Received: 3 January 2024 | Revised: 4 March 2024 | Accepted: 11 May 2024

DOI: 10.1111/stan.12342

ORIGINAL ARTICLE

WILEY

High-dimensional sparse classification using exponential weighting with empirical hinge loss

The Tien Mai

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
The Tien Mai, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim 7034, Norway.
Email: the.t.mai@ntnu.no

Funding information
Norges Forskningsråd, Grant/Award Number: 309960

In this study, we address the problem of high-dimensional binary classification. Our proposed solution involves employing an aggregation technique founded on exponential weights and empirical hinge loss. Through the employment of a suitable sparsity-inducing prior distribution, we demonstrate that our method yields favorable theoretical results on prediction error. The efficiency of our procedure is achieved through the utilization of Langevin Monte Carlo, a gradient-based sampling approach. To illustrate the effectiveness of our approach, we conduct comparisons with the logistic Lasso on simulated data and a real dataset. Our method frequently demonstrates superior performance compared to the logistic Lasso.

KEYWORDS

binary classification, high-dimensionality, Langevin Monte Carlo, PAC-Bayesian inequalities, prediction error, sparsity

1 | INTRODUCTION

Classification in high-dimensional scenarios, where the number of potential explanatory variables (predictors) p significantly exceeds the sample size n, presents a fundamental challenge that transcends disciplines such as statistics and machine learning (Bühlmann & Van De Geer, 2011; Hastie, Tibshirani, Friedman, & Friedman, 2009; Fan, Fan, & Wu, 2010; Giraud, 2021).

This issue holds considerable relevance across various domains, including applications such as disease classification (Chung & Keles, 2010), document classification (Kotte, Rajavelu, & Rajsingh, 2020), and image recognition (Li, Chai, Zhou, & Yin, 2021). The setting of large p, small n introduces a significant challenge known as the "curse of dimensionality." The works in Bickel and Levina (2004) and Fan and Fan (2008) highlighted that, even in simple cases, high-dimensional classification without feature selection can perform as poorly as random guessing. Consequently, the imperative arises to mitigate this issue by reducing feature space dimensionality through the judicious selection of a sparse subset of "meaningful" features.

Numerous methodologies have been suggested to address the challenge of classification in high-dimensional settings, as discussed in works such as Fan et al. (2010) and Giraud (2021). The majority of these approaches center around penalized maximum likelihood estimation. Notably, the statistical package "`glmnet`" (Friedman, Hastie, & Tibshirani, 2010) has successfully implemented the Lasso and elastic net for generalized linear models, showcasing practical effectiveness. In a more recent study of Abramovich and Grinshtein (2018), the authors establish nonasymptotic bounds on misclassification excess risk for procedures based on penalized maximum likelihood. However, probabilistic approaches have received comparatively less attention in tackling this problem.

Diverging from traditional approaches centered on parametric models, we adopt an alternative strategy that involves considering a set of classifiers and selecting the one that yields the best prediction error. This approach is rooted in the principles of statistical learning theory Vapnik, 1998, where the zero-one loss is employed as a measure of prediction error, and the classifier's risk is governed by a PAC (probably approximately correct) bound. Our novel approach combines elements from both Bayesian and machine learning methodologies. More specifically, we consider a pseudo-Bayesian strategy that incorporates a risk concept based on the hinge loss instead of relying on a likelihood function. Because of the computational challenges arising from the nonconvexity of the zero-one loss function, the hinge loss serves as a suitable alternative Zhang, 2004. The hinge loss is well-known for its effectiveness in diverse machine learning tasks and computational efficiency. It is noteworthy that the substitution of loss functions for likelihood has gained popularity in Generalized Bayesian inference in recent years, as evidenced by works such as Matsubara, Knoblauch, Briol, and Oates (2022), Jewson and Rossell (2022), Yonekura and Sugasawa (2023), Medina, Olea, Rush, and Velez (2022), Grünwald and Van Ommen (2017), Bissiri, Holmes, and Walker (2016), Lyddon, Holmes, and Walker (2019), Syring and Martin (2019), Knoblauch, Jewson, and Damoulas (2022), and Hong and Martin (2020).

