QuanDA: Quantile-Based Discriminant Analysis for High-Dimensional Imbalanced Classification

Qian Tang
School of Statistics
University of Minnesota
Minneapolis, MN, 55455
tang1015@umn.edu

Yuwen Gu
Department of Statistics
University of Connecticut
Storrs, Connecticut, 06269
yuwen.gu@uconn.edu

Boxiang Wang*
Department of Statistics and Actuarial Science
University of Iowa
Iowa City, IA, 52246
boxiang-wang@uiowa.edu

Abstract

Binary classification with imbalanced classes is a common and fundamental task, where standard machine learning methods often struggle to provide reliable predictive performance. Although numerous approaches have been proposed to address this issue, classification in low-sample-size and high-dimensional settings still remains particularly challenging. The abundance of noisy features in high-dimensional data limits the effectiveness of classical methods due to overfitting, and the minority class is even difficult to detect because of its severe underrepresentation with low sample size. To address this challenge, we introduce Quantile-based Discriminant Analysis (QuanDA), which builds upon a novel connection with quantile regression and naturally accounts for class imbalance through appropriately chosen quantile levels. We provide comprehensive theoretical analysis to validate QuanDA in ultra-high dimensional settings. Through extensive simulation studies and high-dimensional benchmark data analysis, we demonstrate that QuanDA overall outperforms existing classification methods for imbalanced data, including cost-sensitive large-margin classifiers, random forests, and SMOTE.

1 Introduction

High-dimensional binary classification is a fundamental yet challenging machine learning task, particularly in problems where sample sizes are small and class distributions are heavily imbalanced. The situation commonly arises in many application fields, such as disease diagnostics (Krawczyk et al., 2016; Bae and Yoon, 2015; Azari et al., 2015), where data acquisition is costly due to the involvement of human and animal experiments in clinical studies (Evans and Ildstad, 2001). As a result, the number of features often far exceeds the number of data points, leading to the so-called high-dimensional-low-sample-size (HDLSS) problem (Hall et al., 2005; Aoshima et al., 2018). Besides data scarcity, class imbalance further complicates the challenge: the positive class, such as disease occurrence, is typically much rarer than the negative one. Similar challenges are prevalent in many other applications such as image detection (Kubat et al., 1997), cybersecurity (Cieslak et al., 2006), fraud detection (Wei et al., 2013; Sanz et al., 2014), text categorization (Wu et al., 2014), and fault diagnostics (Wu et al., 2018; Zhu and Song, 2010; Santos et al., 2018).

*The corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

When data are highly imbalanced, identifying the minority class becomes no easier than finding a needle in a haystack. Standard machine learning methods often become ineffective, as they tend to bias heavily toward the majority class and struggle to capture decisive patterns in the minority class. This bias can lead to poor generalization and unreliable prediction, and the challenge is further exacerbated in the HDLSS setting, where the minority class is severely underrepresented with the limited sample size compared to the data dimension. To address this challenging issue, many methods have been proposed in the literature to adjust for the class imbalance. Generally speaking, those methods fall mainly into three categories: (1) data-level adjustment, (2) algorithm-level adjustment, and (3) a combination of the first two categories. Comprehensive reviews include, but are not limited to Weiss (2004); He and Garcia (2009); Fernández (2018); Feng et al. (2021); Rezvani and Wang (2023).

Two straightforward methods falling within the category of data-level adjustment are oversampling and undersampling, where the imbalance is mitigated by randomly duplicating samples from the minority class or eliminating samples from the majority class (Paula et al., 2015). However, in the HDLSS setting, neither approach is suitable. Undersampling can yield an excessively small data set, as too many majority class samples must be discarded to match the minority class size. On the other hand, oversampling often leads to model overfitting, as individual data points may be replicated too many times, thereby distorting the classification boundary (Devi et al., 2020). Rather than replicating the minority class, an alternative strategy is to generate synthetic data to decrease the risk of overfitting. One well-known example of this strategy is the so-called Synthetic Minority Over-sampling TEchnique (SMOTE, Chawla et al., 2002). Many of its variants have also been developed in the literature, such as FSMOTE (Gosain and Sardana, 2019), MSMOTE (Hu et al., 2009), SMOTE-ENN (Muntasir Nishat et al., 2022), and SMOTE-RSB (Ramentol et al., 2012), among others. SMOTE can be further integrated with ensemble learning techniques, such as SMOTEBoost (Chawla et al., 2003) and WSMOTE (Abedin et al., 2023). However, SMOTE has been shown to not perform well in the HDLSS setting due to a strong bias toward the minority class (Blagus and Lusa, 2013).

For algorithm-level adjustment, a widely adopted framework is cost-sensitive learning, which assigns a higher cost for data points that are misclassified in the minority class. This approach is commonly employed in large-margin classifiers, such as the support vector machines (SVMs, Cortes and Vapnik, 1995; Vapnik, 1995). Cost-sensitive SVMs (Lin et al., 2002; Zeng and Zhang, 2023) place a different weight on each data point in empirical hinge loss, and the resulting classifiers are known to be Fisher consistent in terms of cost-sensitive Bayes risk (Lin, 2002, 2004). In the literature, most studies on cost-sensitive large-margin classifiers focus on standard classifiers or kernel machines, for example, Zhang et al. (2016); Shin et al. (2017); Fu et al. (2018), while their performance on the HDLSS setting may not be reliable due to the so-called data pilling issue (Marron et al., 2007; Wang and Zou, 2018). Moreover, the introduction of varying weights also affects the efficiency of their optimization algorithms. In addition to large-margin classifiers, ensemble learning, such as random forest and boosting, can also deal with imbalanced data through algorithm-level adjustments (Chen et al., 2004; Khalilia et al., 2011; Galar et al., 2011; Sağlam and Cengiz, 2022). Adjustments for imbalanced classification at both the data and algorithm levels have been extended to the deep learning framework as well, a survey of which can be found in Johnson and Khoshgoftaar (2019). However, deep learning-based approaches typically require a sufficiently large sample size, which is not the case in the HDLSS setting. Consequently, deep learning is not applicable in this context.

Despite extensive research on imbalanced classification, a critical gap remains: few approaches perform well when the sample size is extremely small and the number of features is extremely large. In the HDLSS setting, overly complex models are prone to overfitting, whereas an underfit model may fail to handle class imbalance and high dimensionality simultaneously. To this end, it is essential to develop a direct yet effective classification method that can handle these challenges simultaneously without relying on too much data.

In this paper, we propose Quantile-based Discriminant Analysis (QuanDA) for imbalanced classification in the HDLSS setting. This method is based on quantile regression, which is widely used in statistics and econometrics to estimate the conditional quantiles of a continuous response given a set of features, providing a comprehensive view of the underlying distribution. The theoretical properties of quantile regression have been extensively studied (Belloni and Chernozhukov, 2011; Wang et al., 2012; Zheng et al., 2015) in the context of high-dimensional regression. Although it may seem counterintuitive to apply quantile regression directly to a classification problem, we show

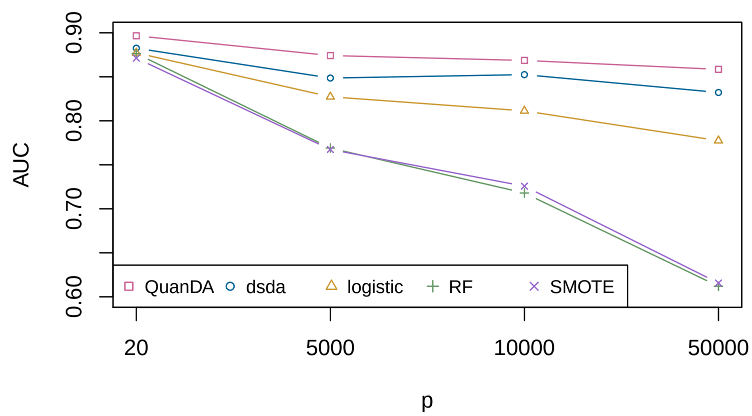

Figure 1: AUC scores comparing QuanDA with direct sparse discriminant analysis (dsda), logistic regression, random forest (RF), and SMOTE on simulated data with $\mathbf{\Sigma} = \mathrm{AR}_5$, $n = 400$ and class imbalance $\pi_0 = 0.9$. See details on data generation in Section 4. The x-axis is plotted with evenly spaced positions for clarity, and we label it with the original values of p to preserve interpretability.

that the population minimizer of the cost-sensitive Bayes risk can be obtained from a conditional quantile of the binary class label. Building upon this connection, we propose QuanDA which directly fits quantile regression to the label for imbalanced classification. This idea echoes the established connection between least-squares regression and linear discriminant analysis (Hastie et al., 2009; Mai et al., 2012), which is studied mainly for balanced classification. A pivotal component that enables QuanDA to perform effectively in the HDLSS setting is the jittering step, which introduces random noise to discrete class labels. This perturbation helps stabilize the algorithm and enhances its robustness when the sample size is limited. While a conceptually related idea was explored by Papandreou and Yuille (2011),the underlying motivations and implementations differ substantially. In their work, stochastic perturbations were applied to continuous energy potentials to induce a discrete label random field derived from an energy-based formulation. We also impose a sparse penalty, such as the lasso (Tibshirani, 1996), to automatically discard irrelevant features.

To give a quick demonstration of our proposed method QuanDA in the HDLSS setting, Figure 1 presents AUC scores that compare QuanDA with several popular classifiers. When the number of features p is small, all the methods achieve satisfactory AUC scores. However, as p increases, QuanDA, dsda, and logistic regression, which are specifically designed for high-dimensional data, exhibit slightly reduced performance but remain relatively stable. In contrast, RF and SMOTE struggle with high-dimensional data, and their performance deteriorates significantly. Overall, QuanDA consistently delivers the highest AUC scores and robustness to increasing dimensionality.

There are several notable features of QuanDA. First, by fitting quantile regression, we can select an appropriate quantile level to account for class imbalance, which is intuitive and straightforward. Second, QuanDA is flexible enough to incorporate various sparse penalties, for example, the group lasso (Yuan and Lin, 2006) or fused lasso (Tibshirani et al., 2005), to handle grouped or spatially structured features. Third, by formulating the problem into a standard quantile regression, QuanDA can be directly fitted using off-the-shelf algorithms for sparse quantile regression, such as hdqr (Tang et al., 2024) and fhdqr (Gu et al., 2018), which eliminates the need to design a specialized solver. We demonstrate through extensive simulations and benchmark data analyzes that QuanDA is highly competitive with the representative classifiers for imbalanced classification in the HDLSS setting, including cost-sensitive large-margin classifiers, random forests, and SMOTE.

The remainder of this paper is organized as follows. Section 2 begins with the connection between the quantile regression and imbalanced classification, which is followed by an introduction to the QuanDA algorithm. Section 3 provides the theoretical analysis. Section 4 presents simulation studies and real-data applications. All technical proofs are provided in the appendix.

2 Methodology

2.1 Background and motivation

We consider a binary classification problem with training data, $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where each $\mathbf{x}_i \in \mathbb{R}^p$ and the binary class label $y_i \in \{0, 1\}$. While $\{-1, 1\}$ encoding is common in large-margin classifiers for geometric interpretation, we opt for the $\{0, 1\}$ encoding for ease of handling class probabilities. The two coding approaches are essentially equivalent after an appropriate transformation.

Let $\pi_1 = P(Y = 1)$ and $\pi_0 = 1 - \pi_1 = P(Y = 0)$ denote the marginal class probabilities. Without loss of generality, we assume $\pi_1 \ll \pi_0$, which means that class 1 forms the *minority class*, whereas class 0 represents the *majority class*. Given the training data, our goal is to build a decision function that assigns a new test data point $\mathbf{x}_{\text{new}}$ to class 1 if $\phi(\mathbf{x}_{\text{new}}) = 1$ or class 0 if $\phi(\mathbf{x}_{\text{new}}) = 0$.

2.1.1 Review of Bayes risk

The performance of a classifier is fundamentally dictated by the Bayes risk, the lowest achieveable classification error given the true underlying distribution. Let $\mathbb{I}(\cdot)$ be the indicator function. The Bayes risk is defined as the expectation of the 0–1 loss on the population level (Lin, 2002):

$$R(\phi) = \mathbb{E}_{\mathbf{X}Y}[\mathbb{I}(Y = 0, \phi(\mathbf{X}) = 1) + \mathbb{I}(Y = 1, \phi(\mathbf{X}) = 0)],$$

whose minimizer gives the Bayes rule, say the theoretically optimal classifier under the 0–1 loss,

$$\phi^\star(\mathbf{X}) = \arg\min_{\phi} R(\phi) = \mathbb{I}(\eta(\mathbf{X}) > 1/2),$$

where $\eta(\mathbf{X}) = P(Y = 1|\mathbf{X})$ is the conditional probability for class 1.

To illustrate why the Bayes risk is unsuitable for direct application for imbalanced classification, note that $R(\phi)$ can be equivalently written as

$$\mathbb{E}_{\mathbf{X}}\left[\frac{\pi_0 g^-(\mathbf{X})}{\pi_0 g^-(\mathbf{X}) + \pi_1 g^+(\mathbf{X})}\mathbb{I}(\phi(\mathbf{X}) = 1) + \frac{\pi_1 g^+(\mathbf{X})}{\pi_0 g^-(\mathbf{X}) + \pi_1 g^+(\mathbf{X})}\mathbb{I}(\phi(\mathbf{X}) = 0)\right],$$

where $g^+(\mathbf{X})$ and $g^-(\mathbf{X})$ are the conditional density of $\mathbf{X}$ given $Y = 1$ and $Y = 0$, respectively. It is easily seen that the Bayes risk tends to be biased toward the majority class when $\pi_1 \ll \pi_0$.

