Project n$^{\underline{o}}$2: Indefinite Kernel Learning - Analyzing Sigmoid Kernel Corrections in Support Vector Machines Advanced Machine Learning (MDS)

Volpi Davide D'Agostino Raffaele Clerici Federico

Date: 16 December 2025

Abstract

We investigate the indefinite sigmoid kernel in Support Vector Machines and evaluate the efficacy of spectral transformation methods designed to enforce positive semi-definiteness. While the non-PSD nature of the sigmoid kernel theoretically compromises convex optimization, it often yields effective discriminative geometries in practice. Using the UCI Breast Cancer Wisconsin dataset, we analyze the kernel's spectral properties and their correlation with classification performance. We compare the unmodified kernel against three correction strategies: spectral shifting, eigenvalue clipping, and normalized clipping. Our results demonstrate that while clipping is the least destructive method, spectral corrections frequently degrade model performance. Crucially, we find that the strongest uncorrected kernels do not usually benefit from PSD enforcement. Also, experiments with a normalized sigmoid kernel yield similar patterns. We conclude that despite the theoretical guarantees of convexification, spectral corrections offer a poor trade-off between computational cost and accuracy when compared to standard PSD kernels like the RBF.

Note: all the notation and abbreviations that will be used can be found in Table 1.

1 Introduction

This study is explicitly exploratory: our goal is to characterize how the eigenvalue spectrum of the sigmoid kernel and its spectral corrections (shifting, clipping, normalized clipping) influence the induced feature-space geometry and the resulting SVM performance. To explore these theoretical insights, we utilize the widely known UCI Breast Cancer Wisconsin dataset [10] as a standard benchmark for binary classification.

Our approach focuses on comparing three strategies for the correction of indefinite kernels: shifting, clipping, and normalized clipping.

Our main contributions include: (i) an analysis of the eigenvalue spectrum of the sigmoid kernel under different parameter settings; (ii) an evaluation of the improvements or degradations introduced by previously proposed fixing strategies; and (iii) a comparative analysis of these experiments using cosine similarity instead of the standard dot product.

The rest of this report is organized as follows: Section 2 states the problem. Section 3 reviews related work on this dataset and analogous problems and establishes the problem context. Section 4 describes the data characteristics, preprocessing pipeline, and exploratory analysis. Section 5 details our experimental protocol and model configurations. Section 6 presents all the results and observations, examining performances and trade-offs across methods.

2 Problem statement

Our objective is to demonstrate whether indefinite kernels, in particular the sigmoid kernel, can still provide an effective representation of the dataset despite lacking positive semi-definiteness guarantees. Furthermore, we investigate if applying kernel-fixing strategies can modify its behaviour, and how. Specifically, we evaluate three modification techniques:

1. **Shifting**, which translates the kernel spectrum by adding a constant $\lambda > 0$.

2. **Clipping**, which enforces a non-negative spectrum by truncating negative eigenvalues.
3. **Normalized clipping**, which rescales the clipped spectrum to preserve relative magnitudes.

3 Related Work

The indefinite nature of the sigmoid kernel was rigorously established by Lin and Lin [5], who demonstrated that it becomes CPD under specific parameter regimes ($\gamma > 0$, small negative c_0) but remains indefinite otherwise, leading to non-convex SVM optimization. They proposed modified SMO algorithms with stationary-point convergence guarantees for non-PSD kernels, though empirical results showed sigmoid kernels generally underperform RBF. Alternative theoretical frameworks have been developed to handle indefiniteness directly: Pekalska et al. [9] embedded indefinite kernels in pseudo-Euclidean spaces, Ong et al. [7] extended theory to reproducing kernel Kreĭn spaces, and Luss and d'Aspremont [6] proposed Robust SVM that treats indefinite kernels as noisy observations of unknown PSD kernels. However, these approaches require specialized algorithms incompatible with standard SVM implementations.

Spectral correction methods modify the eigenvalue spectrum to enforce positive semi-definiteness while maintaining compatibility with convex solvers. Eigenvalue clipping is theoretically distinguished as the unique minimum-Frobenius-norm projection onto the PSD cone and achieves the lowest classification error rates among spectral transformations [2, 4]. Spectral shifting is simpler but modifies all eigenvalues indiscriminately and typically underperforms clipping [4]. Normalized clipping, which rescales the clipped matrix to unit diagonal, approximates Higham's iterative algorithm for nearest correlation matrices [3] but is typically implemented as single-pass normalization for computational efficiency.

Our work differs from previous approaches by conducting a systematic exploration of the sigmoid kernel's behavior across its full parameter space, explicitly correlating spectral properties (eigenvalue distribution, negative mass fraction, CPD status) with classification performance under rigorous cross-validation. Unlike prior benchmarks [2, 4] that evaluate corrections on fixed parameter sets or multiple kernel types, we focus exclusively on sigmoid kernels with dense (γ, c_0) grid search to isolate the effect of indefiniteness degree on correction efficacy. We adopt the standard spectral corrections (shifting, clipping, normalized clipping) for computational tractability but extend evaluation by computing ΔF1 distributions and correlating them with spectral characteristics, providing empirical evidence on whether corrections genuinely enhance performance or merely shift the optimization landscape.

4 Data and Preprocessing

4.1 Data description

The Breast Cancer Wisconsin (Diagnostic) Dataset used in our experiments was introduced by [10] and is hosted in the UCI Machine Learning Repository [1]. The task consists of distinguishing malignant from benign breast tumors based on features computed from digitized images of fine needle aspirates. The target variable is binary, with malignant tumors encoded as 1 and benign tumors as 0.

The dataset contains 30 continuous features derived from ten basic measurements: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, each computed as mean, standard error, and largest value. These features quantify geometric and morphological properties of cell nuclei. The dataset is clean, with no missing values, but presents challenges typical of biomedical data, such as strong feature correlations, heterogeneous scaling, and potential outliers. The classes are moderately imbalanced, with approximately 37% malignant and 63% benign samples.

No extensive exploratory data analysis was performed. The dataset is already clean and fully pre-processed, eliminating the need for standard EDA procedures. Additionally, in-depth EDA was considered unnecessary, as the focus of this study is on evaluating kernel-based methods, and prolonged analyses could have diverted attention from the core experimental goals.

4.2 Preprocessing steps

Minimal preprocessing was applied. Continuous features were standardized (zero mean, unit variance) using `sklearn`'s `StandardScaler` to ensure numerical stability and balanced contributions across features for distance- and gradient-based algorithms. No imputation, outlier removal, transformations, or encoding were necessary, as all features are continuous and the target is already binary. This approach preserves the dataset's structure while maintaining focus on kernel evaluation and classification performance.

5 Methodology

5.1 Experimental Protocol

We utilize the Breast Cancer Wisconsin dataset from `scikit-learn` [10]. The data is split into training (80%) and test (20%) sets using a stratified sampling strategy to maintain class distribution (approx. 63% benign, 37% malignant), with a fixed random seed (42) for reproducibility.