The foundation of our theoretical findings regarding prediction errors relies on the PAC-Bayes bound technique. Initially introduced by McAllester (1998) and Shawe-Taylor and Williamson (1997) to furnish numerical generalization certificates for Bayesian-flavored machine learning algorithms, this technique took a broader applicability turn when Catoni (2004, 2007) realized its utility in proving oracle inequalities and rates of convergence for (generalized)-Bayesian estimators in statistics. This methodology shares strong connections with the "information bounds" presented by Zhang (2006) and Russo and Zou (2019). For an in-depth exploration of this topic, we recommend referring to Guedj (2019) and Alquier (2024). PAC-Bayes bounds have been instrumental in establishing oracle inequalities in various problems, as evidenced by works like Seeger (2002), Langford and Shawe-Taylor (2002), Herbrich and Graepel (2002), Maurer (2004), Dalalyan and Tsybakov (2008), Seldin, Laviolette, Cesa-Bianchi,

Shawe-Taylor, and Auer (2012), Alquier, Ridgway, and Chopin (2016), Seldin and Tishby (2010), Germain, Lacasse, Laviolette, March, and Roy (2015), Mai and Alquier (2015, 2017), and Cottet and Alquier (2018). Our methodology aligns with the principles of PAC-Bayesian theory, offering robust theoretical assurances regarding prediction error for our proposed method.

Utilizing a loss function based on hinge loss offers a solution to overcome certain constraints inherent in traditional likelihood-based Bayesian models, especially in the context of binary response variables. The convex nature of the hinge loss function facilitates ease of optimization. Additionally, our method incorporates a smooth sparsity-promoting prior, previously explored in Dalalyan and Tsybakov (2012a, 2012b). These features enhance the efficiency of our approach, making it amenable to implementation using Langevin Monte Carlo (LMC) method. We advocate a LMC approach for the computation of our proposed method, an emerging technique in high-dimensional Bayesian methods (Dalalyan, 2017; Dalalyan & Riou-Durand, 2020; Durmus & Moulines, 2017, 2019). The LMC method originated in physics with Langevin diffusions (Ermak, 1975) and gained popularity in statistics and machine learning after the seminal paper by Roberts and Tweedie (1996).

Apart from a thorough theoretical investigation, we complement our study by undertaking a comprehensive series of simulations to assess the numerical performance of the method we propose. In the realm of numerical comparisons, our methodology demonstrates comparable outcomes when contrasted with both the logistic Lasso and the Bayesian logistic approach. The outcomes of our simulations reveal noteworthy insights. Notably, our proposed method exhibits a heightened level of robustness in the face of varying sample sizes and sparsity levels, outperforming the logistic Lasso in these scenarios. Furthermore, to validate the practical utility of our method, we conduct an application using a real dataset. The results obtained from this real data example align closely with those derived from the logistic Lasso, emphasizing the consistency and applicability of our proposed method across different datasets. This multifaceted evaluation, on both simulated and real data, collectively underscores the effectiveness and reliability of our proposed method in the realm of high-dimensional classification problems.

The subsequent sections of this paper are structured as outlined below. Section 2 provides an introduction to both the high-dimensional classification problem and the method we propose to address it. In Section 3, we consolidate our theoretical analysis, specifically focusing on the prediction error associated with our proposed method. Section 4 is dedicated to the presentation and discussion of our simulation studies and a real data application. Our discussion on the findings and conclusions of our work unfold in Section 5. For the technical proofs underpinning our analyses, interested readers can refer to Appendix A.