To address class imbalance, Lin et al. (2002) introduce a weighted Bayes risk:

$$R(\phi) = \mathbb{E}_{\mathbf{X}Y}\left[\frac{w_0}{w_0 + w_1}\mathbb{I}(Y = 0, \phi(\mathbf{X}) = 1) + \frac{w_1}{w_0 + w_1}\mathbb{I}(Y = 1, \phi(\mathbf{X}) = 0)\right].$$

With these weights, the corresponding Bayes rule is given by

$$\phi^\star(\mathbf{X}) = \mathbb{I}\left(\eta(\mathbf{X}) > \frac{w_0}{w_0 + w_1}\right), \tag{1}$$

and the optimal Bayes risk is

$$R(\phi^\star) = \mathbb{E}_{\mathbf{X}}\left[\frac{w_0}{w_0 + w_1}(1 - \eta(\mathbf{X}))\mathbb{I}\left(\eta(\mathbf{X}) > \frac{w_0}{w_0 + w_1}\right) + \frac{w_1}{w_0 + w_1}\eta(\mathbf{X})\mathbb{I}\left(\eta(\mathbf{X}) \leq \frac{w_0}{w_0 + w_1}\right)\right]. \tag{2}$$

In the above framework, the most common choice of weights is $w_0 = \pi_1$ and $w_1 = \pi_0$. Given the class imbalanced, say $\pi_1 \ll \pi_0$, a higher penalty is thereby imposed on misclassifying the minority data. The framework can further extended to account for unequal costs of misclassification between the two classes. In particular, the weights can be adjusted to $w_0 = \pi_1 c_1$ and $w_1 = \pi_0 c_0$, where c_0 and c_1 represent the costs of a false positive and a false negative, respectively. In practice, c_1 is set to a higher value than c_0. Incorporating such costs into the weighting scheme leads to the mean-within-group-error criterion (Qiao and Liu, 2009; Qiao et al., 2010). Accordingly, the Bayes rule becomes

$$\phi^\star(\mathbf{X}) = \mathbb{I}\left(\eta(\mathbf{X}) > \frac{w_0}{w_0 + w_1} = \frac{\pi_1 c_1}{\pi_0 c_0 + \pi_1 c_1}\right). \tag{3}$$

2.1.2 From quantile to Bayes risk

We now establish an intuitive connection between the Bayes risk and quantile functions. With $\tau \in (0, 1)$, the quantile function is defined as

$$Q_Y(\tau|\mathbf{X}) = \inf\{y\colon F_{Y|\mathbf{X}}(y) \geq \tau\},$$

where $F_{Y|\mathbf{X}}$ denotes the conditional distribution function of Y given $\mathbf{X}$. Since Y is either 0 or 1, $Q_Y(\tau|\mathbf{X})$ is also binary. Moreover, it follows that $Q_Y(\tau|\mathbf{X}) = 1$ if and only if $P(Y = 1|\mathbf{X}) > 1-\tau$, that is,

$$Q_Y(\tau|\mathbf{X}) = \mathbb{I}(\eta(\mathbf{X}) > 1 - \tau).$$

Consequently, setting $\tau = w_1/(w_0+w_1) = \pi_0$, $Q_Y(\tau|\mathbf{X})$ exactly coincides with the Bayes rule given in equation (1). Likewise, the Bayes rule in equation (3) corresponds to using $\tau = w_1/(w_0 + w_1) = (\pi_1 c_1)/(\pi_0 c_0 + \pi_1 c_1)$.

To estimate the quantile function, it is well known that

$$Q_Y(\tau|\mathbf{X} = \mathbf{x}) \equiv \arg\min_{Q} \mathbb{E}\left[\rho_\tau(Y - Q(\mathbf{X}))|\mathbf{X} = \mathbf{x}\right],$$

where $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$ is the check or pinball loss. Note that the minimum of the objective function is

$$\begin{aligned}\mathbb{E}\left[\rho_\tau(Y - Q_Y(\tau|\mathbf{X}))|\mathbf{X} = \mathbf{x}\right] &= \rho_\tau(-\mathbb{I}(\eta(\mathbf{x}) > 1 - \tau))(1 - \eta(\mathbf{x})) + \rho_\tau(\mathbb{I}(\eta(\mathbf{x}) \leq 1 - \tau))\eta(\mathbf{x}) \\ &= \begin{cases}(1-\tau)(1-\eta(\mathbf{x})), & \text{if } \eta(\mathbf{x}) > 1-\tau, \\ \tau\eta(\mathbf{x}), & \text{if } \eta(\mathbf{x}) \leq 1-\tau.\end{cases}\end{aligned}$$

Therefore, setting $\tau = w_1/(w_0 + w_1)$, we see that the minimum of the population risk under the check loss is equivalent to the optimal Bayes risk given in equation (2).

2.2 Quantile-based discriminant analysis

We now introduce QuanDA based on the connection between quantile regression and Bayes risk. Because quantile regression is designed for continuous responses, directly applying it to binary class labels may lead to numerical instability. To address this, we propose to apply quantile regression on jittered class labels, and we shall show intimate connections to the Bayes risk framework.

Specifically, we craft a jittered response $Z = Y + U$ for a uniform random variable U on $[0, 1)$ that is independent of Y. It can be shown that Z is "almost" continuous, and $Q_Y(\tau|\mathbf{X}) = \lceil Q_{Y+U}(\tau|\mathbf{X}) - 1\rceil$, where $\lceil\cdot\rceil$ is the ceiling function that gives the smallest integer no less than its input. Importantly, the jittering does not actually affect the classification decision. To see this, for a given τ, we define $z = Q_{Y+U}(\tau)$, which gives rise to

$$\begin{aligned}\tau &= P(Y + U \leq z|\mathbf{X}) = P(Y = 0|\mathbf{X})P(U \leq z|\mathbf{X}) + P(Y = 1|\mathbf{X})P(U + 1 \leq z|\mathbf{x}) \\ &= (1 - \eta(\mathbf{X}))P(U \leq z|\mathbf{X}) + \eta(\mathbf{X})P(U + 1 \leq z|\mathbf{X}).\end{aligned} \tag{4}$$

When $\eta(\mathbf{X}) \leq 1 - \tau$, we know $z \leq 1$, because otherwise the right-hand side of equation (4) equals $(1 - \eta(\mathbf{X})) + \eta(\mathbf{X})(z - 1) > \alpha$. Therefore, taking $z \leq 1$, we see

$$Q_{Y+U}(\tau|\mathbf{X}) = z = \frac{\tau}{1 - \eta(\mathbf{X})} < 1.$$

Hence $Q_{Y+U}(\tau|\mathbf{X}) < 1$ is a sufficient condition to give $Q_Y(\tau|X) = 0$.

Likewise, when $\eta(\mathbf{X}) > 1 - \tau$, $z > 1$; otherwise, the right-hand side of equation (4) equals $1 - \eta(\mathbf{x})$, which is less than α. Thus, with $z > 1$, we have

$$Q_{Y+U}(\tau|\mathbf{X}) = z = 1 + \frac{\alpha - (1 - \eta(\mathbf{X}))}{\eta(\mathbf{X})} > 1,$$

which shows that $Q_{Y+U}(\tau|\mathbf{X}) > 1$ is sufficient to give $Q_Y(\tau|X) = 1$.

Consequently, we propose QuanDA by fitting quantile regression on the jitted response Z. To apply QuanDA to the HDLSS setting, we consider linear quantile regression model $Q_Z(\tau|\mathbf{X} = \mathbf{x}) = \alpha^\star(\tau) + \mathbf{x}^\top\boldsymbol{\beta}^\star(\tau)$ and adopt a sparse assumption on $\boldsymbol{\beta}^\star(\tau)$, where

$$(\alpha^\star(\tau), \boldsymbol{\beta}^\star(\tau)) = \arg\min_{\alpha,\boldsymbol{\beta}} \mathbb{E}[\rho_\tau(Z - \alpha - \mathbf{X}^\top\boldsymbol{\beta})], \tag{5}$$

Algorithm 1 Quantile-based Discriminant Analysis

Input: Training data: $\{\mathbf{x}_i, y_i\}_{i=1}^n$, $\tau = \hat{\pi}_1$.
for $r = 1, 2, \ldots, 10$ **do**
  Generate $U^{[r]} \sim \text{Uniform}(0, 1)$.
  Compute $Z^{[r]} = Y + U^{[r]}$.
  for $t = \tau - 0.05, \tau - 0.04, \ldots, \tau + 0.04, \tau + 0.05$ **do**
    For each λ_1, compute $(\widehat{\alpha}^{[r]}(t, \lambda_1), \widehat{\boldsymbol{\beta}}^{[r]}(t, \lambda_1))$ from

$$\min_{\alpha, \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(z_i^{[r]} - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1.$$

    Perform five-fold cross-validation to determine the optimal $\lambda_1^\star(t)^{[r]}$ based on the AUC scores.
  end for
end for
for $t = \tau - 0.05, \tau - 0.04, \ldots, \tau + 0.04, \tau + 0.05$ **do**
  Compute $\widehat{\alpha}(t) = \frac{1}{10}\sum_{r=1}^{10} \widehat{\alpha}^{[r]}(t, \lambda_1^\star(t)^{[r]})$ and $\widehat{\boldsymbol{\beta}}(t) = \frac{1}{10}\sum_{r=1}^{10} \widehat{\boldsymbol{\beta}}^{[r]}(t, \lambda_1^\star(t)^{[r]})$.
  Calculate the AUC scores based on $(\widehat{\alpha}(t), \widehat{\boldsymbol{\beta}}(t))$ to select the best $t^\star$.
end for

and the classification rule of QuanDA is $\phi(\mathbf{X}) = 1$ if $\alpha^\star(\tau) + \mathbf{X}^\top \boldsymbol{\beta}^\star(\tau) > 1$ or $\phi(\mathbf{X}) = 0$ otherwise.

In the sample version, we fit QuanDA using the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. For challenging imbalanced HDLSS classification, traditional dimension-reduction techniques such as principal component analysis (PCA) are often inadequate. Because PCA is unsupervised and does not incorporate class labels during dimensionality reduction, it may result in substantial information loss, particularly for the minority class. A more effective strategy, as suggested by Mai et al. (2012), is to employ supervised methods that impose sparse penalties to perform feature selection directly within the classification framework. Motivated by this idea, we introduce an ℓ_1-penalized formulation to achieve sparsity in the classifier:

$$\left(\widehat{\alpha}^\tau, \widehat{\boldsymbol{\beta}}^\tau\right) = \arg\min_{\alpha, \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(z_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1,$$

where $\|\boldsymbol{\beta}\|_1$ is the ℓ_1 norm of $\boldsymbol{\beta}$. The quantile level τ is determined by $w_1/(w_0 + w_1)$ according to the weighted Bayes risk. The class proportions π_1 and π_0 can be estimated by the sample proportions $\sum_{i=1}^n \mathbb{I}(y_i = 1)/n$ and $\sum_{i=1}^n \mathbb{I}(y_i = 0)/n$, respectively. For a new test data point $\mathbf{x}_{\text{new}}$, the classification rule is $\phi(\mathbf{x}_{\text{new}}) = 1$ if $\widehat{\alpha}^\tau + \mathbf{x}_{\text{new}}^\top \widehat{\boldsymbol{\beta}}^\tau > 1$ or $\phi(\mathbf{x}_{\text{new}}) = 0$ otherwise.

In our implementation of QuanDA, we use the elastic net penalty (Zou and Hastie, 2005), $\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$, instead to stabilize the algorithm.

Remark 1. In the implementation phase, to optimally tune the parameter τ, we first determine the proportion of the minority class, π_0, in the training data set. Next, we define a sequence ranging from $\pi_0 - 0.05$ to $\pi_0 + 0.05$, with increments of 0.01 between consecutive points. The optimal τ value is then selected using five-fold cross-validation. Details are summarized in Algorithm 1.

Remark 2. Another line of research called *quantitative classifiers* (Hennig and Viroli, 2016; Pritchard and Liu, 2020; Berrettini et al., 2024) is conceptually related but fundamentally different from QuanDA. These methods are distance-based classifiers, which assign class labels by computing a distance between each data point and each class, and extend the idea of median-based classifiers (Hall et al., 2009) by using quantiles of these distances to make predictions. Although QuanDA shares the term "quantile-based" classification, its quantile regression perspective differs from these distance-based frameworks. In the appendix, we shall compare QuanDA with quantileDA, a representative quantile-based classifier, and show that QuanDA consistently outperforms quantileDA.

3 Theoretical Studies

By construction, the distribution function of $Z = Y + U$ is continuous, but not smooth. In fact, it does not have continuous derivatives only at $\{0, 1, 2\}$. The standard theory for a quantile estimator would become problematic when the quantile turns out to be one. This can be resolved by assuming that the set of $\mathbf{x}$ for which $Q_Z(\tau|\mathbf{x}) = 1$ has measure zero. This is feasible when there exists at least one continuously distributed covariate and that the conditional quantiles $Q_Z(\tau|\mathbf{x})$ are measurable functions of that covariate. We make the following assumptions to show the estimation consistency of the quantile regression for the high-dimensional model.