We analyze the sigmoid kernel in two forms:

1. **Standard sigmoid kernel**: $K(\mathbf{x}, \mathbf{y}) = \tanh(\gamma\langle \mathbf{x}, \mathbf{y}\rangle + c_0)$
2. **Normalized sigmoid kernel**: $K(\mathbf{x}, \mathbf{y}) = \tanh(\gamma \cdot S_C(\mathbf{x}, \mathbf{y}) + c_0)$, where $S_C(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}$ is the cosine similarity.

For every parameter configuration, we evaluate four kernel matrix variations: the original matrix, and three corrected versions: shifted, clipped, and clipped and normalized.

To assess the behaviour of the CPD kernels, we employ the algorithm proposed in Appendix C, which determines if a kernel is CPD by analyzing curvature within the feasible optimization subspace. We apply a numerical tolerance of $\tau = 10^{-10}$ for all spectral analyses.

We perform a grid search over the kernel parameters γ and c_0. For each (γ, c_0) pair, the SVM regularization parameter C is optimized via 5-fold stratified cross-validation on the training set, selecting the C that maximizes the mean F1 score. The final model is retrained on the full training set and evaluated on the held-out test set.

To analyze the impact of curvature direction, we generate γ values over both positive and negative ranges. The complete search space is detailed in Table 2.

For each (γ, c_0), the kernel matrix on the full training set is analyzed. We record the number of negative eigenvalues and the negative mass fraction:

$$\#\text{neg} = \big|\{i : \lambda_i < -\tau\}\big| \qquad \text{negative mass fraction} = \frac{\sum_{\lambda_i<-\tau} |\lambda_i|}{\sum_i |\lambda_i|}$$

To measure the impact of kernel corrections, we define the relative error reduction (Δ_{F1}):

$$\Delta_{\text{F1}} = \frac{\text{err}_{\text{original}} - \text{err}_{\text{corrected}}}{\text{err}_{\text{original}}} = \frac{(1 - \text{F1}_{\text{original}}) - (1 - \text{F1}_{\text{corrected}})}{1 - \text{F1}_{\text{original}}} = \frac{\text{F1}_{\text{corrected}} - \text{F1}_{\text{original}}}{1 - \text{F1}_{\text{original}}}$$

This metric represents the percentage of classification errors eliminated by the correction method. A positive Δ_{F1} indicates performance improvement, while a negative value indicates degradation. In our results, we distinguish between CPD ($\star$), which are theoretically already convex in the SVM optimization space, and non-CPD ($\bullet$) configurations to highlight the relationship between theoretical validity and empirical performance.

Computations were run on ASUS Zenbook 14 (Intel i7, 11th Gen) and two MacBook Air (M1/M2).

5.2 General Kernel SVM Formulation

We consider the Support Vector Machine (SVM) classification problem in its dual representation. The core concept is to map input vectors into a high-dimensional feature space $\mathcal{H}$ via a mapping $\phi : \mathbb{R}^n \to \mathcal{H}$, seeking a linear separating hyperplane in this space.

Given a dataset of l training pairs $(x_i, y_i)_{i=1}^{l}$ where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$, the optimization objective is a quadratic programming (QP) problem:

$$\min_{\alpha} \quad \frac{1}{2}\alpha^\top Q\alpha - e^\top \alpha \text{ subject to} \qquad y^\top \alpha = 0, \qquad 0 \le \alpha_i \le C, \quad i = 1, \dots, l, \tag{1}$$

where $\alpha \in \mathbb{R}^l$ is the vector of Lagrange multipliers, e is a vector of all ones, and $C > 0$ is the regularization parameter. The matrix $Q \in \mathbb{R}^{l \times l}$ has entries $Q_{ij} = y_i y_j K(x_i, x_j)$, where the kernel function K computes the inner product $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$. Standard SVM theory requires K to be PSD to ensure Equation 1 is convex and possesses a unique global minimum.

5.3 Sigmoid Kernel Analysis

In this study, we specifically analyze the Sigmoid Kernel (Hyperbolic Tangent Kernel), historically significant for its connection to neural networks, defined as:

$$K(x_i, x_j) = \tanh(\gamma x_i^\top x_j + c_0), \tag{2}$$

where $\gamma > 0$ is the scaling parameter and c_0 is the shifting parameter. Unlike polynomial kernels which grow unboundedly, or RBF kernels which asymptotically decay to zero, the Sigmoid kernel saturates:

$$\lim_{x_i^T x_j \to \pm\infty} K(x_i, x_j) = \pm 1. \tag{3}$$

Moreover, unlike the RBF kernel, which depends on pairwise distances and is therefore isotropic, the Sigmoid kernel is purely dot-product–based, making it sensitive to vector orientation and global feature scaling rather than Euclidean proximity.

Crucially, the Sigmoid kernel matrix K is generally *indefinite* (not PSD). This violation of Mercer's theorem implies that Q may possess negative eigenvalues, rendering the objective in Equation 1 non-convex and potentially unbounded ($-\infty$).

However, [5] demonstrated that for specific parameters ($\gamma > 0$ and small $c_0 < 0$), the matrix becomes *Conditionally Positive Definite (CPD)*. A matrix K is CPD if:

$$v^\top K v \ge 0 \quad \forall v \in \mathbb{R}^l \setminus \{0\} \quad \text{such that} \quad \sum_{i=1}^{l} v_i = 0. \tag{4}$$

This condition effectively restores convexity because the SVM dual constraints ($y^\top \alpha = 0$) force the optimization to remain exactly within the subspace where the kernel behaves like a PSD matrix.

Despite this theoretical guarantee, an objective of this work is to systematically apply spectral correction methods to both CPD and non-CPD configurations. We aim to investigate whether transforming a CPD kernel, by essentially forcing global Positive Semi-Definiteness, introduces unwarranted distortions or, conversely, offers empirical benefits. This allows us to decouple the effects of the kernel's indefinite nature from the intrinsic impact of the correction algorithms themselves.

Further work exploring the CPD matrices for different training sample sizes is presented in Appendix A.1.

Dual problem for an SVM with indefinite kernel When the kernel matrix Q is indefinite, the SVM dual objective is no longer a convex quadratic: its Hessian is Q and therefore has both positive and negative eigenvalues. This implies that f may have multiple stationary points (local minima, local maxima, and saddle points), and in some parameter regimes the problem can even be unbounded from below so that no global minimizer exists.

In practice, `scikit-learn` uses the LIBSVM library, which implements an SMO-type decomposition method originally designed under the assumption that Q is positive semidefinite. For indefinite kernels, the same algorithm can still be run: the working-set updates are applied and the procedure converges (under mild conditions) to a KKT stationary point of the dual, but this point is not guaranteed to be a global minimum of f. Lin and Lin [5] show that, after suitable modifications, SMO-type methods can be made to converge to stationary points for general symmetric non-PSD kernel matrices, highlighting that with indefinite kernels the optimization problem one actually solves is a non-convex quadratic program whose solution quality depends on the particular stationary point reached by the algorithm.