2 | PROBLEM AND METHOD

2.1 | Problem statement

We formally consider the following general binary classification with a high-dimensional vector of features $x \in \mathbb{R}^d$ and the outcome class label

$$Y|x = \begin{cases} 1, & \text{with probability } p(x), \\ -1, & \text{with probability } 1 - p(x) \end{cases}.$$

The accuracy of a classifier η is defined by the prediction error, given as

$$R(\eta) = \mathbb{P}(Y \neq \eta(x)).$$

It is well-known that $R(\eta)$ is minimized by the Bayes classifier $\eta^*(x) = \text{sign}(p(x) - 1/2)$ (Vapnik, 1998; Devroye, Györfi, & Lugosi, 1996), that is,

$$R(\eta^*) = \inf\ R(\eta).$$

However, the probability function $p(x)$ is unknown and the resulting classifier $\hat{\eta}(x)$ should be designed from the data D_n: a random sample of n independent observations $(x_1, Y_1), \ldots, (x_n, Y_n)$, with $n < d$. The design points x_i may be considered as fixed or random. The corresponding (conditional) prediction error of $\hat{\eta}$ is

$$R(\hat{\eta}) = \mathbb{P}(Y \neq \hat{\eta}(x)\ |D_n),$$

and the goodness of $\hat{\eta}$ w.r.t. η^* is measured by the excess risk, that is, $\mathbb{E}\ R(\hat{\eta}) - R(\eta^*)$. One could obtain $\hat{\eta}$ by estimating $p(x)$ from the data by some $\hat{p}(x)$ and use a plug-in classifier of the form $\hat{\eta}(x) = \text{sign}(\hat{p}(x) - 1/2)$. A standard approach is to consider one of the most commonly used models—logistic regression, where it is assumed that $p(x) = 1/(1 + e^{-\beta^\top x})$ and $\beta \in \mathbb{R}^d$ is a vector of unknown regression coefficients. The corresponding Bayes classifier is a linear classifier, $\eta^*(x) = \text{sign}(\beta^{*\top} x)$. One then estimates β^* from the data to get $\hat{\beta}$ (e.g., using maximum likelihood), and the resulting linear classifier is $\hat{\eta}(x) = \text{sign}(\hat{\beta}^\top x)$, see for example, Abramovich and Grinshtein (2018).

Another common general (nonparametric) approach for finding a classifier $\hat{\eta}$ from the data is empirical risk minimization, where minimization of a true prediction error $R(\eta)$ is replaced by minimization of the corresponding empirical risk over a given class of classifiers. Consider the class of linear classifiers, the empirical risk is given by:

$$r_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i(\beta^\top x_i) < 0\}.$$

The ability of the classifier to predict a new label given feature x is then assessed by the prediction error

$$R(\beta) = \mathbb{E}\big[\mathbb{1}\{Y(\beta^\top x) < 0\}\big].$$

For the sake of simplicity, we put $R^* := R(\beta^*)$, where β^* is the ideal Bayes classifier.

In this paper, we consider a sparse setting and thus we assume that $s^* < n$ where $s^* = \|\beta^*\|_0$ (the number of nonzero entries).

The main goal in this work is to develop a classifier $\hat{\beta}$ such that its prediction error will be close the ideal Bayes error R^*.

2.2 | The proposed method

The explanation of the prediction risk, represented by $R(\beta)$, is clear; however, managing its empirical equivalent, $r_n(\beta)$, presents computational challenges due to its nonsmooth and nonconvex

characteristics. To tackle this issue, a frequently used approach is to replace the empirical risk with an alternative convex surrogate, as proposed in earlier studies Zhang (2004) and Bartlett, Jordan, and McAuliffe (2006).

In this paper, we primarily focus on the hinge loss, which results in the following hinge empirical risk:

$$r_n^h(\beta) = \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\ (\beta^\top x_i))_+\ ,$$

where $(a)_+ := \max(a, 0), \forall a \in \mathbb{R}$.

We consider an exponentially weighted aggregate (EWA) procedure and define the following pseudo-posterior distribution:

$$\hat{\rho}_\lambda(\beta) \propto \exp[-\lambda r_n^h(\beta)]\pi(\beta), \tag{1}$$

where $\lambda > 0$ is a tuning parameter that will be discussed later and $\pi(\beta)$ is a prior distribution, given in (2), that promotes (approximately) sparsity on the parameter vector β.