- (C1) Y is a binary random variable supported on $\{0, 1\}$ and $\mathbf{X}$ is a random vector in $\mathbb{R}^p$. The mean function of Y given $\mathbf{X}$, $\eta(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$, is strictly between 0 and 1 for almost every realization $\mathbf{X} = \mathbf{x}$.
- (C2) The exists at least one continuously distributed covariate in $\mathbf{X}$.
- (C3) Make $Z = Y + U$, where $U \sim \text{Unif}(0, 1)$ is independent of Y and $\mathbf{X}$. The following restriction on the quantile process of Z given $\mathbf{X} = \mathbf{x}$ holds:

$$Q_Z(\tau|\mathbf{x}) = \alpha^*(\tau) + \mathbf{x}^\top \boldsymbol{\beta}^*(\tau) \;\text{ for } \tau \in (0, 1),$$

  where $\alpha^*(\tau) \in \mathbb{R}$ and $\boldsymbol{\beta}^*(\tau) \in \mathbb{R}^p$. Furthermore, if $\boldsymbol{\beta}^*_{(c)}(\tau)$ denotes the components of $\boldsymbol{\beta}^*(\tau)$ corresponding to the continuous covariates in $\mathbf{X}$, then $\mathbf{X}_{(c)}^\top \boldsymbol{\beta}^*_{(c)}(\tau) \neq \mathbf{0}$.
- (C4) For a.e. $\mathbf{x}$, the density $f_{\varepsilon|\mathbf{x}}(\cdot|\mathbf{x})$ of $\varepsilon_\tau \equiv Z - \mathbf{x}^\top \boldsymbol{\beta}^*(\tau)$ given $\mathbf{X} = \mathbf{x}$ satisfies: (1) $f_{\varepsilon|\mathbf{x}}(u|\mathbf{x})$ is continuously differentiable almost everywhere, and $f_{\varepsilon|\mathbf{x}}(u|\mathbf{x}) \le \bar{f}$ and $f'_{\varepsilon|\mathbf{x}}(u|\mathbf{x}) \le \bar{f}'$ for a.e. u in the support of ε_τ; (2) $f_{\varepsilon|\mathbf{x}}(\alpha^*(\tau) + u|\mathbf{x}) \ge \underline{f} > 0$ for all u in a small neighborhood of zero.
- (C5) Let $\mathcal{A} = \{j \in \{1, \ldots, p\} \colon \beta^*_j(\tau) \neq 0\}$ and $s = \text{card}(\mathcal{A})$. The covariates $\mathbf{X}$ satisfy

$$\kappa_m(u, v) = \inf_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{C}_{u,v},\, (\delta,\, \boldsymbol{\Delta}) \neq \mathbf{0}} \frac{\mathbb{E}[(\delta + \mathbf{X}^\top \boldsymbol{\Delta})^2]}{\|\boldsymbol{\Delta}_{\mathcal{A} \cup \overline{\mathcal{A}}(\boldsymbol{\Delta},\, m)}\|_2^2 + \delta^2} > 0,$$

  where $\mathcal{C}_{u,v} = \big\{(\delta, \boldsymbol{\Delta}) \colon \delta \in \mathbb{R},\, \boldsymbol{\Delta} \in \mathbb{R}^p,\, \|\boldsymbol{\Delta}_{\mathcal{A}^c}\|_1 \le u \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 + v|\delta|\big\}$ for some $u, v > 0$, $\overline{\mathcal{A}}(\boldsymbol{\Delta}, m) \subset \mathcal{A}^c$ is the support of the m largest in absolute value components of $\boldsymbol{\Delta}_{\mathcal{A}^c}$ for integer $m \ge 0$. When $m = 0$, we take $\overline{\mathcal{A}}(\boldsymbol{\Delta}, m) = \emptyset$.
- (C6) The covariates $\mathbf{X}$ satisfy

$$q = \frac{3}{8} \frac{\underline{f}^{3/2}}{\bar{f}'} \inf_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{C}_{u,v},\, (\delta,\, \boldsymbol{\Delta}) \neq \mathbf{0}} \frac{\big[\mathbb{E}(\delta + \mathbf{X}^\top \boldsymbol{\Delta})^2\big]^{3/2}}{\mathbb{E}|\delta + \mathbf{X}^\top \boldsymbol{\Delta}|^3} > 0.$$

Assumption (C1) is standard and ensures that classification is feasible. Assumptions (C2)–(C3) ensure that $Q_Z(\tau|\mathbf{x})$ has zero probability of taking integer values. Assumption (C4) is standard for quantile regression and works for a more general U than a uniform. Indeed, it is also possible to take U to have a beta distribution on $[0, 1)$. Assumption (C5) serves as the sparse identifiability condition or restricted eigenvalue condition, which is commonly imposed in the high-dimensional statistical literature (Candes and Tao, 2007; Bickel et al., 2009). The sparsity nonlinearity coefficient q in Assumption (C6) controls the quality of minorization of the quantile regression empirical loss by a quardratic function over sparse neighborhoods of the true parameter. It is often assumed in the high-dimensional quantile regression (Belloni and Chernozhukov, 2011).

Theorem 3.1. *Under conditions (C1)–(C6), with probability at least* $1 - p(\lambda)$, *where*

$$p(\lambda) = 2\exp\Big(-\frac{n\lambda^2}{2}\Big) + 2p\exp\Big(-\frac{n\lambda^2}{2M_0}\Big) + \exp\Big[-16M_0 \frac{s(1 + \log p)}{\kappa_0(3, 1)}\Big],$$

the lasso estimator $(\widehat{\alpha}_\lambda, \widehat{\boldsymbol{\beta}}_\lambda)$ *of the quantile regression satisfies*

$$\|\widehat{\alpha}_\lambda - \alpha^*\|_2 \le \frac{8}{\underline{f}\sqrt{\kappa_m(3, 1)}} \left[16\sqrt{\frac{2M_0}{\kappa_0(3, 1)}}\sqrt{\frac{1 + \log p}{n}}(2\sqrt{s} + 1) + \lambda\sqrt{\frac{s}{\kappa_0(3, 1)}}\right]$$

and

$$\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|_2 \leq \frac{8}{\underline{f}\sqrt{\kappa_m(3,1)}}\sqrt{1+\frac{18s}{m}+\frac{2}{m}}\cdot\left[16\sqrt{\frac{2M_0}{\kappa_0(3,1)}}\sqrt{\frac{1+\log p}{n}}(2\sqrt{s}+1)+\lambda\sqrt{\frac{s}{\kappa_0(3,1)}}\right],$$

provided that the growth condition

$$32\sqrt{\frac{2M_0}{\kappa_0(3,1)}}\sqrt{\frac{1+\log p}{n}}(2\sqrt{s}+1)+2\lambda\sqrt{\frac{s}{\kappa_0(3,1)}} \leq \underline{f}^{1/2}q$$

holds, where $M_0 = \max_{1\leq j\leq p}\mathbb{E}[X_j^2]$.

By Theorem 3.1, one can typically choose the parameter $\lambda = C\sqrt{\log p/n}$ for the quantile lasso estimator, where $C > \sqrt{2M_0}$ is some constant. For example, one can set $C = 2\sqrt{M_0}$. Note that given the design $\mathbf{X}$, M_0 can readily be obtained. Therefore, in principle, the parameter λ in the quantile lasso regression is tuning free. This is in similar spirit to square-root lasso Belloni et al. (2011). With such choice of λ, we can see that $p(\lambda) = o(1)$ as $n, p \to \infty$, which leads to

$$\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|_2 = \mathcal{O}_P\left(\frac{1}{\sqrt{\kappa_0\kappa_s}}\sqrt{\frac{s\log p}{n}}\right)$$

provided $q^{-1}\sqrt{s\log p/(n\kappa_0)} = o(1)$ and $\kappa_0(s\log p)^{-1} = o(1)$, by taking $m = s$. When κ_0 and κ_s are both positive constants, the quantile lasso estimator achieves the near-optimal rate $\sqrt{s\log p/n}$, which implies that p can be of exponential order of n, i.e., $\log p = \mathcal{O}(n^\gamma)$ for some $0 < \gamma < 1$, provided $s\log p = o(n)$.

4 Numerical Studies

The goal is to demonstrate that QuanDA consistently outperforms all competing methods, including logistic regression (implemented in the R package `glmnet`(Friedman et al., 2010)), the direct sparse discriminant analysis (dsda) (implemented in the R package `dsda` (Mai et al., 2012)), random forest (RF) (implemented in the R package `randomForest` (Liaw et al., 2002)) and SMOTE (implemented in the R package `smotefamily` (Siriseriwan, 2019)). In particular, we show that the performance of QuanDA remains robust even in highly imbalanced scenarios, where other classifiers collapse.

While many other methods have been developed for imbalanced classification, they are not tailored to the HDLSS setting and tend to collapse as illustrated in Figure 1. Hence, we focus our comparison only on a representative set of methods that are either commonly used or designed for HDLSS data.

4.1 Simulations

In the simulations, the dimension p is set to 10,000, with a sample size n of 400 and randomly generated class labels. The data set is highly unbalanced, and the majority class (π_0) comprises 85%, 90%, or 95% of the total, leaving the minority class (π_1) at 15%, 10%, or 5%. The simulation data we generate following the methodology described by Wang et al. (2006). The positive class follows a normal distribution with mean vector $\boldsymbol{\mu}_+$ and covariance matrix $\boldsymbol{\Sigma}$. The mean vector $\boldsymbol{\mu}_+$ is set to 0.7 for the first five features and 0 for the others. The covariance matrix $\boldsymbol{\Sigma}$ is defined as:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^\star_{5\times 5} & \mathbf{0}_{5\times(p-5)} \\ \mathbf{0}_{(p-5)\times 5} & \mathbf{I}_{(p-5)\times(p-5)} \end{pmatrix},$$

where $\boldsymbol{\Sigma}^\star_{5\times 5}$ takes the form of an autoregressive structure (AR$_\rho$) defined as $(\rho^{|i-j|})$, or a compound symmetric structure (CS$_\rho$) expressed as $(\rho + (1-\rho)\mathbb{I}(i=j))$, for $\rho \in \{0.2, 0.5, 0.7\}$. The negative class has the same distribution except for the mean vector $\boldsymbol{\mu}_- = -\boldsymbol{\mu}_+$.

We randomly split the simulation data into a training set of size 200 and a test set of size 200. In QuanDA method, we tuned the parameter λ by five-fold cross-validation and used the optimal τ from the candidate list given in Algorithm 1. To address imbalanced classification, we perform weighted logistic regression and weighted random forest, where the weights are determined based on the class proportions. Specifically, the weight for the majority class is set to $1/\pi_0$ and for the minority class, it

Table 1: The AUC scores of QuanDA for different combinations of λ_1 and λ_2, based on simulated data with $\mathbf{\Sigma} = \text{AR}_5$, $n = 400$, and $p = 10{,}000$. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses.

π_0	λ_1	λ_2	AUC	λ_1	λ_2	AUC
0.85	0.1	0.1	0.873 (0.041)	0.1	0.01	0.870 (0.042)
	0.01	0.1	0.798 (0.048)	0.01	0.01	0.805 (0.048)
0.9	0.1	0.1	0.843 (0.087)	0.1	0.01	0.834 (0.098)
	0.01	0.1	0.750 (0.065)	0.01	0.01	0.758 (0.070)
0.95	0.1	0.1	0.506 (0.043)	0.1	0.01	0.506 (0.043)
	0.01	0.1	0.682 (0.118)	0.01	0.01	0.674 (0.111)

Table 2: The comparison of AUC scores using simulated data with $n = 400$ and $p = 10000$. These scores are averaged over 50 independent runs with standard errors given in parentheses. More comparison results based on F1, G-means, and PRAUC are shown in the supplemental file.