5.4 Correction Methods

Shifting The shifting correction starts from the eigen-decomposition of the symmetric kernel matrix $K = U\Lambda U^\top$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ may contain negative eigenvalues because of the indefinite sigmoid kernel. In order to make the matrix usable in standard SVM solvers without discarding information, we add a constant $\delta > 0$ to all eigenvalues, choosing $\delta = |\min_i \lambda_i|$, and define $K' = U(\Lambda + \delta I)U^\top = K + \delta I$. Operationally, this is a simple spectral shift that guarantees K' is positive semidefinite (PSD) while preserving the relative ordering and directions of eigenvectors.

Computational complexity: Computing the eigendecomposition requires $O(n^3)$ time and $O(n^2)$ space. Finding $\delta = |\min_i \lambda_i|$ costs $O(n)$. Applying the shift exploits the algebraic property $K' = U(\Lambda + \delta I)U^\top = K + \delta I$, requiring only $O(n)$ time to add δ to the diagonal entries of K without any matrix reconstruction. The total complexity is $O(n^3)$, dominated by eigendecomposition.

Clipping The clipping correction also relies on the eigen-decomposition $K = U\Lambda U^\top$, but instead of shifting the whole spectrum it modifies only the negative part. We construct a clipped spectrum $\Lambda_{\mathrm{clip}} = \mathrm{diag}(\max(\lambda_1, 0), \ldots, \max(\lambda_n, 0))$ and set $K_{\mathrm{clip}} = U\Lambda_{\mathrm{clip}}U^\top$. This guarantees that K_{clip} is positive semidefinite while leaving all eigenvalues that were already non-negative unchanged. Importantly, this operation coincides with the orthogonal projection of K onto the PSD cone with respect to the Frobenius norm,

$$K_{\mathrm{clip}} = \arg\min_{M \succeq 0} \|K - M\|_F,$$

so K_{clip} is, in a precise sense, the closest PSD kernel matrix to the original one. From a practical standpoint, the method consists of a single eigen-decomposition followed by thresholding of the spectrum and reconstruction of K_{clip}, after which training proceeds with a standard SVM algorithm for PSD kernels.

Computational complexity: The eigendecomposition requires $O(n^3)$ time and $O(n^2)$ space. Thresholding the eigenvalues costs $O(n)$. Reconstructing $K_{\mathrm{clip}} = U\Lambda_{\mathrm{clip}}U^\top$ requires computing $U\Lambda_{\mathrm{clip}}$ (column scaling, $O(n^2)$) followed by $(U\Lambda_{\mathrm{clip}})U^\top$ (matrix multiplication, $O(n^3)$), for a total reconstruction cost of $O(n^3)$. The overall complexity is $O(n^3)$, dominated by eigendecomposition and reconstruction.

Normalized Clipping In the normalized clipping variant we first apply spectral clipping to obtain K_{clip} and then perform a diagonal normalization so that all training points have unit self-similarity. After symmetrizing K_{clip}, each entry is rescaled as

$$\widetilde{K}_{ij} = \frac{(K_{\mathrm{clip}})_{ij}}{\sqrt{(K_{\mathrm{clip}})_{ii}\,(K_{\mathrm{clip}})_{jj}}},$$

which yields a matrix $\widetilde{K}$ that remains positive semidefinite and satisfies $\widetilde{K}_{ii} = 1$ for all i. Our goal here is to enforce a unit diagonal in order to turn K_{clip} into a correlation-like matrix and therefore study how a clipped kernel behaves once all self-similarities are fixed to 1. In this way we can disentangle the effect of clipping itself from trivial rescalings of the kernel values and directly compare performance across different parameter settings (γ, c_0) under a common "correlation" normalization.

Computational complexity: Normalized clipping inherits the $O(n^3)$ cost of clipping (eigen-decomposition plus reconstruction). Extracting the diagonal requires $O(n)$ time, and rescaling all n^2 entries according to $\widetilde{K}_{ij} = \frac{(K_{\mathrm{clip}})_{ij}}{\sqrt{(K_{\mathrm{clip}})_{ii}\,(K_{\mathrm{clip}})_{jj}}}$ requires $O(n^2)$ time. Since the cubic cost dominates, the overall complexity remains $O(n^3)$.

6 Analysis of kernel corrections

6.1 Spectral properties across the parameter space

We selected the parameter space (γ, c_0) by analyzing the dot product distribution in order to avoid saturation of the sigmoid kernel to $\{-1, +1\}$. The three heatmaps in Figure 1 provide complementary perspectives on the eigenvalue distribution of the kernel Gram matrix, revealing a clear correspondence with theoretical predictions from Lin and Lin [5]. The distribution of inner products $\langle \phi(x_i), \phi(x_j) \rangle = \tanh(\gamma \langle x_i, x_j \rangle + c_0)$ determines these spectral properties, and our parameter selection strategy successfully identifies regions where the kernel matrix exhibits CPD behavior.

Heatmap 1 displays E_{neg}, the ratio of negative mass to total mass. This metric quantifies the relative contribution of negative eigenvalues to the spectral norm. The starred regions, concentrated in the quadrant where $\gamma > 0$ and $c_0 < 0$, exhibit values approaching zero (yellow), indicating minimal negative mass. This aligns precisely with Theorem 4 from Lin and Lin [5], which establishes that for $\gamma > 0$ (corresponding to $a > 0$ in their notation) and sufficiently negative c_0 (corresponding to $r < 0$), the kernel matrix becomes conditionally positive definite (CPD).

Heatmap 2 shows f_{neg}, the fraction of eigenvalues that are negative. The dark region (low fraction) in the parameter space where $\gamma \approx 0.001$–0.01 and $c_0 < -0.1$ demonstrates that CPD matrices dominate this region. Theoretically, a CPD matrix can have at most one negative eigenvalue (Theorem 3), which would manifest as $f_{\text{neg}} \approx 1/n$ for an $n \times n$ matrix.

Heatmap 3 presents N_{neg}, the absolute count of negative eigenvalues. This heatmap provides the same information as Heatmap 2 in absolute terms rather than as a fraction, and serves as a control to verify the consistency of our spectral analysis.

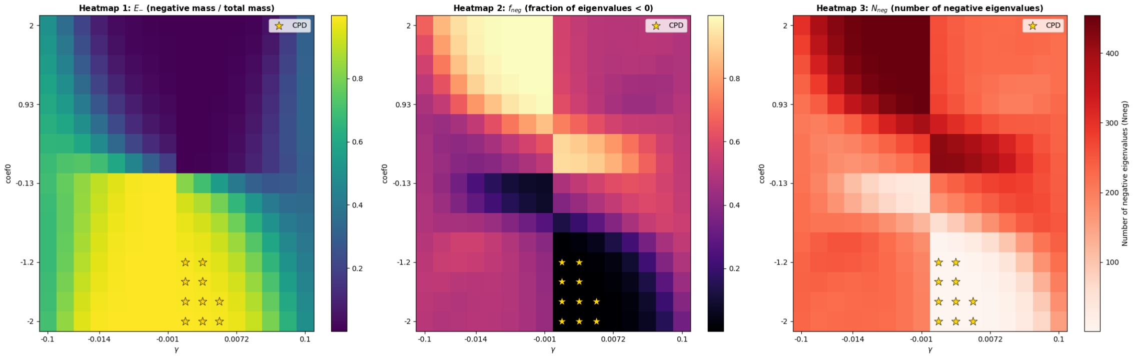

Figure 1: Spectral properties of the sigmoid kernel matrix across the parameter space (γ, c_0). **Left:** E_{neg}, ratio of negative to total eigenvalue mass. **Center:** f_{neg}, fraction of negative eigenvalues. **Right:** N_{neg}, count of negative eigenvalues. Stars indicate the values that correspond to CPD matrices.