The EWA procedure has found application in various contexts in prior works (Dalalyan, Grappin, & Paris, 2018; Dalalyan and Tsybakov (2008, 2012b); Dalalyan (2020)). The term $\hat{\rho}_\lambda$ is also referred to as the Gibbs posterior (Alquier et al., 2016; Catoni, 2007). The incorporation of $\hat{\rho}_\lambda$ is driven by the minimization problem presented in Lemma 1, rather than strictly adhering to conventional Bayesian principles. Notably, there is no necessity for a likelihood function or a complete model; only the empirical risk based on the hinge loss function is crucial. This approach aligns with the evolving trend in the Generalized Bayesian method in contemporary literature, where the likelihood is often replaced with a power version or a loss-based method, as seen in works such as Bissiri et al. (2016), Knoblauch et al. (2022), Grünwald and Van Ommen (2017), Hong and Martin (2020), and Matsubara et al. (2022).

However, in this manuscript, we consistently denote π as the prior and $\hat{\rho}_\lambda$ as the pseudo-posterior. The rationale behind the EWA can be summarized as follows: when comparing two parameters, b_1 and b_2, if $r_n^h(b_1) < r_n^h(b_2)$, then $\exp[-\lambda r_n^h(b_1)] > \exp[-\lambda r_n^h(b_2)]$ for any $\lambda > 0$. This implies that, in comparison to π, $\hat{\rho}_\lambda$ assigns more weight to the parameter with a smaller hinge empirical risk. Consequently, the adjustment in the distribution favors the parameter value associated with a smaller in-sample hinge empirical risk. The tuning parameter λ dictates the degree of this adjustment, and its selection will be further investigated in subsequent sections.

2.3 | A sparsity-inducing prior

Given a positive number C_1, for all $\beta \in B_1(C_1) := \{\beta \in \mathbb{R}^d : \|\beta\|_1 \le C_1\}$, we consider the following prior,

$$\pi(\beta) \propto \prod_{i=1}^{d}\frac{1}{(\tau^2 + \beta_i^2)^2}, \tag{2}$$

where $\tau > 0$ is a tuning parameter. For technical reason, we assume that $C_1 > 2d\tau$.

Initially, it is noteworthy that C_1 serves as a regularization constant, typically assumed to be very large. Consequently, π essentially takes the form of a product of d rescaled Student's distributions. To be more precise, the distribution of π closely approximates that of $S\tau\sqrt{2}$, where S denotes a random vector with independent and identically distributed (i.i.d) components drawn from the Student's t-distribution with 3 degrees of freedom. One can choose a very small τ, smaller than $1/n$, resulting in the majority of components in τS being in close proximity to zero. However, owing to the heavy-tailed nature of the Student's t-distribution, a few components of τS are significantly distant from zero. This particular characteristic imparts the prior with the ability to encourage sparsity.

It is worth noting that this type of prior has been previously examined in the context of aggregating estimators Dalalyan and Tsybakov (2012a, 2012b). In this work, we further investigate the applicability of this prior in sparse classification, specifically with hinge empirical loss. Moreover, various authors have underscored the significance of heavy-tailed priors in addressing sparsity, see for example, Seeger (2008), Johnstone and Silverman (2004), Rivoirard (2006), Abramovich, Grinshtein, and Pensky (2007), Carvalho, Polson, and Scott (2010), Castillo and van der Vaart (2012), and Castillo and Mismer (2018).

3 | THEORETICAL RESULTS

Let $r_n^* := r_n(\beta^*)$. We will require the following assumptions in order to state our main results.

Assumption 1. We assume that there is a constant $C_{\mathrm{x}} > 0$ such that $\sum_{i=1}^n \|x_i\|_2/n \leq C_{\mathrm{x}}$.

Assumption 2. We assume that there is a constant $C' > 0$ such that $r_n^h(\beta^*) \leq (1 + C')r_n^*$.