	QuanDA	dsda	logistic	RF	SMOTE
			$\pi_0 = 0.85$		
AR_2	*0.923* (0.032)	0.922 (0.032)	0.915 (0.036)	0.786 (0.057)	0.770 (0.068)
AR_5	*0.872* (0.042)	0.861 (0.047)	0.842 (0.069)	0.763 (0.055)	0.767 (0.051)
AR_7	*0.849* (0.045)	0.840 (0.048)	0.821 (0.069)	0.749 (0.056)	0.746 (0.063)
CS_2	0.947 (0.028)	*0.949* (0.025)	0.945 (0.029)	0.786 (0.061)	0.791 (0.056)
CS_5	*0.907* (0.037)	*0.907* (0.036)	0.893 (0.043)	0.770 (0.051)	0.774 (0.050)
CS_7	*0.874* (0.042)	0.868 (0.046)	0.846 (0.070)	0.757 (0.056)	0.756 (0.057)
			$\pi_0 = 0.9$		
AR_2	*0.913* (0.041)	0.908 (0.046)	0.890 (0.075)	0.723 (0.075)	0.730 (0.087)
AR_5	*0.855* (0.059)	0.852 (0.061)	0.811 (0.122)	0.718 (0.075)	0.726 (0.069)
AR_7	*0.830* (0.065)	0.821 (0.082)	0.778 (0.122)	0.711 (0.080)	0.715 (0.087)
CS_2	*0.938* (0.037)	*0.938* (0.038)	0.923 (0.073)	0.743 (0.080)	0.746 (0.071)
CS_5	*0.894* (0.048)	0.890 (0.051)	0.869 (0.078)	0.730 (0.068)	0.728 (0.073)
CS_7	*0.856* (0.061)	0.854 (0.062)	0.810 (0.122)	0.719 (0.074)	0.730 (0.067)
			$\pi_0 = 0.95$		
AR_2	*0.827* (0.115)	0.813 (0.143)	0.714 (0.175)	0.661 (0.102)	0.675 (0.107)
AR_5	*0.770* (0.116)	0.738 (0.148)	0.660 (0.169)	0.658 (0.095)	0.648 (0.098)
AR_7	*0.740* (0.118)	0.715 (0.150)	0.645 (0.155)	0.649 (0.092)	0.652 (0.101)
CS_2	*0.850* (0.118)	0.850 (0.132)	0.750 (0.184)	0.657 (0.098)	0.662 (0.101)
CS_5	*0.803* (0.121)	0.791 (0.149)	0.695 (0.170)	0.659 (0.097)	0.646 (0.103)
CS_7	*0.772* (0.109)	0.744 (0.146)	0.668 (0.160)	0.645 (0.098)	0.653 (0.101)

is set to $1/\pi_1$. For both logistic regression and dsda, we also employ five-fold cross-validation to select the optimal parameter λ_1, given that $\lambda_2 = 0.01$.

Tables 2 summarizes the simulation results, averaged from 50 independent runs. We observe that our method consistently achieves the highest AUC scores in the all the simulation settings. Moreover, the advantages of our approach become increasingly evident as the imbalance ratio grows. The empirical performance confirms the effectiveness of QuanDA in addressing imbalanced classification.

We then conduct a sensitivity analysis of the hyperparameters in our algorithm. QuanDA involves two regularization parameters λ_1 and λ_2. We begin by evaluating the AUC scores of QuanDA under various combinations of λ_1 and λ_2, using simulated data generated with an AR_5 covariance structure, $n = 400$, and $p = 10{,}000$. To assess performance under different levels of class imbalance, we consider $\pi_0 \in \{0.85, 0.9, 0.95\}$. Table 1 show that, for a fixed value of λ_1, changes in λ_2 have little impact on QuanDA's performance. In contrast, varying λ_1 while holding λ_2 fixed has a more noticeable effect. Based on our numerical experiments, cross-validation typically provides a reliable choice for tuning λ_1.

Additional simulation results are provided in the supplementary materials, including comparisons of QuanDA with other widely used methods based on PRAUC, F1 score, and G-mean; see Tables S.1–S.7.

Table 3: The comparison of AUC scores on benchmark HDLSS data. The method achieving the highest AUC for each data set is italicized.

	QuanDA	dsda	logistic	RF	SMOTE
breast$_{(42,22283)}$	*0.949* (0.066)	0.926 (0.116)	0.944 (0.088)	0.891 (0.101)	0.895 (0.096)
leuk$_{(72,7128)}$	0.993 (0.016)	0.987 (0.044)	0.991 (0.028)	*0.997* (0.007)	0.996 (0.009)
LSVT$_{(126,309)}$	*0.909* (0.047)	0.901 (0.054)	0.898 (0.052)	0.884 (0.060)	0.884 (0.058)
ovarian$_{(253,15154)}$	*1.000* (0.000)	1.000 (0.002)	1.000 (0.002)	1.000 (0.001)	1.000 (0.001)
prostate$_{(102,6033)}$	*0.969* (0.029)	0.965 (0.025)	0.963 (0.027)	0.942 (0.047)	0.940 (0.049)

4.2 Benchmark Data Applications

In this section, we demonstrate the performance of QuanDA using seven benchmark high-dimensional data (Mai and Zou, 2015; Sorace and Zhan, 2003; Graham et al., 2010; Alon et al., 1999; Golub et al., 1999; Singh et al., 2002; Tsanas et al., 2013). All the benchmark data are available at the UCI Machine Learning Repository (Kelly et al., 2023). Those data sets have a varying dimensionality, ranging from 309 to 22,283. Each data set is partitioned into two parts: 70% is used for training and the remaining 30% for testing. The model fitting and parameter tuning are performed on the training set, after which the classification accuracy of the model is assessed on the test set.

Table 3 reports the average AUC scores from 50 independent repetitions. It shows that our method QuanDA overall outperforms all the other four methods in both metrics in those benchmark data examples.

5 Conclusion and Limitations

In this work, we have developed Quantile-based Discriminant Analysis (QuanDA), a method specifically designed to address imbalanced classification problems in high-dimensional, low-sample-size (HDLSS) settings. Extensive numerical studies demonstrate that QuanDA overall outperforms widely used imbalanced classification solvers, including cost-sensitive large-margin classifiers, random forests, and SMOTE. This shows that QuanDA is competitive in the challenging HDLSS scenarios.

Although this work focuses on binary classification, QuanDA can be naturally extended to multi-class settings. One straightforward approach is the one-vs-one method (Friedman, 1996; Hastie and Tibshirani, 1998), which decomposes a K-class problem into $K(K-1)/2$ pairwise binary classification tasks. A binary classifier is applied to each pair, and predictions are aggregated using majority voting. Another potential direction involves adapting the multiclass sparse discriminant analysis (msda) framework proposed by Mai and Zou (2015). Some preliminary results are presented in Section S2 in the supplementary material, while a comprehensive study including rigorous theoretical analysis is left for future work.

The strong empirical performance of QuanDA results in part from its simple structure, which makes it particularly suited for the HDLSS setting. Extending QuanDA to kernel learning and deep learning frameworks would be some promising future work for handling unstructured data.

In addition, a key contribution of this work is a novel connection between quantile regression and imbalanced classification. This connection enables the direct application of extensive quantile regression variants on imbalanced classification problems. For example, Qiao et al. (2023) proposed transfer learning to leverage external information for fitting quantile regression. The same framework can be directly integrated into our Algorithm 1 to transfer information to improve classification accuracy. Meta-learning for quantile regression (Fakoor et al., 2023) also shows promise for extending meta-learning techniques to imbalanced classification through quantile regression.

6 Acknowledgment

We would like to thank four anonymous reviewers and AC of NeurIPS 2025 for their constructive comments. Tang's research is supported by IRSA Faragher Distinguished Postdoctoral Fellowship. Wang's research is supported by National Institute Health grant R01GM163244-01.

References

M. Z. Abedin, C. Guotai, P. Hajek, and T. Zhang. Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*, 9(4):3559–3579, 2023.

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

M. Aoshima, D. Shen, H. Shen, K. Yata, Y.-H. Zhou, and J. S. Marron. A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics*, 60(1):4–19, 2018.

A. Azari, V. P. Janeja, and S. Levin. Imbalanced learning to predict long stay emergency department patients. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 807–814. IEEE, 2015.

S.-H. Bae and K.-J. Yoon. Polyp detection via imbalanced learning and discriminative feature learning. *IEEE Transactions on Medical Imaging*, 34(11):2379–2393, 2015.

A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

M. Berrettini, C. Hennig, and C. Viroli. The quantile-based classifier with variable-wise parameters. *arXiv preprint arXiv:2404.13589*, 2024.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

R. Blagus and L. Lusa. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14:1–16, 2013.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. SMOTEBoost: Improving prediction of the minority class in boosting. In *Proceedings of the Principles of Knowledge Discovery in Databases, PKDD-2003, Cavtat-Dubrovnik, Croatia*, pages 107–119. Springer, 2003.

C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. Technical Report 666, University of California, Berkeley, 2004.

D. Cieslak, N. Chawla, and A. Striegel. Combating imbalance in network intrusion datasets. In *2006 IEEE International Conference on Granular Computing*, pages 732–737, 2006.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

D. Devi, S. K. Biswas, and B. Purkayastha. A review on solution to class imbalance problem: Undersampling approaches. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 626–631. IEEE, 2020.

C. J. Evans and S. Ildstad. *Small Clinical Trials: Issues and Challenges*. National Academies Press (US), 2001.

R. Fakoor, T. Kim, J. Mueller, A. J. Smola, and R. J. Tibshirani. Flexible model aggregation for quantile regression. *Journal of Machine Learning Research*, 24(162):1–45, 2023.

Y. Feng, M. Zhou, and X. Tong. Imbalanced classification: A paradigm-based review. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(5):383–406, 2021.

A. Fernández. *Learning From Imbalanced Data Sets*. Springer, 2018.

J. H. Friedman. Another approach to polychotomous classification. *Technical Report, Statistics Department, Stanford University*, 1996.

J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

S. Fu, S. Zhang, and Y. Liu. Adaptively weighted large-margin angle-based classifiers. *Journal of multivariate analysis*, 166:282–299, 2018.

M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

A. Gosain and S. Sardana. Farthest SMOTE: a modified SMOTE approach. In *Computational Intelligence in Data Mining: Proceedings of the International Conference on CIDM 2017*, pages 309–320. Springer, 2019.

K. Graham, A. de Las Morenas, A. Tripathi, C. King, M. Kavanah, J. Mendez, M. Stone, J. Slama, M. Miller, G. Antoine, et al. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *British Journal of Cancer*, 102(8):1284–1293, 2010.

M. Grandini, E. Bagli, and G. Visani. Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*, 2020.

Y. Gu, J. Fan, L. Kong, S. Ma, and H. Zou. ADMM for high-dimensional sparse penalized quantile regression. *Technometrics*, 60(3):319–331, 2018.

P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3):427–444, 2005.

P. Hall, D. Titterington, and J.-H. Xue. Median-based classifiers for high-dimensional data. *Journal of the American Statistical Association*, 104(488):1597–1608, 2009.

T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2): 451–471, 1998.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

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

C. Hennig and C. Viroli. Quantile-based classifiers. *Biometrika*, 103(2):435–446, 2016.

S. Hu, Y. Liang, L. Ma, and Y. He. MSMOTE: Improving classification performance when training data is imbalanced. In *2009 Second International Workshop on Computer Science and Engineering*, volume 2, pages 13–17. IEEE, 2009.

J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

M. Kelly, R. Longjohn, and K. Nottingham. The UCI machine learning repository. School of Information and Computer Sciences, University of California, Irvine, 2023.

M. Khalilia, S. Chakraborty, and M. Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11:1–13, 2011.

K. Knight. Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, 26(2):755–770, 1998.

B. Krawczyk, M. Galar, Ł. Jeleń, and F. Herrera. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, 38:714–726, 2016.

M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the 14th International Conference on Machine Learning*, volume 97, pages 179–186, 1997.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces, Isoperimetry and Processes*. Springer-Verlag Berlin Heidelberg, 1991.

A. Liaw, M. Wiener, et al. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

Y. Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68 (1):73–82, 2004.

Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

Q. Mai and H. Zou. Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 135:175–188, 2015.

Q. Mai, H. Zou, and M. Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

J. S. Marron, M. J. Todd, and J. Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.

M. Muntasir Nishat, F. Faisal, I. Jahan Ratul, A. Al-Monsur, A. M. Ar-Rafi, S. M. Nasrullah, M. T. Reza, and M. R. H. Khan. A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. *Scientific Programming*, 2022(1):3649406, 2022.

G. Papandreou and A. L. Yuille. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *2011 International Conference on Computer Vision*, pages 193–200. IEEE, 2011.

B. Paula, L. Torgo, and R. Ribeiro. A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*, 2015.

D. A. Pritchard and Y. Liu. Composite quantile-based classifiers. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(4):337–353, 2020.

S. Qiao, Y. He, and W. Zhou. Transfer learning for high-dimensional quantile regression with statistical guarantee. *Transactions on Machine Learning Research*, 2023.

X. Qiao and Y. Liu. Adaptive weighted learning for unbalanced multicategory classification. *Biometrics*, 65(1):159–168, 2009.

X. Qiao, H. H. Zhang, Y. Liu, M. J. Todd, and J. S. Marron. Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association*, 105(489):401–414, 2010.

E. Ramentol, Y. Caballero, R. Bello, and F. Herrera. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. *Knowledge and Information Systems*, 33:245–265, 2012.

S. Rezvani and X. Wang. A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143:110415, 2023.

F. Sağlam and M. A. Cengiz. A novel SMOTE-based resampling technique trough noise detection and the boosting procedure. *Expert Systems with Applications*, 200:117023, 2022.

P. Santos, J. Maudes, and A. Bustillo. Identifying maximum imbalance in datasets for fault diagnosis of gearboxes. *Journal of Intelligent Manufacturing*, 29:333–351, 2018.

J. A. Sanz, D. Bernardo, F. Herrera, H. Bustince, and H. Hagras. A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. *IEEE Transactions on Fuzzy Systems*, 23(4):973–990, 2014.

S. J. Shin, Y. Wu, H. H. Zhang, and Y. Liu. Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, 104(1):67–81, 2017.

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

W. Siriseriwan. Smotefamily: A collection of oversampling techniques for class imbalance problem based on SMOTE. *R package version*, 1(1):15, 2019.