6.2 Impact of correction methods

Figure 2 reveals that the vast majority of corrections yield negligible performance changes, with distributions heavily concentrated around zero for all three methods. However, the tails of these distributions tell us something: shift and clipnorm exhibit numerous cases of severe performance degradation, with Δ_{F1} values plunging below -1000%. This indicates that kernels which originally contributed meaningfully to classification are completely disrupted by the correction process: their geometric structure is so fundamentally altered that the SVM can no longer exploit them effectively.

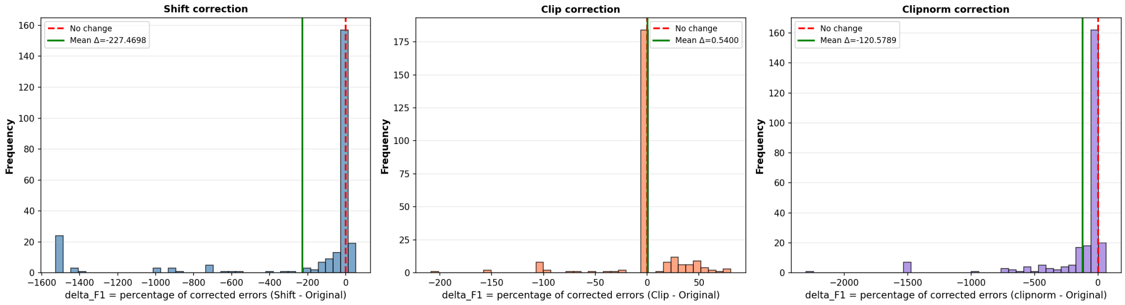

Figure 2: Distribution of relative error reduction (Δ_{F1}) across all parameter combinations for three correction methods for the standard sigmoid kernel. Negative values indicate performance degradation.

The clip method stands out as notably different. While it also shows degradations, these are substantially more contained compared to shift and clipnorm. Moreover, clip is the method that produces the largest positive improvements, in few cases reaching +80%. This suggests that clipping negative eigenvalues, rather than shifting or normalizing the entire spectrum, is the least destructive intervention when a correction is necessary.

Note: In the following analyses and plots, we apply a mask to exclude parameter combinations where $\Delta_{\mathrm{F1}} = 0$ exactly. This removes cases where the correction had literally no effect and allows us to focus on instances where corrections altered performance, whether positively or negatively.

Figure 3 decomposes correction efficacy by spectral properties: the fraction of negative eigenvalues (x-axis) versus the negative mass fraction (y-axis). A notable characteristic across all three panels is that most points lie approximately along the diagonal, indicating that when a kernel has few negative eigenvalues, those eigenvalues also tend to carry significant mass, and vice versa. This diagonal concentration may reflect a spectral property of non-PSD sigmoid kernels in our parameter grid.

The shift correction (left panel) exhibits a clear pattern tied to spectral structure. The most significant improvements (green circles) occur in the high-eigenvalue-count, low-mass regime, so kernels with many negative eigenvalues but individually small magnitudes. This behavior is intuitive: shifting the spectrum by a constant value $\lambda_{\min}$ has minimal distortion when $|\lambda_{\min}|$ is small, preserving the kernel's discriminative geometry while ensuring positive semidefiniteness. Conversely, the most severe degradations (red circles, including CPD stars) concentrate in the low-eigenvalue-count, high-mass regime, where a few negative eigenvalues dominate the spectrum. In these cases, the shift magnitude is large, drastically altering the kernel matrix. So, forcing a large shift onto an already-effective kernel destroys its structure, leading to significant performance loss.

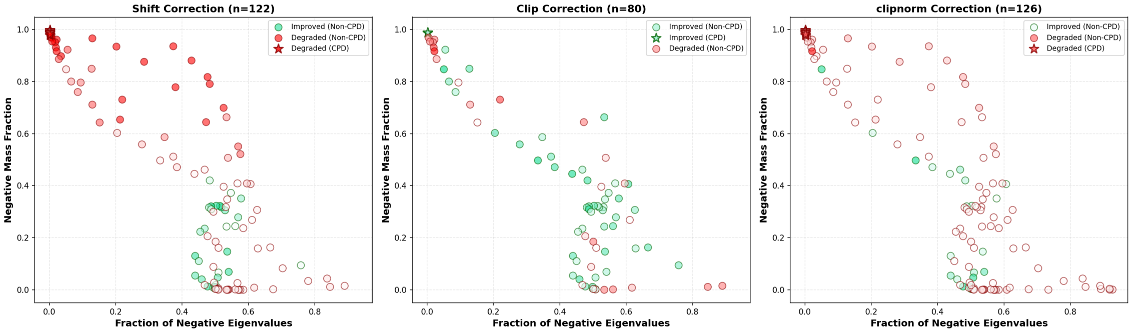

Figure 3: Spectral properties versus correction efficacy for the standard sigmoid kernel. Green/red circles indicate improved/degraded non-CPD kernels; stars mark CPD kernels. Sample sizes: shift ($n = 122$), clip ($n = 80$), clipnorm ($n = 126$).

The clip method (center panel) shows a similar but substantially weaker pattern. While improvements still favor the high-count, low-mass region, the distinction is far less pronounced. More interestingly, some kernels in the low-count, high-mass regime, including CPD kernels, experience improvements (green stars and circles in the top left part of the plot). This behavior aligns with the clipping strategy: rather than shifting the entire spectrum, clipping selectively zeroes out a few pathological negative directions while leaving the rest of the kernel intact. For kernels that are already well-performing (e.g., CPD or near-CPD), removing a small number of residual negative eigenvalues can provide gains without destabilizing the overall geometry. This explains why clip is the least destructive method and occasionally improves already-good kernels, when there are few negative eigenvalues. Overall, anyway, the majority of the CPD matrices do not appear in the plot, meaning that the performance of those matrices is usually kept the same.

The clipnorm method (right panel) reverts to a shift-like pattern: improvements are sparse and confined to the high-count, low-mass region, while degradations dominate elsewhere. This suggests that the normalization step following the clipping is often counterproductive. While the initial clipping operation might correctly remove pathological directions, the subsequent rescaling to unit trace distorts the kernel's scale in ways that disrupt the SVM's optimization landscape. Many kernels that were successfuly "repaired" by clipping are subsequently "broken" by normalization, resulting in performance degradation comparable to or worse than shift. This observation may underscore the importance of preserving the kernel's intrinsic scale: even after correcting spectral violations, rescaling can undo the benefits of correction.