Remark 1. Assumption 1 on the design matrix above is less stringent when contrasted with those outlined in the reference Abramovich and Grinshtein (2018). More specifically, the Weighted Restricted Eigenvalue condition, a requirement concerning the design matrix to obtain the result for the logistic Slope in Abramovich and Grinshtein (2018), is not required for the results presented in our study. Conversely, Assumption 2 may be regarded as analogous to conditions required in Abramovich and Grinshtein (2018), where an upper bound on $|\beta^{*\top} x_i|$ for all $i = 1, \ldots, n$ is required.

3.1 | Bounds on prediction risk

Theorem 1. *Assume that Assumptions 1 and 2 hold. We have, for* $\lambda = \sqrt{n\log(nd)}$, $\tau = 1/(n\sqrt{d})$ *and with probability at least* $1 - 2\epsilon, \epsilon \in (0,1)$, *that for all* β^* *such that* $\|\beta^*\|_1 \leq C_1 - 2d\tau$ *and* $\|\beta^*\|_0 \leq s^*$,

$$\int R d\hat{\rho}_\lambda \leq (1 + 2C')R^* + c\frac{s^*\sqrt{\log(n\sqrt{d}/s^*)}}{\sqrt{n}} + \frac{\log(1/\epsilon)}{\sqrt{n\log(nd)}},$$

where c is a universal constant depending only on C', C_1, C_x.

The proof for the aforementioned theorem and subsequent results can be found in Appendix A, where we employ the "PAC-Bayesian bounds" technique from Catoni (2007) as our primary technical arguments. Initially introduced in Shawe-Taylor and Williamson (1997); McAllester (1998), PAC-Bayesian bounds serve as a method to offer empirical bounds on the prediction risk of Bayesian-type estimators. However, as extensively discussed in Catoni (2003, 2004, 2007), this technique also provides a set of powerful technical tools for establishing nonasymptotic bounds. For a thorough exploration of PAC-Bayes bounds, along with recent surveys and advancements, readers are encouraged to consult the following references: Guedj (2019); Alquier (2024).

Remark 2. Theorem 1 establishes a connection between the integrated prediction risk of our approach and the minimum achievable risk, attained by the Bayes classifier β^*. The assumption regarding the boundedness of the parameter significantly influences our technical proofs, a common feature in PAC-Bayes literature, but it could potentially be mitigated through alternative methodologies as suggested by Alquier and Ridgway (2020); Alquier (2024).

In addition to the outcome outlined in Theorem 1, we can derive a result for, $\hat{\beta} \sim \hat{\rho}_\lambda$, a stochastic classifier sampled from our suggested pseudo-posterior (1). The following result is occasionally referred to as the contraction rate of the pseudo-posterior.

Theorem 2. *Under the same assumptions for Theorem 1, and the same definition for* τ *and* λ, *let* ε_n *be any sequence in* $(0,1)$ *such that* $\varepsilon_n \to 0$ *when* $n \to \infty$. *Define*

$$\Theta_n = \left\{ \beta \in \mathbb{R}^d \,:\, R \le (1+2C')R^* + c\frac{s^*\sqrt{\log(n\sqrt{d}/s^*)}}{\sqrt{n}} + \frac{\log(1/\varepsilon_n)}{\sqrt{n\log(nd)}} \right\}.$$

Then

$$\mathbb{E}\left[\mathbb{P}_{\hat{\beta}\sim\hat{\rho}_\lambda}(\hat{\beta} \in \Theta_n)\right] \ge 1 - 2\varepsilon_n \xrightarrow[n\to\infty]{} 1.$$

The primary challenges for any classifier manifest in the vicinity of the boundary $\{x : p(x) = 1/2\}$, or equivalently, a hyperplane $\beta^\top x = 0$ for the logistic regression model, where accurate prediction of the class label becomes particularly challenging. However, in regions where $p(x)$ is sufficiently away from $1/2$, referred to as the margin or low-noise condition, there exists potential for improving the bounds on prediction risk. The improvement of the obtained risk bounds is done under the additional low-noise or margin assumption.