J. M. Sorace and M. Zhan. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, 4:1–13, 2003.

Q. Tang, Y. Zhang, and B. Wang. Finite smoothing algorithm for high-dimensional support vector machines and quantile regression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47865–47884. PMLR, 2024.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108, 2005.

A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig. Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):181–190, 2013.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

B. Wang and H. Zou. Another look at distance-weighted discrimination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):177–198, 2018.

L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16: 589–615, 2006.

L. Wang, Y. Wu, and R. Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012.

W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16:449–475, 2013.

T. Weiss. Correlates of posttraumatic growth in married breast cancer survivors. *Journal of Social and Clinical Psychology*, 23(5):733–746, 2004.

Q. Wu, Y. Ye, H. Zhang, M. K. Ng, and S.-S. Ho. ForesTexter: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems*, 67:105–116, 2014.

Z. Wu, W. Lin, and Y. Ji. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. *IEEE Access*, 6:8394–8402, 2018.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

L. Zeng and H. H. Zhang. Sparse learning and class probability estimation with weighted support vector machines. *arXiv preprint arXiv:2312.10618*, 2023.

C. Zhang, Y. Liu, J. Wang, and H. Zhu. Reinforced angle-based multicategory support vector machines. *Journal of Computational and Graphical Statistics*, 25(3):806–825, 2016.

Q. Zheng, L. Peng, and X. He. Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics*, 43(5):2225, 2015.

Z.-B. Zhu and Z.-H. Song. Fault diagnosis based on imbalance modified kernel fisher discriminant analysis. *Chemical Engineering Research and Design*, 88(8):936–951, 2010.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

SUPPLEMENTARY MATERIAL

S1 Additional Numerical results

S1.1 Simulations

All numerical experiments in this work were carried out on an Intel(R) Xeon(R) Gold 6430 (3.40 GHz) processor.

For high-dimensional imbalanced classification problems, relying solely on the AUC score is insufficient to fully evaluate the effectiveness of a method. To offer a more comprehensive assessment, we compare our proposed approach with several baseline methods using additional performance metrics, including the area under the precision-recall curve (PRAUC), the F1 score, and the geometric mean (G-mean). Tables S.1 to S.3 show that QuanDA consistently outperforms all competing methods across all evaluation metrics. Notably, quantileDA is excluded from AUC and PRAUC comparisons, as it only produces predicted class labels rather than probability scores. Consequently, AUC and PRAUC cannot be computed for this method. Furthermore, random forest, SMOTE, and quantileDA yield F1 scores and G-means of zero in certain settings, as they predict all samples as belonging to the majority class, failing to identify any instances of the minority class.

Tables S.4–S.7 report the AUC, PRAUC, F1, and G-mean scores for QuanDA and several widely used methods for imbalanced classification, evaluated on simulated data with $n = 400$ and $p = 5000$. QuanDA consistently achieves superior performance across all evaluation metrics compared to the competing methods.

S1.2 Real-data analysis

We further evaluate QuanDA against other methods using real-world datasets, focusing on PRAUC, F1, and G-mean scores. As shown in Tables **??** and **??**, QuanDA consistently demonstrates superior performance across all metrics, while quantileDA exhibits the least effective results among the compared methods.

Table S.1: The PRAUC scores for five imbalanced classification solvers were evaluated using simulated data with $n = 400$ and $p = 10000$. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses.

	QuanDA	dsda	logistic	RF	SMOTE
			$\pi_0 = 0.85$		
AR_2	*0.756* (0.086)	0.750 (0.088)	0.733 (0.100)	0.431 (0.098)	0.413 (0.110)
AR_5	*0.626* (0.103)	0.600 (0.108)	0.572 (0.123)	0.390 (0.088)	0.380 (0.100)
AR_7	*0.573* (0.101)	0.552 (0.107)	0.525 (0.120)	0.378 (0.086)	0.361 (0.092)
CS_2	*0.829* (0.072)	0.827 (0.075)	0.819 (0.086)	0.439 (0.100)	0.451 (0.118)
CS_5	*0.713* (0.100)	0.707 (0.098)	0.683 (0.115)	0.412 (0.101)	0.428 (0.086)
CS_7	*0.629* (0.103)	0.613 (0.112)	0.580 (0.133)	0.389 (0.108)	0.382 (0.105)
			$\pi_0 = 0.9$		
AR_2	*0.657* (0.136)	0.638 (0.160)	0.604 (0.174)	0.272 (0.083)	0.286 (0.098)
AR_5	*0.588* (0.140)	0.501 (0.144)	0.456 (0.182)	0.269 (0.105)	0.274 (0.101)
AR_7	*0.452* (0.140)	0.443 (0.149)	0.396 (0.172)	0.265 (0.096)	0.263 (0.088)
CS_2	*0.737* (0.129)	0.734 (0.149)	0.696 (0.178)	0.294 (0.103)	0.296 (0.093)
CS_5	*0.600* (0.145)	0.589 (0.163)	0.551 (0.174)	0.288 (0.115)	0.280 (0.097)
CS_7	*0.508* (0.142)	0.503 (0.154)	0.453 (0.188)	0.264 (0.101)	0.273 (0.086)
			$\pi_0 = 0.95$		
AR_2	0.353 (0.211)	*0.368* (0.224)	0.262 (0.239)	0.129 (0.081)	0.144 (0.104)
AR_5	*0.243* (0.151)	0.243 (0.176)	0.184 (0.181)	0.120 (0.071)	0.116 (0.066)
AR_7	*0.211* (0.147)	0.207 (0.152)	0.150 (0.144)	0.122 (0.071)	0.121 (0.079)
CS_2	0.406 (0.228)	*0.452* (0.258)	0.321 (0.271)	0.121 (0.078)	0.141 (0.091)
CS_5	0.302 (0.197)	*0.336* (0.216)	0.219 (0.205)	0.140 (0.089)	0.118 (0.070)
CS_7	*0.248* (0.141)	0.243 (0.183)	0.184 (0.177)	0.132 (0.086)	0.124 (0.070)

Table S.2: The F1 scores for six imbalanced classification solvers were evaluated using simulated data with $n = 400$ and 10000. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses. Notably, random forest, SMOTE, and quantileDA yield F1 scores of zero in this setting.

	QuanDA	dsda	logistic
		$\pi_0 = 0.85$	
AR_2	*0.663* (0.069)	0.648 (0.109)	0.635 (0.079)
AR_5	*0.561* (0.071)	0.500 (0.097)	0.531 (0.084)
AR_7	*0.528* (0.072)	0.440 (0.107)	0.494 (0.087)
CS_2	0.724 (0.072)	*0.730* (0.085)	0.711 (0.078)
CS_5	*0.634* (0.076)	0.608 (0.104)	0.600 (0.089)
CS_7	*0.563* (0.074)	0.513 (0.094)	0.528 (0.095)
		$\pi_0 = 0.9$	
AR_2	*0.586* (0.088)	0.512 (0.199)	0.539 (0.126)
AR_5	*0.482* (0.097)	0.376 (0.161)	0.447 (0.112)
AR_7	*0.444* (0.085)	0.334 (0.143)	0.404 (0.104)
CS_2	*0.655* (0.084)	0.599 (0.201)	0.614 (0.118)
CS_5	*0.538* (0.089)	0.472 (0.185)	0.503 (0.124)
CS_7	*0.486* (0.091)	0.389 (0.168)	0.448 (0.103)
		$\pi_0 = 0.95$	
AR_2	*0.399* (0.161)	0.193 (0.238)	0.309 (0.189)
AR_5	*0.314* (0.136)	0.126 (0.169)	0.225 (0.172)
AR_7	*0.288* (0.134)	0.109 (0.157)	0.196 (0.172)
CS_2	*0.453* (0.179)	0.290 (0.271)	0.370 (0.203)
CS_5	*0.358* (0.151)	0.175 (0.219)	0.265 (0.189)
CS_7	*0.313* (0.141)	0.141 (0.178)	0.225 (0.177)

Table S.3: The G-mean scores for six imbalanced classification solvers were evaluated using simulated data with $n = 400$ and $p = 10000$. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses. Notably, random forest, SMOTE, and quantileDA yield G-mean scores of zero in this setting.

	QuanDA	dsda	logistic
		$\pi_0 = 0.85$	
AR_2	*0.844* (0.047)	0.747 (0.093)	0.808 (0.072)
AR_5	*0.788* (0.052)	0.629 (0.084)	0.747 (0.087)
AR_7	*0.761* (0.053)	0.576 (0.097)	0.721 (0.091)
CS_2	*0.872* (0.047)	0.813 (0.072)	0.854 (0.071)
CS_5	*0.828* (0.055)	0.716 (0.090)	0.787 (0.088)
CS_7	*0.783* (0.056)	0.640 (0.079)	0.740 (0.100)
		$\pi_0 = 0.9$	
AR_2	*0.812* (0.069)	0.632 (0.205)	0.775 (0.123)
AR_5	*0.757* (0.070)	0.513 (0.179)	0.719 (0.133)
AR_7	*0.730* (0.073)	0.473 (0.163)	0.685 (0.166)
CS_2	*0.848* (0.062)	0.700 (0.205)	0.811 (0.103)
CS_5	*0.789* (0.078)	0.598 (0.189)	0.750 (0.127)
CS_7	*0.758* (0.088)	0.522 (0.181)	0.723 (0.124)
		$\pi_0 = 0.95$	
AR_2	*0.690* (0.152)	0.262 (0.306)	0.572 (0.310)
AR_5	*0.602* (0.188)	0.194 (0.244)	0.456 (0.327)
AR_7	*0.592* (0.150)	0.165 (0.222)	0.399 (0.334)
CS_2	*0.705* (0.162)	0.381 (0.332)	0.596 (0.317)
CS_5	*0.657* (0.158)	0.246 (0.287)	0.502 (0.319)
CS_7	*0.623* (0.126)	0.213 (0.249)	0.446 (0.328)

Table S.4: The AUC scores for five imbalanced classification solvers were evaluated using simulated data with $n = 400$ and $p = 5000$. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses.

	QuanDA	dsda	logistic	RF	SMOTE
			$\pi_0 = 0.85$		
AR_2	*0.931* (0.025)	0.925 (0.028)	0.919 (0.029)	0.830 (0.051)	0.829 (0.051)
AR_5	*0.882* (0.034)	0.871 (0.036)	0.861 (0.041)	0.801 (0.049)	0.804 (0.051)
AR_7	*0.857* (0.038)	0.845 (0.042)	0.831 (0.048)	0.781 (0.059)	0.782 (0.065)
CS_2	*0.956* (0.018)	0.953 (0.020)	0.949 (0.023)	0.845 (0.044)	0.838 (0.054)
CS_5	*0.916* (0.029)	0.909 (0.031)	0.901 (0.034)	0.818 (0.047)	0.816 (0.057)
CS_7	*0.882* (0.035)	0.874 (0.039)	0.863 (0.044)	0.806 (0.055)	0.804 (0.055)
			$\pi_0 = 0.9$		
AR_2	*0.923* (0.039)	0.908 (0.052)	0.891 (0.081)	0.793 (0.066)	0.793 (0.075)
AR_5	*0.874* (0.047)	0.849 (0.073)	0.827 (0.098)	0.769 (0.071)	0.767 (0.075)
AR_7	*0.846* (0.060)	0.816 (0.092)	0.794 (0.105)	0.744 (0.085)	0.755 (0.080)
CS_2	*0.948* (0.031)	0.940 (0.039)	0.930 (0.055)	0.804 (0.071)	0.790 (0.089)
CS_5	*0.907* (0.043)	0.888 (0.061)	0.868 (0.096)	0.777 (0.081)	0.786 (0.071)
CS_7	*0.873* (0.050)	0.853 (0.072)	0.830 (0.098)	0.755 (0.084)	0.763 (0.070)
			$\pi_0 = 0.95$		
AR_2	*0.874* (0.081)	0.833 (0.125)	0.760 (0.159)	0.706 (0.120)	0.690 (0.115)
AR_5	*0.821* (0.090)	0.780 (0.129)	0.698 (0.147)	0.689 (0.103)	0.683 (0.112)
AR_7	*0.793* (0.094)	0.747 (0.133)	0.674 (0.142)	0.681 (0.105)	0.680 (0.107)
CS_2	*0.901* (0.079)	0.871 (0.108)	0.792 (0.163)	0.713 (0.108)	0.687 (0.104)
CS_5	*0.855* (0.088)	0.813 (0.131)	0.745 (0.149)	0.689 (0.125)	0.693 (0.104)
CS_7	*0.821* (0.089)	0.778 (0.137)	0.693 (0.141)	0.675 (0.108)	0.676 (0.105)

Table S.5: The PRAUC scores for five imbalanced classification solvers were evaluated using simulated data with $n = 400$ and $p = 5000$. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses.