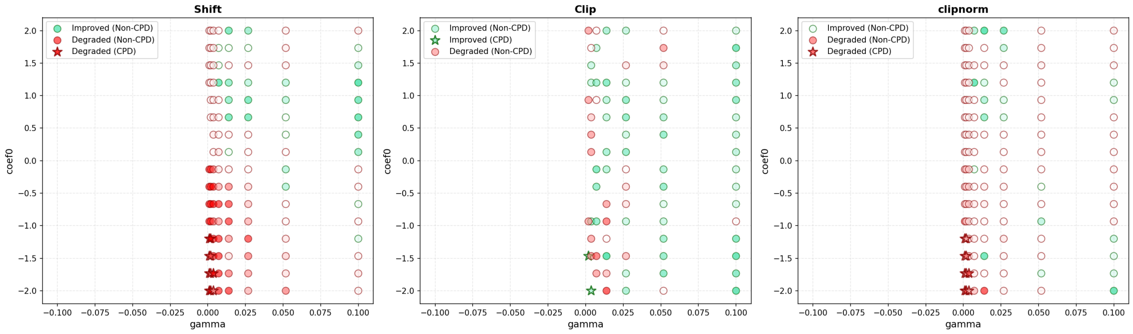

Figure 4: Spatial distribution of correction efficacy across the standard sigmoid kernel parameter space (γ, c_0).Color indicates improvement (green), degradation (red), or CPD status (stars).

Figure 4 reveals the spatial manifestation of patterns established in previous analyses. The concentration of shift degradations in the low-γ, negative-c_0 region directly reflects the spectral regime identified in Figure 1 (few but large negative eigenvalues) and the behavior in Figure 3 (well-performing kernels destroyed by correction). Clip's broader improvement distribution aligns with its efficacy in moderate-negative-fraction regimes (Figure 1), while clipnorm's shift-like spatial pattern confirms that normalization reintroduces the problems observed spectroscopically. These spatial patterns are thus the direct consequence of the interplay between parameter choices and spectral structure. It is very interesting to note that the whole region $\gamma < 0$ results in exactly no change under any of the transformations, that is why those points were not considered, as stsated before.

6.3 Impact of corrections on original kernel performance

Figure 5 examines the relationship between original (uncorrected) F1 score and the performance change induced by each correction method. We retained only non-zero changes ($\Delta_{\text{F1}} \neq 0$) to focus on cases where corrections meaningfully alter performance. The three methods exhibit markedly different behaviors across the performance spectrum.

Shift method (left column) produces severe degradations for many kernels, with Δ_{F1} dropping below -1000% in numerous cases. The vast majority of kernels suffer severe performance collapse. This reveals that when a non-PSD sigmoid kernel achieves strong classification performance (baseline F1 > 0.90), its negative eigenvalues are likely not pathological artifacts but features of its geometry. The shift correction, by enforcing positive semi-definiteness through eigenvalue translation, destroys the very spectral properties that enabled effective classification.

Clip method (center column) shows more moderate degradations and fewer severe failures. The zoomed view reveals an intriguing structure: points trace a curved trajectory as baseline F1 varies, suggesting that the correction's impact depends non-trivially on the kernel's initial spectral properties. Crucially, several non-CPD kernels achieve modest improvements ($\Delta_{\text{F1}} > 0$), indicating that clipping can occasionally enhance performance when the original kernel matrix contains pathological but correctable structure.

Clipnorm method (right column) exhibits behavior intermediate between shift and clip. While it produces severe degradations for some kernels, the pattern is less extreme than shift. The zoomed view shows that visible CPD kernels (stars) are few, as most experience negligible changes and are filtered out. Non-CPD kernels display more variable responses, with the normalization step appearing to preserve some useful information even after clipping, though not consistently.

The critical observation is that shift and clipnorm corrections consistently degrade high-performing kernels. Kernels achieving F1 > 0.90 rarely benefit from correction; instead, they predominantly show substantial performance losses. This suggests that high-performing kernels, whether strictly CPD or possessing favorable spectral properties despite being non-PSD, already encode effective discriminative geometry. Attempting to enforce Mercer compliance by modifying their eigenvalue structure disrupts this geometry, resulting in degraded classification performance. The only correction method that seems to enhance or at least preserve the classification power is clipping, which often results in performance improvements or negligible changes, with the advantage of obtaining a convex optimization surface.

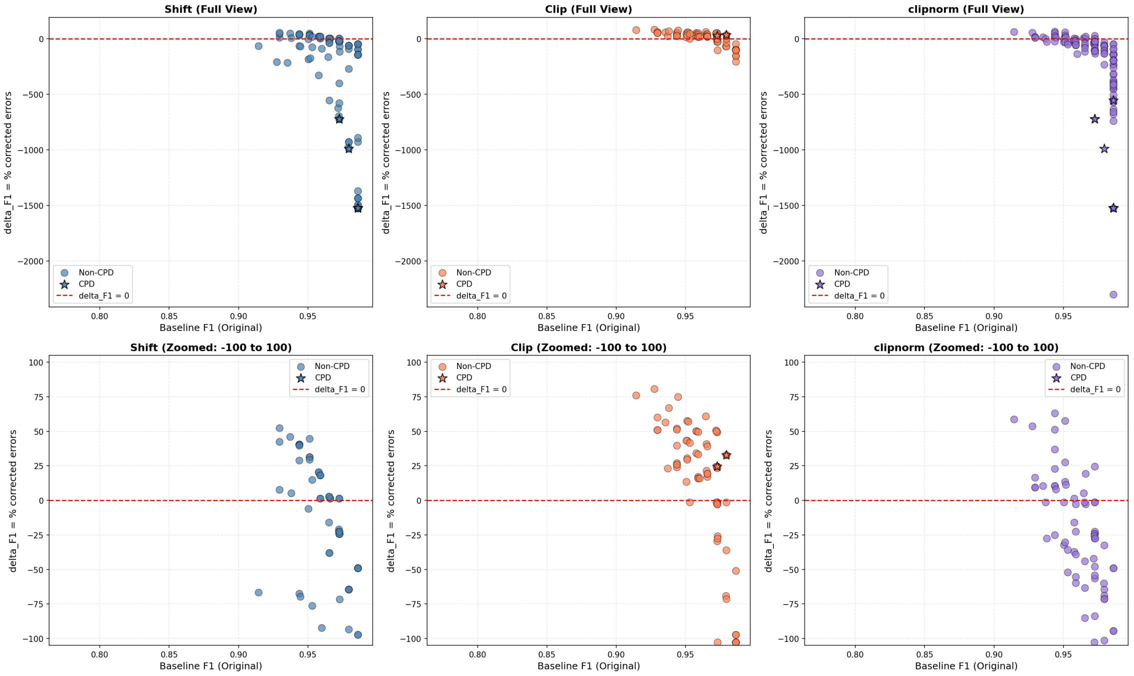

Figure 5: Performance change Δ_{F1} versus original F1 score for shift, clip, and clipnorm corrections with the standard sigmoid kernel. Top: full view. Bottom: zoomed to $|\Delta_{\mathrm{F1}}| \leq 100$. Stars indicate CPD kernels, circles indicate non-CPD kernels. Dashed line marks $\Delta_{\mathrm{F1}} = 0$.

