(\bar{\theta}_M - \theta_0) + \frac{1}{2}(\bar{\theta}_M - \theta_0)^\top \nabla^2 V\big(\tilde{\theta}_M\big)(\bar{\theta}_M - \theta_0).$$

Insert the optimizer expansion from Corollary 3.2. The linear term generated by the leading displacement is exactly the first-order validation score. The remaining linear contribution is bounded by the product of the validation-gradient norm and the optimizer remainder. The quadratic Taylor term is bounded using the validation Hessian bound and the global displacement estimate in Corollary 3.1. Hence

$$\big|R_{V,M}\big| \le \tau^2 \left[\|\nabla V(\theta_0)\| C_{\theta,M} + \frac{L_V G_M^2}{2\mu_M^2} \right] =: \varepsilon_{V,M}, \qquad C_{\theta,M} := \frac{A_M G_M}{\mu_M^2} + \frac{T_M G_M^2}{2\mu_M^3}.$$

which proves the response expansion and its uncertainty radius. The pairwise ordering implication follows by placing each true validation change in the symmetric interval centered at its first-order score with the stated radius. The two one-method statements follow by comparing that interval with zero.

A.5 Finite generated-sample identity

Let the realized synthetic sample be

$$Z_s = (1 - \lambda_s) v_{I_s} + \lambda_s v_{J_s}, \qquad s = 1, \dots, m.$$

Define the realized endpoint masses

$$\hat{a}_l = \frac{1}{m} \sum_{s=1}^{m} [(1 - \lambda_s)\mathbf{1}\{I_s = l\} + \lambda_s \mathbf{1}\{J_s = l\}],$$

and

$$\hat{B}_{\text{sel}}(\theta) = \sum_{l=1}^{N} (\hat{a}_l - u_l)\, g_\theta(v_l),$$

$$\hat{B}_{\text{curv}}(\theta) = \frac{1}{m} \sum_{s=1}^{m} \big[g_\theta(Z_s) - (1 - \lambda_s) g_\theta\big(v_{I_s}\big) - \lambda_s g_\theta\big(v_{J_s}\big) \big].$$

The realized gradient perturbation satisfies the exact identity

$$\frac{1}{m} \sum_{s=1}^{m} g_\theta\,(Z_s) - \sum_{l=1}^{N} u_l\, g_\theta(v_l) = \hat{B}_{\text{sel}}(\theta) + \hat{B}_{\text{curv}}(\theta).$$

No independence assumption is needed. If the triples are conditionally independent draws from the generator law, then

$$\mathbb{E}\big(\hat{B}_{\text{sel}} \mid D\big) = B_{\text{sel}}^M, \qquad \mathbb{E}\big(\hat{B}_{\text{curv}} \mid D\big) = B_{\text{curv}}^M.$$

This unbiasedness applies to the signature at a fixed θ. It does not imply

$$\mathbb{E}\big[\arg\min R_M^{(m)} \mid D\big] = \arg\min \mathbb{E}\big[R_M^{(m)} \mid D\big],$$

which is generally false because optimization is nonlinear.

Remark. Conditional on a realized synthetic sample, define the realized objective perturbation and realized method signature analogously to their expected counterparts. Whenever the regularity and nonsingularity assumptions of Theorem 3 hold for the realized objective path, Theorem 3 and Corollaries 3.1-3.3 apply conditionally, with the expected perturbation and signature replaced by the corresponding realized quantities.

A.6 Logistic interpolation-curvature bound

Let $x, y \in \mathbb{R}^d$, $d = y - x$, $z_\lambda = (1 - \lambda)x + \lambda y$, and $\tilde{d} = (0, d^\top)^\top$. Suppose

$$\max_{0 \le t \le 1} \| \tilde{z}_t \| \le R.$$

For the positive-class logistic gradient,

$$g_\theta(x) = (\sigma(\theta^\top \tilde{x}) - 1)\tilde{x},$$

the edgewise interpolation remainder satisfies

$$\|g_\theta(z_\lambda) - (1-\lambda)g_\theta(x) - \lambda g_\theta(y)\| \le \lambda(1-\lambda)\left[\frac{1}{4}|\theta^\top \tilde{d}|\,\|\,\tilde{d}\,\| + \frac{R}{12\sqrt{3}}(\theta^\top \tilde{d})^2\right].$$

To prove the claim, use the Green-kernel identity and the expression

$$h_\theta''(t) = \sigma''(\theta^\top \tilde{z}_t)(\theta^\top \tilde{d})^2 \tilde{z}_t + 2\sigma'(\theta^\top \tilde{z}_t)(\theta^\top \tilde{d})\tilde{d}.$$

The exact scalar bounds

$$|\sigma'(s)| \le \frac{1}{4}, \qquad |\sigma''(s)| \le \frac{1}{6\sqrt{3}} \quad \text{for all } s,$$

combined with $\int_0^1 K_\lambda(t)dt = \lambda(1-\lambda)/2$, give the result. The two vector components may cancel, so the combined norm bound is not asserted to be globally optimal.

Appendix B. Additional empirical details

The 12-dataset experiment uses 3 folds, 5 methods, 3 ratios, and 2 learners. Every learner-method-ratio combination contributes 36 dataset-fold comparisons, giving 1,080 optimizer-response comparisons in total. Table B1 gives learner- and method-level optimizer diagnostics.