	QuanDA	dsda	logistic	RF	SMOTE
			$\pi_0 = 0.85$		
AR_2	*0.777* (0.065)	0.763 (0.080)	0.748 (0.075)	0.528 (0.110)	0.524 (0.113)
AR_5	*0.650* (0.078)	0.627 (0.089)	0.609 (0.090)	0.456 (0.107)	0.468 (0.097)
AR_7	*0.596* (0.080)	0.574 (0.096)	0.550 (0.095)	0.418 (0.109)	0.423 (0.117)
CS_2	*0.846* (0.051)	0.842 (0.058)	0.832 (0.061)	0.558 (0.111)	0.544 (0.128)
CS_5	*0.738* (0.073)	0.723 (0.083)	0.701 (0.085)	0.503 (0.092)	0.489 (0.116)
CS_7	*0.654* (0.081)	0.635 (0.094)	0.614 (0.096)	0.478 (0.104)	0.470 (0.109)
			$\pi_0 = 0.9$		
AR_2	*0.682* (0.126)	0.655 (0.141)	0.625 (0.152)	0.376 (0.116)	0.373 (0.113)
AR_5	*0.555* (0.119)	0.507 (0.140)	0.477 (0.156)	0.331 (0.116)	0.323 (0.106)
AR_7	*0.491* (0.113)	0.449 (0.141)	0.416 (0.151)	0.303 (0.110)	0.291 (0.091)
CS_2	*0.764* (0.115)	0.751 (0.127)	0.726 (0.141)	0.381 (0.126)	0.381 (0.131)
CS_5	*0.634* (0.131)	0.603 (0.150)	0.569 (0.163)	0.340 (0.114)	0.351 (0.121)
CS_7	*0.548* (0.124)	0.514 (0.146)	0.480 (0.159)	0.327 (0.109)	0.308 (0.097)
			$\pi_0 = 0.95$		
AR_2	*0.428* (0.195)	0.390 (0.208)	0.315 (0.210)	0.186 (0.138)	0.167 (0.122)
AR_5	*0.315* (0.169)	0.281 (0.160)	0.218 (0.173)	0.149 (0.093)	0.144 (0.127)
AR_7	*0.263* (0.154)	0.248 (0.159)	0.186 (0.160)	0.137 (0.093)	0.142 (0.087)
CS_2	*0.527* (0.191)	0.471 (0.217)	0.372 (0.227)	0.185 (0.125)	0.167 (0.107)
CS_5	*0.400* (0.177)	0.346 (0.197)	0.269 (0.188)	0.162 (0.114)	0.165 (0.108)
CS_7	*0.305* (0.173)	0.286 (0.177)	0.197 (0.163)	0.145 (0.104)	0.146 (0.100)

Table S.6: The F1 scores for six imbalanced classification solvers were evaluated using simulated data with $n = 400$ and $p = 5000$. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses. Notably, random forest, SMOTE, and quantileDA yield F1 scores of zero in this setting.

	QuanDA	dsda	logistic
		$\pi_0 = 0.85$	
AR_2	*0.670* (0.072)	0.648 (0.109)	0.653 (0.071)
AR_5	*0.562* (0.073)	0.521 (0.090)	0.544 (0.071)
AR_7	*0.525* (0.069)	0.463 (0.094)	0.500 (0.074)
CS_2	*0.733* (0.065)	0.732 (0.089)	0.725 (0.059)
CS_5	*0.632* (0.076)	0.613 (0.091)	0.618 (0.070)
CS_7	*0.569* (0.076)	0.530 (0.091)	0.550 (0.072)
		$\pi_0 = 0.9$	
AR_2	*0.584* (0.093)	0.547 (0.151)	0.577 (0.108)
AR_5	*0.478* (0.099)	0.403 (0.160)	0.453 (0.093)
AR_7	*0.444* (0.101)	0.342 (0.150)	0.419 (0.108)
CS_2	*0.655* (0.091)	0.626 (0.141)	0.641 (0.116)
CS_5	*0.548* (0.101)	0.496 (0.159)	0.527 (0.109)
CS_7	*0.484* (0.101)	0.408 (0.176)	0.456 (0.101)
		$\pi_0 = 0.95$	
AR_2	*0.413* (0.144)	0.279 (0.224)	0.293 (0.180)
AR_5	*0.337* (0.127)	0.187 (0.177)	0.234 (0.143)
AR_7	*0.292* (0.122)	0.146 (0.183)	0.200 (0.121)
CS_2	*0.471* (0.155)	0.325 (0.265)	0.354 (0.201)
CS_5	*0.393* (0.142)	0.228 (0.212)	0.258 (0.155)
CS_7	*0.335* (0.134)	0.181 (0.191)	0.222 (0.134)

Table S.7: The G-mean scores for six imbalanced classification solvers were evaluated using simulated data with $n = 400$ and $p = 5000$. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses. Notably, random forest, SMOTE, and quantileDA yield G-means of zero in this setting.

	QuanDA	dsda	logistic
		$\pi_0 = 0.85$	
AR_2	*0.843* (0.037)	0.744 (0.094)	0.820 (0.060)
AR_5	*0.782* (0.045)	0.644 (0.075)	0.754 (0.070)
AR_7	*0.754* (0.048)	0.594 (0.078)	0.730 (0.085)
CS_2	*0.874* (0.033)	0.809 (0.078)	0.855 (0.049)
CS_5	*0.822* (0.039)	0.718 (0.081)	0.801 (0.058)
CS_7	*0.787* (0.046)	0.649 (0.076)	0.758 (0.066)
		$\pi_0 = 0.9$	
AR_2	*0.802* (0.072)	0.664 (0.155)	0.793 (0.093)
AR_5	*0.752* (0.077)	0.535 (0.175)	0.721 (0.103)
AR_7	*0.731* (0.075)	0.474 (0.183)	0.707 (0.138)
CS_2	*0.845* (0.066)	0.732 (0.128)	0.821 (0.109)
CS_5	*0.787* (0.081)	0.621 (0.171)	0.755 (0.102)
CS_7	*0.744* (0.087)	0.535 (0.200)	0.727 (0.104)
		$\pi_0 = 0.95$	
AR_2	*0.698* (0.124)	0.381 (0.289)	0.568 (0.290)
AR_5	*0.652* (0.142)	0.280 (0.249)	0.505 (0.287)
AR_7	*0.610* (0.147)	0.216 (0.256)	0.462 (0.282)
CS_2	*0.731* (0.134)	0.418 (0.323)	0.581 (0.299)
CS_5	*0.681* (0.122)	0.320 (0.279)	0.527 (0.296)
CS_7	*0.652* (0.123)	0.266 (0.260)	0.488 (0.295)

S2 Extension to Imbalanced Multi-class Classification

Although QuanDA is primarily designed for binary classification tasks, it can be extended to imbalanced multi-class classification under HDLSS settings using strategies. In this section, we present a simple example to illustrate the effectiveness of QuanDA in handling multi-class classification using the one-vs-one strategy. We leave the development and evaluation of more advanced extensions to future work.

Motivated by the simulation design described in Section 4, we consider a three-class imbalanced classification problem in which each class follows a multivariate normal distribution with a common covariance matrix Σ. The class-specific mean vectors differ only in the first five features, with values set to $0.7, 0$ and -0.7 , respectively. Two imbalance scenarios are examined. In the first scenario, the majority class (π_0) accounts for 80% of the total sample, while each of the two minority classes represents 10%. In the second scenario, the majority class comprises 90% of the total sample, and each minority class accounts for 5%. The total sample size is set to $n = 100$ or $n = 400$, and the feature dimension is $p = 10,000$. We compare the performance of QuanDA with several competing methods, including MSDA, weighted random forest, and SMOTE. The macro F1 score (Grandini et al., 2020) is employed as the evaluation metric to assess the multi-class classification performance of each method. Table S.8 shows that QuanDA achieves the highest macro F1 score among the compared methods. In contrast, weighted random forest and SMOTE fails to identify the minority classes, predicting all samples as belonging to the majority class.

Table S.8: The macro F1 scores for three imbalanced classification solvers were evaluated using simulated data with $p = 10000$. These scores represent the average results obtained over 50 independent runs with standard errors given in parentheses.

$n = 400$	QuanDA	msda	RF	SMOTE
		8:1:1		
AR_2	0.515 (0.073)	0.369 (0.031)	0.297 (0.000)	0.297 (0.000)
AR_5	0.490 (0.069)	0.334 (0.023)	0.297 (0.000)	0.297 (0.000)
AR_7	0.479 (0.071)	0.318 (0.021)	0.297 (0.000)	0.297 (0.000)
CS_2	0.524 (0.068)	0.401 (0.029)	0.297 (0.000)	0.297 (0.000)
CS_5	0.506 (0.072)	0.346 (0.020)	0.297 (0.000)	0.297 (0.000)
CS_7	0.488 (0.061)	0.322 (0.017)	0.297 (0.000)	0.297 (0.000)
		9:0.5:0.5		
AR_2	0.460 (0.067)	0.319 (0.016)	0.317 (0.000)	0.317 (0.000)
AR_5	0.438 (0.064)	0.323 (0.028)	0.317 (0.000)	0.317 (0.000)
AR_7	0.425 (0.065)	0.324 (0.028)	0.317 (0.000)	0.317 (0.000)
CS_2	0.478 (0.068)	0.325 (0.031)	0.317 (0.000)	0.317 (0.000)
CS_5	0.460 (0.070)	0.325 (0.033)	0.317 (0.000)	0.317 (0.000)
CS_7	0.435 (0.071)	0.322 (0.023)	0.317 (0.000)	0.317 (0.000)

S3 Proof of Theorem 3.1

For ease of notation, we fix $\tau \in (0, 1)$ and drop all subscript τ wherever no confusion arises. Let

$$Q_n(\alpha, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(z_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta}).$$

For $\lambda > 0$, define the lasso estimator of the quantile regression by

$$(\widehat{\alpha}_\lambda, \widehat{\boldsymbol{\beta}}_\lambda) := \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\arg\min}\, Q_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{6}$$

Let

$$\nu_n(\alpha, \boldsymbol{\beta}) = Q_n(\alpha, \boldsymbol{\beta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*) - \mathbb{E}\big[Q_n(\alpha, \boldsymbol{\beta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*)\big].$$

For some $r > 0$, set

$$\mathcal{G}_{u,v,r} = \{(\delta, \boldsymbol{\Delta}) \in \mathcal{C}_{u,v} : n^{-1} \sum_{i=1}^{n} (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2 \le r^2\}.$$

Also, define
$$e(u,v,r)=\sup_{(\delta,\,\mathbf{\Delta})\in\mathcal{G}_{u,v,r}}|\nu_n(\alpha^*+\delta,\boldsymbol{\beta}^*+\mathbf{\Delta})|.$$
Let $F_{\varepsilon|\mathbf{x}}$ and $f_{\varepsilon|\mathbf{x}}$ be the distribution and density functions of ε_τ, respectively. We shall write them as F and f for simplicity of notation in the proofs. The following proofs are based on a given $\mathbf{X}$ (i.e., conditional on $\mathbf{X}$), but can be easily modified for a stochastic $\mathbf{X}$.

Lemma S3.1. *Under conditions (C1)–(C3), with probability at least*
$$1-2\exp\Big(-\frac{n\lambda^2}{2}\Big)-2p\exp\Big(-\frac{n\lambda^2}{2M_0}\Big),$$
the lasso estimator $(\widehat{\boldsymbol{\alpha}}_\lambda,\widehat{\boldsymbol{\beta}}_\lambda)$ of the quantile regression satisfies
$$(\widehat{\delta}^\lambda,\widehat{\mathbf{\Delta}}^\lambda)\in\mathcal{C}_{3,1}=\{(\delta,\mathbf{\Delta})\colon\delta\in\mathbb{R},\ \mathbf{\Delta}\in\mathbb{R}^p,\ \|\mathbf{\Delta}_{\mathcal{A}^c}\|_1\le 3\|\mathbf{\Delta}_{\mathcal{A}}\|_1+|\delta|\},$$
where $\widehat{\delta}^\lambda=\widehat{\alpha}_\lambda-\alpha^*$ and $\widehat{\mathbf{\Delta}}^\lambda=\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*$.