6.4 Normalized Sigmoid Kernel

To isolate angular relationships from magnitude effects, we evaluate a normalized variant using cosine similarity:

$$K_{\mathrm{norm}}(x_i, x_j) = \tanh\left(\gamma \cdot \frac{\langle x_i, x_j \rangle}{\|x_i\| \|x_j\|} + c_0\right), \tag{5}$$

where γ now directly controls sensitivity to angular alignment rather than being confounded with vector norms, avoiding the saturation of the Gram matrix to $\{-1, +1\}$. The complete parameter grid can be found in Table 2.

Figure 6 shows that normalization reduces extreme degradations but does not fundamentally improve correction efficacy. Comparing with Figure 2, the severe left tail decay is shortened for all methods, yet mean performance remains similar or slightly worse. Clip remains the least destructive method, with its distribution concentrated near zero.

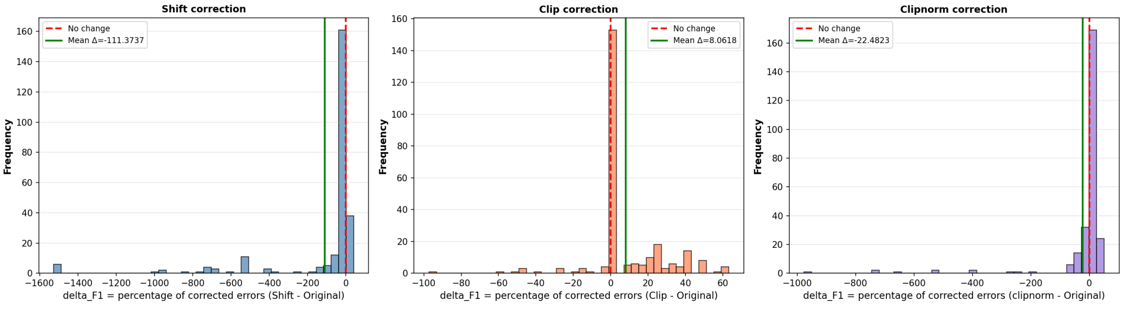

Figure 6: Distribution of relative error reduction (Δ_{F1}) across all parameter combinations for three correction methods for the normalized sigmoid kernel. Negative values indicate performance degradation.

Figure 7 confirms that normalization preserves the fundamental spectral-performance relationships. The diagonal concentration (negative eigenvalue count correlates with negative mass), shift's degradation of low-count/high-mass kernels, clip's superior behavior, and clipnorm's pathology all persist unchanged. This demonstrates that these patterns arise from intrinsic sigmoid kernel geometry, not input scaling artifacts.

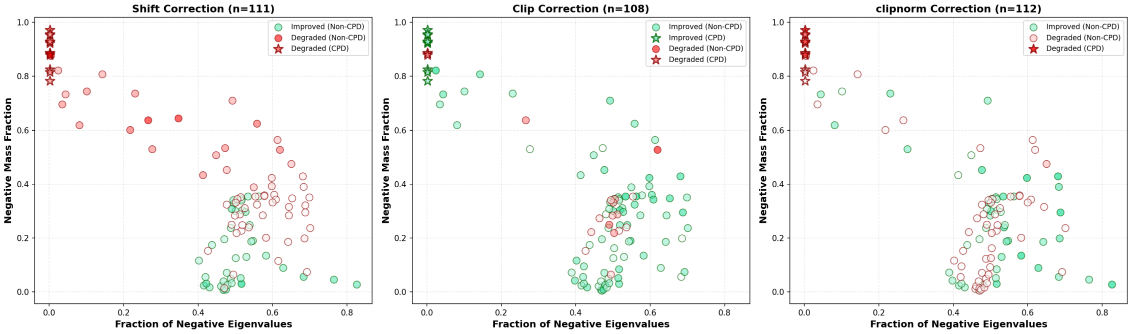

Figure 7: Spectral properties versus correction efficacy for the normalized sigmoid kernel. Green/red circles indicate improved/degraded non-CPD kernels; stars mark CPD kernels. Sample sizes: shift ($n = 94$), clip ($n = 89$), clipnorm ($n = 109$).

The spatial distribution (Figure 8) and baseline-versus-correction relationships (Figure 9) mirror their standard counterparts: shift degrades CPD regions, clip is least destructive, clipnorm inherits shift pathology, and high-performing kernels (F1 > 0.90) consistently degrade under shifting and normalized clipping but not in standard clipping. The only notable difference is slightly sharper boundaries in parameter space, likely due to the constrained input range $[-1, 1]$ eliminating magnitude-dependent variations.

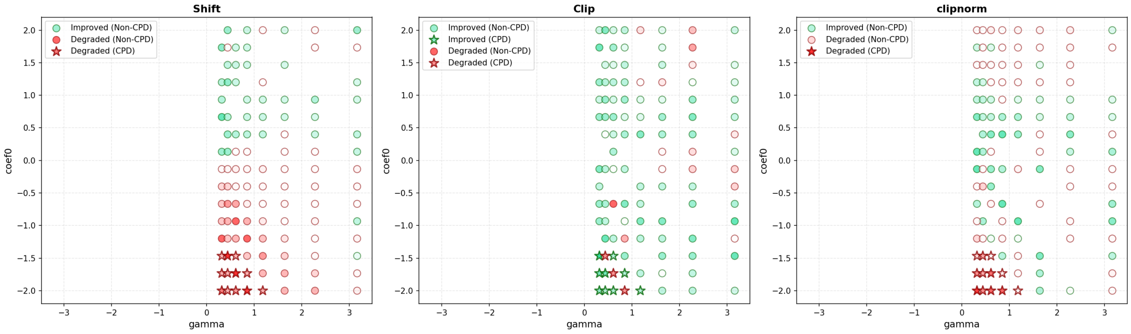

Figure 8: Spatial distribution of correction efficacy across the normalized sigmoid kernel parameter space (γ, c_0). Color indicates improvement (green), degradation (red), or CPD status (stars).