3.2 | Improved bounds under the margin condition

In this work, we make use of the following margin assumption as in Mammen and Tsybakov (1999), see also Tsybakov (2004); Bartlett et al. (2006).

Assumption 3 (Low-noise/Margin assumption). We assume that there is a constant $C \geq 1$ such that:

$$\mathbb{E}\left[\left(\mathbb{1}_{Y(\beta^\top x)\leq 0} - \mathbb{1}_{Y(\beta^{*\top} x)\leq 0}\right)^2\right] \leq C[R(\beta) - R^*].$$

Theorem 3. *Assume that Assumptions 1, 2, 3 hold. We have, for* $\lambda = 2n/(3C+2)$, $\tau = 1/(n\sqrt{d})$ *and with probability at least* $1-2\epsilon, \epsilon \in (0,1)$, *that for all* β^* *such that* $\|\beta^*\|_1 \leq C_1 - 2d\tau$ *and* $\|\beta^*\|_0 \leq s^*$,

$$\int R d\hat{\rho}_\lambda \leq (1+3C')R^* + C_{C,C_1,C',C_x}\frac{s^*\log(n\sqrt{d}/s^*) + \log(1/\epsilon)}{n},$$

where C_{C,C_1,C',C_x} *is a universal constant depending only on* C, C', C_1, C_x.

The proof can be found in Appendix A.

Remark 3. The prediction bounds derived in Theorems 3 and 1 represent novel contributions to the field. These bounds explicitly depend on s^*, signifying the adaptability of our method in scenarios characterized by sparsity. It is crucial to highlight that the outcomes of these main theorems exhibit adaptive characteristics, indicating that the estimator's performance is independent of s^*, the sparsity of β^*. In instances where the true sparsity s^* is very small, the prediction error aligns closely with the Bayes error, denoted as R^*, even when dealing with a relatively small sample size. This outcome is commonly denoted as an "oracle inequality," suggesting that our estimator performs comparably to a scenario where knowledge of the sparsity of β^* is accessible through an oracle.

Remark 4. Compared to Theorem 1, the bound in Theorem 3 is faster and of order $1/n$ rather than $1/\sqrt{n}$. These bounds allow to compare the out-of-sample error of our method to the optimal one, R^*.

Let us now consider the noiseless case where $Y = \text{sign}(\beta^{*\top}x)$ almost surely. Then, $R^* = 0$ and we have that

$$\mathbb{E}\left[\left(\mathbb{1}_{Y(\beta^\top x)\leq 0} - \mathbb{1}_{Y(\beta^{*\top} x)\leq 0}\right)^2\right] = \mathbb{E}\left[\mathbb{1}^2_{Y(\beta^\top x)\leq 0}\right] = \mathbb{E}\left[\mathbb{1}_{Y(\beta^\top x)\leq 0}\right] = R(\beta) = R(\beta) - R^*.$$

Thus, the margin assumption is satisfied with $C = 1$. We now state a corollary in the noiseless case for Theorem 3.

Corollary 1. *In the case of noiseless, that is,* $Y = \text{sign}(\beta^{*\top}x)$, *we have, for* $\lambda = 2n/5$, $\tau = 1/(nd)$ *and with probability at least* $1-2\epsilon, \epsilon \in (0,1)$, *that for all* β^* *such that* $\|\beta^*\|_1 \leq C_1 - 2d\tau$ *and* $\|\beta^*\|_0 \leq s^*$,

$$\int R d\hat{\rho}_\lambda \leq C'\frac{s^*\log\left(\frac{n\sqrt{d}}{s^*}\right) + \log(1/\epsilon)}{n},$$

where $C' := C_{1,C_1,C',C_x}$ *is a universal constant depending only on* C_1, C', C_x.