Proof of Lemma S3.1. Let
$$\zeta=-\frac{1}{n}\sum_{i=1}^n\big[\tau-I(\varepsilon_i\le\alpha^*)\big],$$
and $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_p)^\top$, where
$$\xi_j=-\frac{1}{n}\sum_{i=1}^n\big[\tau-I(\varepsilon_i\le\alpha^*)\big]x_{ij},\ 1\le j\le p.$$
Note that $(\zeta,\boldsymbol{\xi}^\top)^\top\in\partial Q_n(\alpha^*,\boldsymbol{\beta}^*)$, where the subdifferential is taken with respect to α and $\boldsymbol{\beta}$. By convexity of $Q_n(\alpha,\boldsymbol{\beta})$ and optimality of $(\widehat{\alpha}_\lambda,\widehat{\boldsymbol{\beta}}_\lambda)$, we have
$$\begin{aligned}0&\ge Q_n(\widehat{\alpha}_\lambda,\widehat{\boldsymbol{\beta}}_\lambda)-Q_n(\alpha^*,\boldsymbol{\beta}^*)+\lambda(\|\widehat{\boldsymbol{\beta}}_\lambda\|_1-\|\boldsymbol{\beta}^*\|_1)\\&\ge\zeta(\widehat{\alpha}_\lambda-\alpha^*)+\boldsymbol{\xi}^\top(\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*)+\lambda(\|\widehat{\boldsymbol{\beta}}_\lambda\|_1-\|\boldsymbol{\beta}^*\|_1)\\&\ge-|\zeta|\cdot|\widehat{\alpha}_\lambda-\alpha^*|-\|\boldsymbol{\xi}\|_\infty\cdot\|\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*\|_1\\&\qquad+\lambda\Big(\sum_{j\in\mathcal{A}^c}|\hat{\beta}_{\lambda,j}-\beta_j^*|-\sum_{j\in\mathcal{A}}|\hat{\beta}_{\lambda,j}-\beta_j^*|\Big),\end{aligned}$$
which implies that
$$(\lambda-\|\boldsymbol{\xi}\|_\infty)\sum_{j\in\mathcal{A}^c}|\hat{\beta}_{\lambda,j}-\beta_j^*|\le(\lambda+\|\boldsymbol{\xi}\|_\infty)\sum_{j\in\mathcal{A}}|\hat{\beta}_{\lambda,j}-\beta_j^*|+|\zeta|\cdot|\widehat{\alpha}_\lambda-\alpha^*|.\tag{7}$$
Under event $\mathcal{E}=\{|\zeta|\le\lambda/2,\ \|\boldsymbol{\xi}\|_\infty\le\lambda/2\}$, it follows from (7) that
$$\|\widehat{\mathbf{\Delta}}^\lambda_{\mathcal{A}^c}\|_1\le3\|\widehat{\mathbf{\Delta}}^\lambda_{\mathcal{A}}\|_1+|\widehat{\delta}^\lambda|.$$
The lemma then follows from Hoeffding's inequality
$$\begin{aligned}\Pr(\mathcal{E})&\ge1-\Pr\Big(|\zeta|>\frac{\lambda}{2}\Big)-\Pr\Big(\|\boldsymbol{\xi}\|_\infty>\frac{\lambda}{2}\Big)\\&\ge1-\Pr\Big(\Big|-\frac{1}{n}\sum_{i=1}^n[\tau-I(\varepsilon_i\le\alpha^*)]\Big|>\frac{\lambda}{2}\Big)\\&\qquad-\sum_{j=1}^p\Pr\Big(\Big|-\frac{1}{n}\sum_{i=1}^n x_{ij}[\tau-I(\varepsilon_i\le\alpha^*)]\Big|>\frac{\lambda}{2}\Big)\\&\ge1-2\exp\Big(-\frac{n\lambda^2}{2}\Big)-2p\exp\Big(-\frac{n\lambda^2}{2M_0}\Big),\end{aligned}$$
where $M_0=\max_{1\le j\le p}\mathbb{E}\left[X_j^2\right]$. $\square$

Lemma S3.2. *For $u, v, r, t > 0$, under conditions (C1)–(C6), with probability at least* $1 - \exp\left[-nt^2/(32r^2)\right]$, *we have*

$$e(u,v,r) \le 4\sqrt{\frac{2M_0}{\kappa_0(u,v)}}\sqrt{\frac{1+\log p}{n}}\Big[(1+u)\sqrt{s}+(1+v)\Big]r + t$$

when $p \ge 3$. *It follows immediately that, if one takes*

$$t = 4\sqrt{\frac{2M_0}{\kappa_0(u,v)}}\sqrt{\frac{1+\log p}{n}}\Big[(1+u)\sqrt{s}+(1+v)\Big]r,$$

then with probability at least $1 - \exp\left[-M_0(1+u)^2 s(1+\log p)/\kappa_0(u,v)\right]$, *we have*

$$e(u,v,r) \le 8\sqrt{\frac{2M_0}{\kappa_0(u,v)}}\sqrt{\frac{1+\log p}{n}}\Big[(1+u)\sqrt{s}+(1+v)\Big]r.$$

Proof of Lemma S3.2. First, note that the check loss $\rho_\tau(\cdot)$ is Lipschitz continuous with Lipschitz constant $\max(\tau,\, 1-\tau)$. Let $\delta = \alpha - \alpha^*$, $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, and define

$$\begin{aligned} U_i(\boldsymbol{\delta}, \boldsymbol{\Delta}) &= \rho_\tau(z_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta}) - \rho_\tau(z_i - \alpha^* - \mathbf{x}_i^\top \boldsymbol{\beta}^*) \\ &= \rho_\tau(r_i^* - \delta - \mathbf{x}_i^\top \boldsymbol{\Delta}) - \rho_\tau(r_i^*), \end{aligned}$$

where $r_i^* = z_i - \alpha^* - \mathbf{x}_i^\top \boldsymbol{\beta}^* = \varepsilon_i - \alpha^*$, $1 \le i \le n$. It follows immediately that

$$e(u,v,r) = \sup_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{G}_{u,v,r}} \left| \frac{1}{n}\sum_{i=1}^n \big[U_i(\delta, \boldsymbol{\Delta}) - \mathbb{E}U_i(\delta, \boldsymbol{\Delta})\big] \right|.$$

By Lipschitz continuity of the check loss, it follows that

$$\begin{aligned} |U_i(\delta, \boldsymbol{\Delta})| &\le |\rho_\tau(r_i^* - \delta - \mathbf{x}_i^\top \boldsymbol{\Delta}) - \rho_\tau(r_i^*)| \\ &\le \max(\tau, 1-\tau)|\delta + \mathbf{x}_i^\top \boldsymbol{\Delta}| \le |\delta + \mathbf{x}_i^\top \boldsymbol{\Delta}|,\ 1 \le i \le n. \end{aligned} \tag{8}$$

Applying Massart's concentration inequality Bühlmann and van de Geer (2011), we have

$$\Pr(e(u,v,r) \ge \mathbb{E}[e(u,v,r)] + t) \le \exp\left(-\frac{n^2 t^2}{8 b_n^2(u,v,r)}\right), \tag{9}$$

where $b_n^2(u,v,r) = \sup_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{G}_{u,v,r}} \sum_{i=1}^n \mathrm{var}(U_i(\delta, \boldsymbol{\Delta}))$. First, we derive an upper bound on $b_n^2(u,v,r)$. Note that by (8) and the Cauchy–Schwarz inequality

$$\begin{aligned} b_n^2(u,v,r) &= \sup_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{G}_{u,v,r}} \sum_{i=1}^n \mathbb{E}\big[U_i(\delta, \boldsymbol{\Delta}) - \mathbb{E}(U_i(\delta, \boldsymbol{\Delta}))\big]^2 \\ &\le 4 \sup_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{G}_{u,v,r}} \sum_{i=1}^n (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2 \le 4nr^2. \end{aligned}$$

Next, we show an upper bound on $\mathbb{E}\big[e(u,v,r)\big]$. Applying the symmetrization procedure van der Vaart and Wellner (1996) and the contraction principle Ledoux and Talagrand (1991), we have

$$\begin{aligned} \mathbb{E}[e(u,v,r)] &\le 2\mathbb{E}\left[\sup_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{G}_{u,v,r}} \frac{1}{n}\left|\sum_{i=1}^n \xi_i U_i(\delta, \boldsymbol{\Delta})\right|\right] \\ &\le \frac{2}{n}\mathbb{E}\left[\sup_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{G}_{u,v,r}} \left|\sum_{i=1}^n \xi_i \{\rho_\tau(r_i^* - \delta - \mathbf{x}_i^\top \boldsymbol{\Delta}) - \rho_\tau(r_i^*)\}\right|\right] \\ &\le \frac{4}{n}\mathbb{E}\left[\sup_{(\delta,\, \boldsymbol{\Delta}) \in \mathcal{G}_{u,v,r}} \left|\sum_{i=1}^n \xi_i (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})\right|\right], \end{aligned} \tag{10}$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables that are independent of $\varepsilon_1, \ldots, \varepsilon_n$ and $\Pr(\xi_i = \pm 1) = 0.5$.

For $(\delta,\, \boldsymbol{\Delta}) \in \mathcal{G}_{u,v,r}$, by condition (C1) and Cauchy–Schwarz inequality, we have

$$r^2 \geq \kappa_0(u,v)(\delta^2 + \|\boldsymbol{\Delta}_{\mathcal{A}}\|_2^2) \geq \kappa_0(u,v)\delta^2 + \frac{\kappa_0(u,v)}{s}\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1^2, \tag{11}$$

which implies that $|\delta| \leq r/\sqrt{\kappa_0(u,v)}$ and $\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 \leq r\sqrt{s/\kappa_0(u,v)}$. Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$. For any $t \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}\exp(tX_j^\top \boldsymbol{\xi}) &= \prod_{i=1}^n \Big[\frac{1}{2}(e^{tx_{ij}} + e^{-tx_{ij}})\Big] \\
&\leq \prod_{i=1}^n \exp\Big(\frac{1}{2}t^2 x_{ij}^2\Big) = \exp\Big(\frac{t^2}{2}\sum_{i=1}^n x_{ij}^2\Big),\ 0 \leq j \leq p.
\end{aligned}$$

Letting $t > 0$, by Jensen's inequality we have

$$\begin{aligned}
&\exp\big(t\mathbb{E}\big[\|\mathbb{X}^\top \boldsymbol{\xi}\|_\infty\big]\big) = \exp\Big(t\mathbb{E}\max_{0\leq j\leq p}|X_j^\top \boldsymbol{\xi}|\Big) \leq \mathbb{E}\exp\Big(t\max_{0\leq j\leq p}|X_j^\top \boldsymbol{\xi}|\Big) \\
&= \mathbb{E}\Big[\max_{0\leq j\leq p}\exp(t|X_j^\top \boldsymbol{\xi}|)\Big] \leq \mathbb{E}\max_{0\leq j\leq p}\Big(e^{tX_j^\top \boldsymbol{\xi}} + e^{-tX_j^\top \boldsymbol{\xi}}\Big) \\
&\leq \sum_{j=0}^p \mathbb{E}\Big(e^{tX_j^\top \boldsymbol{\xi}} + e^{-tX_j^\top \boldsymbol{\xi}}\Big) \leq 2\sum_{j=0}^p \exp\Big(\frac{t^2}{2}\|X_j\|_2^2\Big) \\
&\leq 2(1+p)\exp\Big(\frac{t^2}{2}\max_{0\leq j\leq p}\|X_j\|_2^2\Big) = 2(1+p)\exp\Big(\frac{nM_0}{2}t^2\Big),
\end{aligned}$$

which implies that

$$\mathbb{E}\big(\|\mathbb{X}^\top \boldsymbol{\xi}\|_\infty\big) \leq \frac{1}{t}\big[\log 2 + \log(1+p)\big] + \frac{nM_0}{2}t,\ t > 0.$$

Taking $t = \sqrt{2[\log 2 + \log(1+p)]/(nM_0)}$, we obtain

$$\mathbb{E}\big(\|\mathbb{X}^\top \boldsymbol{\xi}\|_\infty\big) \leq \sqrt{2nM_0[\log 2 + \log(1+p)]} \leq \sqrt{2M_0}\cdot\sqrt{n(1+\log p)} \tag{12}$$

as long as $p \geq 3$. It then follows from (10), (12) and the Hölder's inequality that

$$\begin{aligned}
\mathbb{E}[e(u,v,r)] &\leq \frac{4}{n}\mathbb{E}\big(\|\mathbb{X}^\top \boldsymbol{\xi}\|_\infty\big)\cdot \sup_{(\delta,\,\boldsymbol{\Delta})\in\mathcal{G}_{u,v,r}}\big(|\delta| + \|\boldsymbol{\Delta}\|_1\big) \\
&\leq \frac{4\sqrt{2M_0}}{n}\sqrt{n(1+\log p)}\sup_{(\delta,\,\boldsymbol{\Delta})\in\mathcal{G}_{u,v,r}}\big(|\delta| + \|\boldsymbol{\Delta}\|_1\big) \\
&\leq \frac{4\sqrt{2M_0}}{n}\sqrt{n(1+\log p)}\sup_{(\delta,\,\boldsymbol{\Delta})\in\mathcal{G}_{u,v,r}}\big[(1+v)|\delta| + (1+u)\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1\big] \\
&\leq 4\sqrt{\frac{2M_0}{\kappa_0(u,v)}}\sqrt{\frac{1+\log p}{n}}\Big[(1+u)\sqrt{s} + (1+v)\Big]r.
\end{aligned}$$

The lemma then follows from (9). $\square$

Lemma S3.3. *Under conditions (C1)–(C6), for any $(\delta, \boldsymbol{\Delta}) \in \mathcal{C}_{u,v}$, we have*

$$\begin{aligned}
&\mathbb{E}\big[Q_n(\alpha^* + \boldsymbol{\delta}, \boldsymbol{\beta}^* + \boldsymbol{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*)\big] \\
&\geq \min\Big\{\frac{\underline{f}}{4n}\sum_{i=1}^n(\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2,\ \underline{f}^{1/2}q\Big[\frac{1}{n}\sum_{i=1}^n(\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2\Big]^{1/2}\Big\}.
\end{aligned}$$

Proof of Lemma S3.3. By Knight's identity Knight (1998), we have for any two scalars r and s,

$$|r - s| - |r| = -s\big[I(r > 0) - I(r < 0)\big] + 2\int_0^s \big[I(r \leq t) - I(r \leq 0)\big]\mathrm{d}\,t.$$

It follows that for any $\tau \in (0,1)$,

$$\begin{aligned}
&\rho_\tau(r-s) - \rho_\tau(r) = (\tau - 0.5)\big[(r-s) - r\big] + 0.5\big[|r-s| - |r|\big] \\
&= (0.5-\tau)s - 0.5s\big[I(r>0) - I(r<0)\big] + \int_0^s \big[I(r \le t) - I(r \le 0)\big]\mathrm{d}\,t \\
&= s\big[I(r<0) - \tau\big] + \int_0^s \big[I(r \le t) - I(r \le 0)\big]\mathrm{d}\,t.
\end{aligned} \tag{13}$$