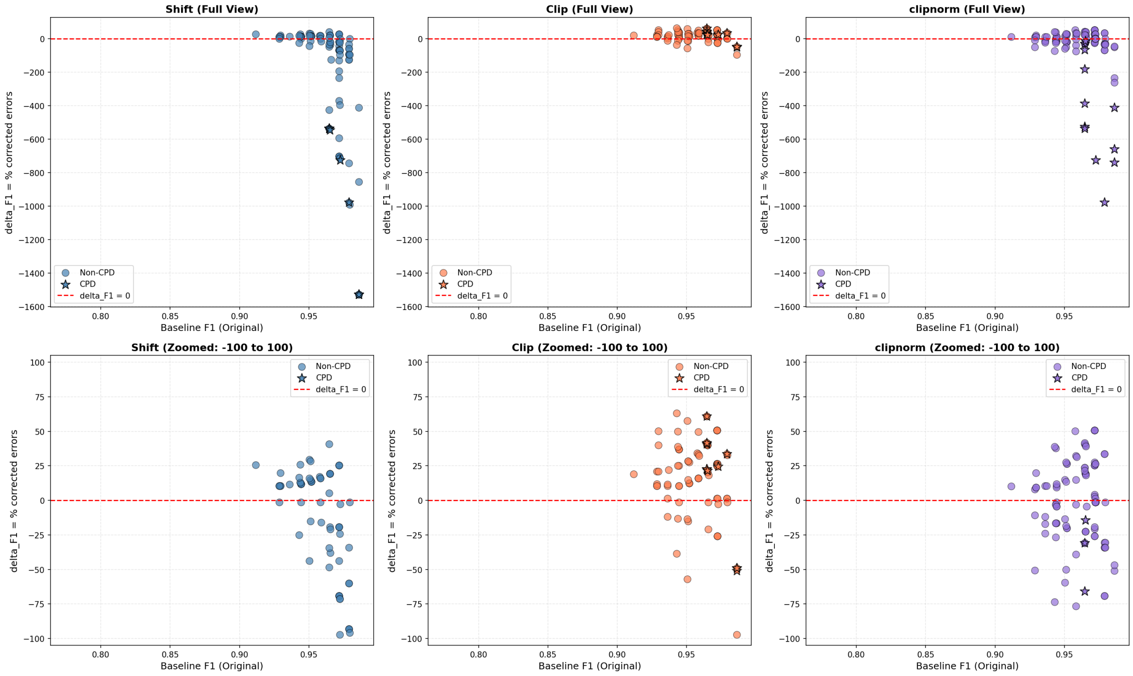

Figure 9: Performance change Δ_{F1} versus original F1 score for shift, clip, and clipnorm corrections with the normalized sigmoid kernel. Top: full view. Bottom: zoomed to $|\Delta_{\mathrm{F1}}| \leq 100$. Stars indicate CPD kernels, circles indicate non-CPD kernels. Dashed line marks $\Delta_{\mathrm{F1}} = 0$.

Normalization of the dot product provides only marginal improvements, mainly by reducing severe failures. Still, it does not resolve the central limitation of indefinite sigmoid kernels, namely the lack of reliably improved discriminative power. The fact that all major empirical phenomena replicate across both the standard and normalized variants (including spectral-performance correlations, high-F1 regime degradation, the superiority of clipping, and the clipnorm pathology) suggests that these effects are driven primarily by the underlying geometry of the sigmoid mapping rather than by feature scaling. The only significant difference with respect to the standard kernel is the pattern in the clipping and shifting plots in Fig 9 which are not showing a clear curved pattern as they were in the standard kernel setting. In this sense, the normalized formulation can be seen as a robustness check: it tests whether the observed behaviors persist under a controlled change in similarity scaling, without fixing the kernel's indefiniteness or its downstream consequences for SVM optimization and generalization.

7 Conclusions

This work explored the behaviour of the sigmoid kernel through its application within an SVM on a binary prediction task. We analyzed different parameter combinations and investigated how simple spectral fixes (shifting, clipping, and normalized clipping) affect the kernel's behaviour.

Overall, most parameter settings showed little improvement after correction, while the tails of the distributions indicated that shifting and normalized clipping could, in some configurations, substantially reduce performance. Among the considered fixes, eigenvalue clipping was typically the least disruptive and was also the method that most often produced the largest positive changes when improvements occurred.

From a theoretical perspective, our experiments are consistent with the fact that sigmoid kernels can be CPD only in specific parameter regimes, and that non-PSD kernels can lead to non-convex dual objectives in standard SVM formulations. Empirically, the relationship between spectral properties and performance appeared non-trivial: the degree and the "shape" of indefiniteness (e.g., a few large negative eigenvalues versus many small ones) seemed to influence how corrections affected the resulting classifier.

The normalized sigmoid kernel did not show any substantial differences from the classic version, suggesting that the observed phenomena may not be affected by input scaling.

In conclusion, the sigmoid kernel exhibited complex behaviour and introduced practical difficulties when aiming for strong performance. Moreover, analyzing its spectrum and applying correction methods can become computationally expensive relative to the potential improvements, making the RBF kernel a more suitable choice in our setting.

7.1 Limitations

This study is limited by its scope (single dataset, one split/protocol, and computationally expensive full spectral corrections), so the observations should not be interpreted as universal claims about sigmoid kernels or kernel repair methods, as this is an exploratory analysis. Future work should validate these patterns across datasets and repeated resampling, and compare against other optimization algorithms designed to work directly with indefinite kernels against the enforcing of PSDness as a preprocessing step.

8 Declarations on GenAI

The authors would like to acknowledge the use of ChatGPT [8] for assistance with spell checking, debugging, and generating well-formatted plots. The tool was used solely to improve clarity and presentation; all analysis and results were performed independently by the authors.

References

[1] Dheeru Dua and Casey Graff. Uci machine learning repository, 2017.

[2] Suicheng Gu and Yuhong Guo. Learning svm classifiers with indefinite kernels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 942–948, 2012.

[3] Nicholas J Higham. Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343, 2002.

[4] Mingqing Hu, Yiqiang Chen, and James Tin-Yau Kwok. Building sparse multiple-kernel svm classifiers. *IEEE Transactions on Neural Networks*, 20(5):827–839, 2009.

[5] Hsuan-Tien Lin and Chih-Jen Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. *Neural Comput*, 3(1-32):16, 2003.

[6] Ronny Luss and Alexandre d'Aspremont. Support vector machine classification with indefinite kernels. *Advances in neural information processing systems*, 20, 2007.

[7] Dino Oglic and Thomas Gärtner. Learning in reproducing kernel kreın spaces. In *International conference on machine learning*, pages 3859–3867. PMLR, 2018.

[8] OpenAI. Chatgpt. [Large language model] https://chat.openai.com/, 2025.

[9] Elzbieta Pekalska, Pavel Paclik, and Robert PW Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of machine learning research*, 2(Dec):175–211, 2001.

[10] William H Wolberg, W Nick Street, and Olvi L Mangasarian. Breast cancer wisconsin (diagnostic) dataset. *UCI Machine Learning Repository*, 1992.

A Additional Results

A.1 Further Studies on CPD Matrices with different training sizes

To assess how stable the sigmoid kernel behavior is with respect to the amount of training data, four additional runs were first carried out by varying the training-set size (20%, 40%, 60%, and 80%), as an exploratory study on subsets. The 80/20 split corresponds to the default configuration used throughout the rest of the project. During this analysis, an interesting fact about the fraction of CPD matrices emerged. As shown in Fig. 10, the observed fraction of hyperparameter configurations yielding a CPD Gram matrix decreases as the training size increases. Importantly, this “percentage” is an *observed fraction* over the tested (γ, c_0) grid for a given split: for each fixed split and hyperparameter pair, the CPD check is deterministic, but the fraction of CPD outcomes can change when the sampled training set changes. Therefore, results obtained on smaller training subsets may not generalize to the 80/20 setting used in the project, and can make CPD appear more prevalent than it is under the final regime.