Remark 5. According to Corollary 1, the bound on the misclassification excess risk for our proposed method follows an order of $s^*\log(de/s^*)/n$. Meanwhile, under

Assumption 3, the study in Abramovich and Grinshtein (2018) established a minimax lower bound for the misclassification excess risk, which is of the order $s^* \log(de/s^*)/n$. Notably, this lower bound is also achieved by the logistic Slope estimator in that same paper under additional Weighted Restricted Eigenvalue condition. As a result, in the noiseless case, our rate is demonstrated to be minimax-optimal in a setting that $n \leq e\sqrt{d}$, by noting that $\frac{n\sqrt{d}}{s^*} = \frac{n}{e\sqrt{d}}\frac{de}{s^*}$.

In analogy to Theorem 2, given additional Assumption 3, we can establish that a stochastic classifier, $\hat{\beta} \sim \hat{\rho}_\lambda$, drawn from our proposed pseudo-posterior in Equation (1) exhibits a fast rate. This particular finding is sometimes denominated as the contraction rate of the pseudo-posterior.

Theorem 4. *Under the same assumptions for Theorem 3, and the same definition for τ and λ, let ε_n be any sequence in $(0,1)$ such that $\varepsilon_n \to 0$ when $n \to \infty$. Define*

$$\Omega_n = \left\{ \beta \in \mathbb{R}^d \, : \, R \leq (1+3C')R^* + C_{C,C_1,C',C_x} \frac{s^* \log(n\sqrt{d}/s^*) + \log(1/\varepsilon_n)}{n} \right\}.$$

Then

$$\mathbb{E}\left[\mathbb{P}_{\hat{\beta}\sim\hat{\rho}_\lambda}(\hat{\beta} \in \Omega_n)\right] \geq 1 - 2\varepsilon_n \xrightarrow[n\to\infty]{} 1.$$

The results presented in Theorems 2 and 4 are also novel findings, to the best of our knowledge.

Remark 6. In this section, we demonstrate the existence of specific values for the tuning parameters λ and τ in our proposed method in theoretical results for prediction errors. It is important to acknowledge, however, that these values may not be the most suitable for practical applications. In practical applications, cross-validation can be employed to appropriately fine-tune these parameters. Nevertheless, the theoretical values identified in our analysis provide valuable insights into the expected magnitude of these tuning parameters when applied in practical situations.

3.3 | Sharp rates with known sparsity s^*

In the present section, we operate under the assumption that s^*, denoting the number of nonzero coefficients in β^*, is a known quantity. This assumption allows us to obtain results that sharply align with the rates established in Abramovich and Grinshtein (2018) (in the noiseless case). The first refinement is observed in the context of Theorem 1, where the tuning parameters λ and τ are intricately dependent on the specific value of s^*.

Proposition 1. *Assume that Assumptions 1 and 2 hold. We have, for $\tau = s^*/(n\sqrt{d})$ and $\lambda = \sqrt{ns^* \log(de/s^*)}$, with probability at least $1-2\epsilon$, $\epsilon \in (0,1)$, that for all β^* such that $\|\beta^*\|_1 \leq C_1 - 2d\tau$ and $\|\beta^*\|_0 \leq s^*$,*

$$\int R d\hat{\rho}_\lambda \leq (1+2C')R^* + c\frac{\sqrt{s^* \log(de/s^*)}}{\sqrt{n}} + \frac{\log(1/\epsilon)}{\sqrt{ns^* \log(de/s^*)}},$$

where c is a universal constant depending only on C_1, C', C_x.

Remark 7. The proof of Proposition 1 is given in Appendix A.4. In the noiseless case, i.e. $R^* = 0$, the rate in Proposition 1 is exactly matched the result from theorem 2 in Abramovich and Grinshtein (2018) that proved a lower bound of order $\sqrt{s^* \log(de/s^*)/n}$.

The subsequent refinement becomes apparent within the framework of Theorem 3, wherein the dependency on the value of s^* is now explicitly confined to the tuning parameter τ, while other parameters remain unaffected.