Let $r_i^* = z_i - \alpha^* - \mathbf{x}_i^\top \boldsymbol{\beta} = \varepsilon_i - \alpha^*$, $1 \le i \le n$. By (13) and the mean value theorem, we have for some $\bar{u}_{i,t}$ between 0 and t,

$$\begin{aligned}
&\mathbb{E}\big[Q_n(\alpha^* + \delta, \boldsymbol{\beta}^* + \boldsymbol{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*)\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\rho_\tau(r_i^* - \delta - \mathbf{x}_i^\top \boldsymbol{\Delta}) - \rho_\tau(r_i^*)\big] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\big[I(\varepsilon_i \le \alpha^*) - \tau\big] + \int_0^{\delta + \mathbf{x}_i^\top \boldsymbol{\Delta}} \big[I(\varepsilon_i \le \alpha^* + t) - I(\varepsilon_i \le \alpha^*)\big]\,\mathrm{d}t\right\} \\
&= \frac{1}{n}\sum_{i=1}^n \int_0^{\delta + \mathbf{x}_i^\top \boldsymbol{\Delta}} \big[F(\alpha^* + t) - F(\alpha^*)\big]\,\mathrm{d}t \\
&= \frac{1}{n}\sum_{i=1}^n \int_0^{\delta + \mathbf{x}_i^\top \boldsymbol{\Delta}} \Big[tf(\alpha^*) + \frac{t^2}{2} f'(\alpha^* + \bar{u}_{i,t})\Big]\,\mathrm{d}t \\
&\ge \frac{1}{2n}\sum_{i=1}^n f(\alpha^*)(\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2 - \frac{\bar{f}'}{6n}\sum_{i=1}^n |\delta + \mathbf{x}_i^\top \boldsymbol{\Delta}|^3 \\
&\ge \frac{1}{2n}\underline{f}\sum_{i=1}^n (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2 - \frac{\bar{f}'}{6n}\sum_{i=1}^n |\delta + \mathbf{x}_i^\top \boldsymbol{\Delta}|^3.
\end{aligned} \tag{14}$$

For $(\delta, \boldsymbol{\Delta}) \in \mathcal{C}_{u,v}$, note that if

$$\left[\frac{1}{n}\sum_{i=1}^n (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2\right]^{1/2} \le \frac{4}{\underline{f}^{1/2}} q, \tag{15}$$

then by condition (C6) we get

$$\frac{\bar{f}'}{6n}\sum_{i=1}^n |\delta + \mathbf{x}_i^\top \boldsymbol{\Delta}|^3 \le \frac{1}{4n}\underline{f}\sum_{i=1}^n (\delta + \mathbf{x}_i^\top)^2,$$

which, together with (14), implies that for all $(\boldsymbol{\delta}, \boldsymbol{\Delta}) \in \mathcal{G}_{u,\,v,\,4\underline{f}^{-1/2}q}$,

$$\mathbb{E}\big[Q_n(\alpha^* + \delta, \boldsymbol{\beta}^* + \boldsymbol{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*)\big] \ge \frac{\underline{f}}{4n}\sum_{i=1}^n (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2.$$

When (15) does not hold, one can similarly apply the technique in the proof of Lemma 4 of Belloni and Chernozhukov (2011) to show that for any $(\boldsymbol{\delta}, \boldsymbol{\Delta}) \in \mathcal{C}_{u,v}$,

$$\begin{aligned}
&\mathbb{E}\big[Q_n(\alpha^* + \delta, \boldsymbol{\beta}^* + \boldsymbol{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*)\big] \\
&\ge \min\left\{\frac{\underline{f}}{4n}\sum_{i=1}^n (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2,\ \underline{f}^{1/2} q\left[\frac{1}{n}\sum_{i=1}^n (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2\right]^{1/2}\right\}.
\end{aligned}$$

This completes the lemma. ☐

Proof of Theorem 3.1. Let $\mathcal{G}^* = \{(\delta, \boldsymbol{\Delta}) \in \mathcal{C}(3,1) \colon n^{-1}\sum_{i=1}^n (\delta + \mathbf{x}_i^\top \boldsymbol{\Delta})^2 = r_*^2\}$, where

$$r_* = 8\underline{f}^{-1}\left[16\sqrt{\frac{2M_0}{\kappa_0(3,1)}}\sqrt{\frac{1+\log p}{n}}(2\sqrt{s}+1) + \lambda\sqrt{\frac{s}{\kappa_0(3,1)}}\right].$$

Moreover, let $\widehat{\delta}^{\lambda} = \widehat{\alpha}_{\lambda} - \alpha^*$ and $\widehat{\mathbf{\Delta}}^{\lambda} = \widehat{\boldsymbol{\beta}}_{\lambda} - \boldsymbol{\beta}^*$. If we can show that

$$\min_{(\delta,\,\mathbf{\Delta})\in\mathcal{G}^*} Q_n(\alpha^*+\delta, \boldsymbol{\beta}^*+\mathbf{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*) + \lambda(\|\boldsymbol{\beta}^*+\mathbf{\Delta}\|_1 - \|\boldsymbol{\beta}^*\|_1) > 0, \tag{16}$$

then by convexity of Q_n, this implies that $n^{-1}\sum_{i=1}^n(\hat{\delta}^{\lambda} + \mathbf{x}_i^\top\widehat{\mathbf{\Delta}}^{\lambda})^2 \leq r_*^2$ under the event $\mathcal{E} = \{(\widehat{\delta}^{\lambda}, \widehat{\mathbf{\Delta}}^{\lambda}) \in \mathcal{C}(3,1)\}$. To show (16), first note that by Lemma S3.2, with probability at least $1-\exp\big[-16M_0 s(1+\log p)/\kappa_0(3,1)\big]$, we have for all $(\delta, \mathbf{\Delta}) \in \mathcal{G}^*$,

$$\begin{aligned}
&Q_n(\alpha^*+\delta, \boldsymbol{\beta}^*+\mathbf{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*) + \lambda(\|\boldsymbol{\beta}^*+\mathbf{\Delta}\|_1 - \|\boldsymbol{\beta}^*\|_1)\\
&\geq \mathbb{E}\big[Q_n(\alpha^*+\delta, \boldsymbol{\beta}^*+\mathbf{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*)\big] - e(3,1,r_*) + \lambda\Bigg(\sum_{j\in\mathcal{A}^{\mathrm{c}}}|\Delta_j| - \sum_{j\in\mathcal{A}}|\Delta_j|\Bigg)\\
&\geq \mathbb{E}\big[Q_n(\alpha^*+\delta, \boldsymbol{\beta}^*+\mathbf{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*)\big] + \lambda\Bigg(\sum_{j\in\mathcal{A}^{\mathrm{c}}}|\Delta_j| - \sum_{j\in\mathcal{A}}|\Delta_j|\Bigg)\\
&\quad - 16\sqrt{\frac{2M_0}{\kappa_0(3,1)}}\sqrt{\frac{1+\log p}{n}}\big(2\sqrt{s}+1\big)r_*.
\end{aligned} \tag{17}$$

On the one hand, by Lemma S3.3, for any $(\delta, \mathbf{\Delta}) \in \mathcal{G}^*$, we have

$$\mathbb{E}\big[Q_n(\alpha^*+\delta, \boldsymbol{\beta}^*+\mathbf{\Delta}) - Q_n(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\big] \geq \min\big\{\underline{f} r_*^2/4,\ \underline{f}^{1/2} q r_*\big\}.$$

On the other hand, by condition (C1) and (11), we can see that

$$\|\mathbf{\Delta}_{\mathcal{A}}\|_1 \leq r_*\sqrt{s/\kappa_0(3,1)}.$$

Thus, it follows from (17) and the growth condition that

$$\begin{aligned}
&Q_n(\alpha^*+\delta, \boldsymbol{\beta}^*+\mathbf{\Delta}) - Q_n(\alpha^*, \boldsymbol{\beta}^*) + \lambda(\|\boldsymbol{\beta}^*+\mathbf{\Delta}\|_1 - \|\boldsymbol{\beta}^*\|_1)\\
&\geq \frac{\underline{f}}{4}r_*^2 - \Bigg[16\sqrt{\frac{2M_0}{\kappa_0(3,1)}}\sqrt{\frac{1+\log p}{n}}\big(2\sqrt{s}+1\big) + \lambda\sqrt{s/\kappa_0(3,1)}\Bigg]r_* > 0
\end{aligned}$$

for all $(\delta, \mathbf{\Delta}) \in \mathcal{G}^*$ by our choice of r_*. By Lemma S3.1 and convexity of Q_n, this implies that with probability at least

$$\Pr(\mathcal{E}) - \exp\big[-16M_0 s(1+\log p)/\kappa_0(3,1)\big] \geq 1 - p_1(\lambda),$$

we have $(\widehat{\delta}^{\lambda}, \widehat{\mathbf{\Delta}}^{\lambda}) \in \mathcal{C}(3,1)$ and

$$n^{-1}\sum_{i=1}^n(\hat{\delta}^{\lambda} + \mathbf{x}_i^\top\widehat{\mathbf{\Delta}}^{\lambda})^2 \leq r_*^2.$$

This, by condition (C1), further implies that

$$r_*^2 \geq \kappa_m(3,1)\Big[|\hat{\delta}^{\lambda}|^2 + \|\widehat{\mathbf{\Delta}}^{\lambda}_{\mathcal{A}\cup\overline{\mathcal{A}}(\widehat{\mathbf{\Delta}}^{\lambda},m)}\|_2^2\Big].$$

As a result, we obtain that

$$|\widehat{\delta}^{\lambda}| \leq \frac{r_*}{\sqrt{\kappa_m(3,1)}}$$

and that

$$\|\widehat{\mathbf{\Delta}}^{\lambda}_{\mathcal{A}\cup\overline{\mathcal{A}}(\widehat{\mathbf{\Delta}}^{\lambda},m)}\|_2 \leq \frac{r_*}{\sqrt{\kappa_m(3,1)}}. \tag{18}$$

Note that the jth largest in absolute value component of $\widehat{\mathbf{\Delta}}_{\mathcal{A}^{\mathrm{c}}}$ is bounded by $\|\widehat{\mathbf{\Delta}}_{\mathcal{A}^{\mathrm{c}}}\|_1/j$. Therefore, it follows that

$$\begin{aligned}
&\Big\|\widehat{\mathbf{\Delta}}_{(\mathcal{A}\cup\overline{\mathcal{A}}(\widehat{\mathbf{\Delta}}^{\lambda},m))^{\mathrm{c}}}\Big\|_2^2 \leq \sum_{j=m+1}^{p}\frac{\|\widehat{\mathbf{\Delta}}^{\lambda}_{\mathcal{A}^{\mathrm{c}}}\|_1^2}{j^2} \leq \frac{1}{m}\|\widehat{\mathbf{\Delta}}^{\lambda}_{\mathcal{A}^{\mathrm{c}}}\|_1^2\\
&\leq \frac{1}{m}\big[3\|\widehat{\mathbf{\Delta}}^{\lambda}_{\mathcal{A}}\|_1 + |\widehat{\delta}^{\lambda}|\big]^2 \leq \frac{18s}{m}\|\widehat{\mathbf{\Delta}}^{\lambda}_{\mathcal{A}}\|_2^2 + \frac{2}{m}|\widehat{\delta}^{\lambda}|^2\\
&\leq \frac{18s}{m}\|\widehat{\mathbf{\Delta}}^{\lambda}_{\mathcal{A}\cup\overline{\mathcal{A}}(\widehat{\mathbf{\Delta}}^{\lambda},m)}\|_2^2 + \frac{2}{m}|\widehat{\delta}^{\lambda}|^2,
\end{aligned}$$

which implies that

$$\|\widehat{\mathbf{\Delta}}^{\lambda}\|_2^2 \leq \Big(1+\frac{18s}{m}\Big)\|\widehat{\mathbf{\Delta}}^{\lambda}_{\mathcal{A}\cup\overline{\mathcal{A}}(\widehat{\mathbf{\Delta}}^{\lambda},m)}\|_2^2 + \frac{2}{m}|\widehat{\delta}^{\lambda}|^2$$
$$\leq \frac{r_*^2}{\kappa_m(3,1)}\Big(1+\frac{18s}{m}+\frac{2}{m}\Big).$$

This completes the proof of Theorem 3.1. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: The claims presented in the abstract and introduction accurately represent the contributions and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: We discuss the limitations of our work in the last section (Section 5) of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes] .

Justification: The assumptions needed for theoretical results are included in the Section 3, and complete proof of each result is given in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] .

   Justification: We have included details of the experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes] .

Justification: The implementation of QuanDA is available at `https://anonymous.4open.science/status/QuanDA-57FE`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: We have included the details of experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes] .

   Justification: For the experiments on synthetic data, we report error bars and the experimental settings for the random data.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] .

   Justification: We have included the details of the computational resources in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

   Justification: We have conformed to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: Our work does not have a direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The models or the data used in the paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: Our work does not use existing assets.881

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: Our work does not involve research with human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: Our work does not involve research with human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: Our work does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.