To further characterize this effect, Fig. 11 reports heatmaps of the difference between the 80% and 20% configurations for three spectral indicators: the negative-mass fraction, the fraction of negative eigenvalues, and the absolute number of negative eigenvalues, each evaluated over the explored (γ, c_0) grid. Regions with positive differences correspond to increased spectral indefiniteness when moving from 20% to 80% training data, while negative differences indicate the opposite trend.

This observation leads us to hypothesize a probabilistic dependency where larger datasets might be less likely to sustain the CPD property. While a formal proof is beyond the scope of this work, the data suggests that increasing the sample size introduces additional constraints that make preserving the CPD nature of the kernel progressively more difficult.

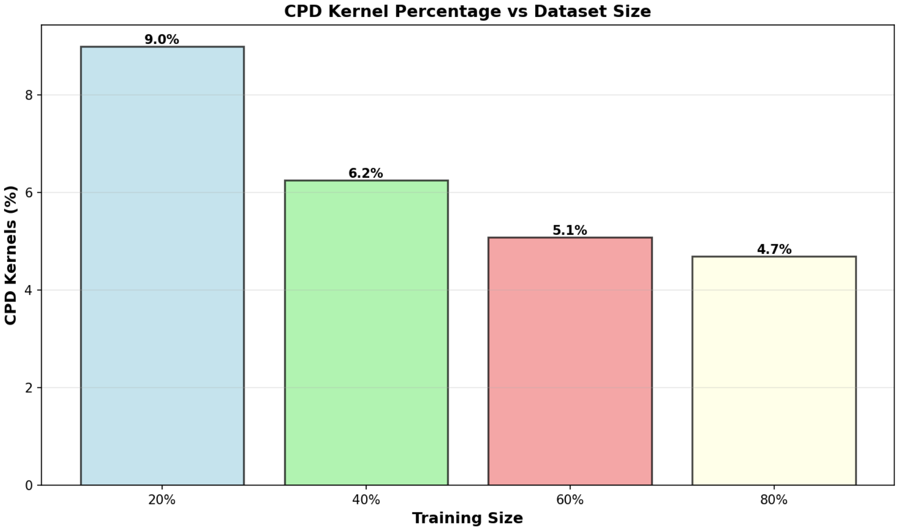

Figure 10: CPD kernel percentage as a function of training-set size.

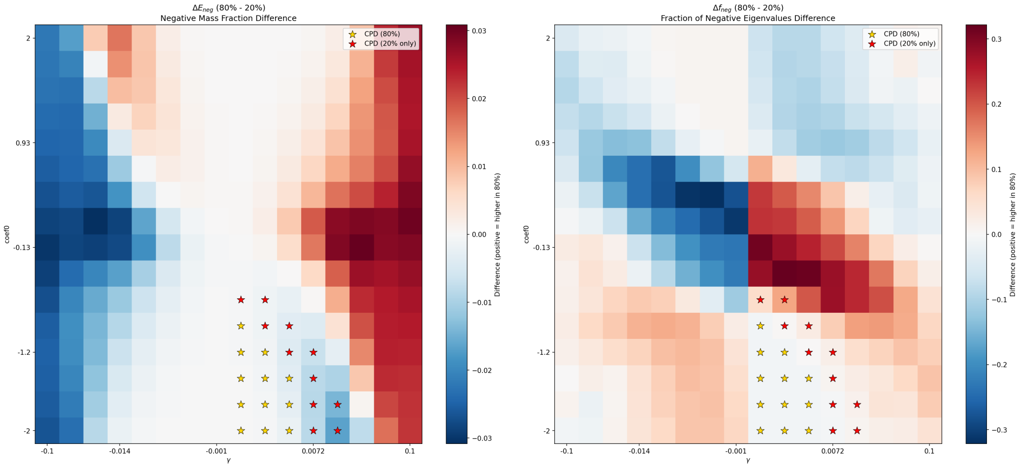

Figure 11: Difference in spectral indefiniteness indicators between the 80% and 20% training configurations over the (γ, c_0) grid: negative-mass fraction (left), fraction of negative eigenvalues (right). Markers indicate CPD configurations under each split.

B Notation

Table 1: Model abbreviations used throughout the text

Full text	**Abbreviation**
Support Vector Machine	SVM
Positive Semi Definite	PSD
Sequential Minimal Optimization	SMO
Conditionally Positive Definite	CPD
Radial Basis Function	RBF
Explorative Data Analysis	EDA
Cross-Validation	CV

C CPD Matrices Algorithm

Algorithm 1 Check if a matrix K is Conditionally Positive Definite (CPD)

```
Require: Matrix $K \in \mathbb{R}^{n \times n}$, tolerance tol
Ensure: True if $K$ is CPD, otherwise False
1: $n \leftarrow$ number of rows of $K$
2: $V \leftarrow$ orthonormal basis for the null space of $e^\top$ where $V^\top e = 0$
3: $K_{\text{proj}} \leftarrow V^\top K V$
4: $\lambda_{\min} \leftarrow$ smallest eigenvalue of $K_{\text{proj}}$
5: if $\lambda_{\min} >$ tol then
6:     return True
7: else
8:     return False
9: end if
```

First, it defines the constraint vector $\mathbf{e}$ (Step 1) and constructs an orthonormal basis $\mathbf{V}$ for the null space of $\mathbf{e}^\top$ (Step 2), effectively capturing all directions where $\sum v_i = 0$. In Step 3, the kernel matrix is projected onto this subspace via the transformation $\tilde{\mathbf{K}} = \mathbf{V}^\top \mathbf{K} \mathbf{V}$. Finally, the algorithm computes the eigenvalues of the reduced matrix $\tilde{\mathbf{K}}$ (Step 4). If the smallest eigenvalue $\lambda_{\min}$ is non-negative (within a numerical tolerance ϵ), the matrix is confirmed to be Conditionally Positive Definite (CPD).

D Implementation Details

Experiments are implemented in Python 3 using `scikit-learn` and `NumPy`. Given the computational intensity (256 combinations × 4 kernel variants × 10 C values), we utilize `joblib` for parallel processing ($n\_jobs = -1$). Eigenvalue computations utilize `np.linalg.eigvalsh` for numerical stability, with values $\lambda_i < -\tau$ treated as strictly negative.

D.1 Search Spaces

Table 2: Hyperparameter search spaces. γ values include both positive and negative ranges to explore different curvature regimes.

Parameter	**Kernel Type**	**Search Space**	**Grid Size**
γ	Standard	$\pm[10^{-3}, 10^{-1}]$ (log-spaced)	16
γ	Normalized	$\pm[10^{-0.5}, 10^{0.5}]$ (log-spaced)	16
c_0	Both	$[-2, 2]$ (linear-spaced)	16
C	Both	$[10^{-3}, 10^{3}]$ (log-spaced)	10
Total combinations per experiment			**256**