Learner	Method	Med. cos.	Med. rel. err.	Med. sel. norm	Med. curv. norm
Logistic	ROS	0.996	0.112	0.147	0.000
Logistic	Global uniform	0.998	0.096	0.109	0.081
Logistic	Risk-weighted	0.997	0.139	0.311	0.011
Logistic	SMOTE	0.998	0.102	0.133	0.008
Logistic	Borderline-SMOTE	0.995	0.149	0.477	0.012
Probit	ROS	0.997	0.120	0.239	0.000
Probit	Global uniform	0.998	0.106	0.176	0.164
Probit	Risk-weighted	0.997	0.134	0.579	0.024
Probit	SMOTE	0.998	0.106	0.226	0.018
Probit	Borderline-SMOTE	0.996	0.150	0.784	0.029

Table B1 Method-level optimizer diagnostics

The empirical finite-sample selection fraction is defined by

$$\frac{\|\,\hat{B}_{\mathrm{sel}}\,\|}{\|\,\hat{B}_{\mathrm{sel}}\,\| + \|\,\hat{B}_{\mathrm{curv}}\,\|},$$

with the natural convention when both norms vanish. The cancellation ratio is

$$\frac{\|\,\hat{B}_{\mathrm{sel}} + \hat{B}_{\mathrm{curv}}\,\|}{\|\,\hat{B}_{\mathrm{sel}}\,\| + \|\,\hat{B}_{\mathrm{curv}}\,\|}.$$

Values near one indicate little cancellation; smaller values indicate opposing mechanism vectors. These diagnostics are descriptive and do not replace the exact vector identity.

B.1 Method-screening protocol and timing

For every dataset-fold-ratio group, the five methods are evaluated under identical preprocessing, regularization, and matched-control normalization. Candidate scores use the realized finite-generator signature. Full candidate fits are retained only for calculating oracle rankings and selection regret. Exact top-one accuracy counts equality of the predicted and observed minimizers after deterministic method-name tie breaking; top-two accuracy records whether the predicted winner belongs to the two smallest observed validation risks.

The measured screening time is generation plus signature construction and one Hessian solve for each candidate. Full candidate-evaluation time is generation plus complete fitting of the corresponding learner for each candidate. The matched-control fit is excluded from both totals because it is common to the two workflows. Timing comparisons are learner-specific wall-clock measurements from the same implementation and hardware and should not be interpreted as hardware-independent complexity constants.

Statements and Declarations

Funding. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.
Competing Interests. The authors have no relevant financial or non-financial interests to disclose.

Author Contributions. Both authors contributed to study conception, theoretical analysis, experiments, interpretation, and manuscript preparation. Both authors approved the final manuscript.
Data Availability Statement. The benchmark datasets analyzed in this study are publicly available from their original repositories. The dataset names and experimental protocol are specified in Section 5.1. The fold-level result files supporting the findings of this study are available from the corresponding author upon reasonable request.
Code Availability. The preprocessing, oversampling, mechanism-diagnostic, component-ablation, validation-screening, and cross-validation code is available from the corresponding author upon reasonable request.

References

1. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16:321–357. https://doi.org/10.1613/jair.953
2. Fernandez A, Garcia S, Herrera F, Chawla NV (2018) SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research 61:863-905. https://doi.org/10.1613/jair.1.11192
3. Elreedy D, Atiya AF (2019) A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information Sciences 505:32-64. https://doi.org/10.1016/j.ins.2019.07.070
4. Elreedy D, Atiya AF, Kamalov F (2024) A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning 113:4903-4923. https://doi.org/10.1007/s10994-022-06296-4
5. Sakho A, Malherbe E, Scornet E (2026) Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants. In: Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS 2026), Proceedings of Machine Learning Research, vol 300.
6. Ahmad T, Kalan MM, Portier F, Stupfler G (2025) Concentration and excess risk bounds for imbalanced classification with synthetic oversampling. Advances in Neural Information Processing Systems 38:174853-174894.
7. Carratino L, Cisse M, Jenatton R, Vert JP (2022) On Mixup regularization. Journal of Machine Learning Research 23(325):1-31.
8. Zou Y, Verma V, Mittal S, Tang WH, Pham H, Kannala J, Bengio Y, Solin A, Kawaguchi K (2023) MixupE: understanding and improving Mixup from directional derivative perspective. In: Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, PMLR 216:2597-2607.
9. Esmaili Mallory M, Huang KH, Austern M (2025) Universality of high-dimensional logistic regression and a novel CGMT under dependence with applications to data augmentation. In: Proceedings of the Thirty-Eighth Conference on Learning Theory, PMLR 291:1799-1918.
10. Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning, PMLR 70:1885-1894.
11. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239
12. Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 5:221–232. https://doi.org/10.1007/s13748-016-0094-0
13. Chen W, Yang K, Yu Z, Shi Y, Chen CLP (2024) A survey on imbalanced learning: latest research, applications and future directions. Artificial Intelligence Review 57:137. https://doi.org/10.1007/s10462-024-10759-6
14. Lin F, Dai Y, Liu Y (2026) On probability estimation for unbalanced classification. Journal of Statistical Theory and Practice 20:57. https://doi.org/10.1007/s42519-026-00572-5
15. Yang Y, Khorshidi HA, Aickelin U (2026) Handling extreme imbalanced multi-class data with MCDO-BR: a diversity-based synthetic over-sampling technique. International Journal of Data Science and Analytics 22:214. https://doi.org/10.1007/s41060-026-01176-9
16. Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Advances in Intelligent Computing, Lecture Notes in Computer Science, vol 3644, pp 878–887. https://doi.org/10.1007/11538059_91