Proposition 2. *Assume that Assumptions 1, 2, and 3 hold. We have, for* $\lambda = 2n/(3C + 2)$, $\tau = s^*/(n\sqrt{d})$ *and with probability at least* $1 - 2\epsilon, \epsilon \in (0, 1)$, *that for all* β^* *such that* $\|\beta^*\|_1 \leq C_1 - 2d\tau$ *and* $\|\beta^*\|_0 \leq s^*$,

$$\int R d\hat{\rho}_\lambda \leq (1 + 3C')R^* + C_{C,C_1,C',C_x} \frac{s^* \log(de/s^*) + \log(1/\epsilon)}{n},$$

where C_{C,C_1,C',C_x} *is a universal constant depending only on* C, C_1, C', C_x.

Remark 8. The proof of Proposition 2 is given in Appendix A.4. In the noiseless case, i.e. $R^* = 0$, the rate in Proposition 2 is exactly matched the result from Theorem 4 in Abramovich and Grinshtein (2018), which is $s^* \log(de/s^*)/n$.

Remark 9. It is emphasized that all the oracle inequalities presented in the paper are not sharp; in other words, instead of $R(\hat{\beta}) \leq R^* + \ldots$, we establish $R(\hat{\beta}) \leq (1 + \delta)R^* + \ldots$ with $\delta > 0$. Therefore, the pursuit of achieving sharp oracle inequalities for our method remains a significant unresolved issue, which we defer to future investigations.

4 | NUMERICAL STUDIES

4.1 | Implementation and comparison of methods

4.1.1 | Implementation

In this section, we introduce the use of the LMC algorithm as a method for sampling from the pseudo-posterior. The LMC algorithm is a gradient-based method for sampling from a distribution.

First, a constant step-size unadjusted LMC algorithm, as described in Durmus and Moulines (2019), is proposed. The algorithm starts with an initial matrix β_0 and uses the recursion:

$$\beta_{s+1} = \beta_s - h\nabla \log \hat{\rho}_\lambda(\beta_s) + \sqrt{2h}E_s \qquad s = 0, 1, \ldots, \tag{3}$$

where $h > 0$ is the step-size and $E_0, E_1, \ldots$ are independent random vectors with i.i.d standard Gaussian entries. It is essential to exercise caution when selecting the step size h, as an insufficiently small value may lead to the summation exploding, as highlighted in Roberts and Stramer (2002). As an alternative method to ensure convergence to the desired distribution,

one can incorporate a Metropolis–Hastings (MH) correction into the algorithm. However, this approach tends to slow down the algorithm due to the additional acceptance/rejection step required at each iteration.

The updating rule presented in (3) is now regarded as a proposal for a new candidate.

$$\tilde{\beta}_{s+1} = \beta_s - h\nabla \log \hat{\rho}_\lambda(\beta_s) + \sqrt{2h}E_s, \qquad s = 0, 1, \ldots,$$

This proposal is accepted or rejected in accordance with the MH algorithm, with the following probability:

$$\min\left\{1, \frac{\hat{\rho}_\lambda(\tilde{\beta}_{s+1})q(\beta_s|\tilde{\beta}_{s+1})}{\hat{\rho}_\lambda(\beta_s)q(\tilde{\beta}_{s+1}|\beta_s)}\right\},$$

where $q(x'|x) \propto \exp\left(-\|x' - x + h\nabla \log \hat{\rho}_\lambda(x)\|^2/(4h)\right)$ is the transition probability density from x to x'. This is recognized as the Metropolis-adjusted Langevin algorithm (MALA), ensuring convergence to the (pseudo) posterior. In contrast to the random-walk Metropolis-Hastings (MH), MALA typically suggests moves toward regions with higher probability, enhancing the likelihood of acceptance. The selection of the step-size h for MALA aims to achieve an acceptance rate of approximately 0.5, as recommended by Roberts and Rosenthal (1998). In the same configuration, the step-size for LMC is chosen to be smaller than those for MALA.

4.1.2 | Comparison of methods

We will evaluate the efficacy, in term of prediction error, of our proposed methodologies by comparing them to Bayesian approaches that utilize logistic regression, as elucidated in Section 2.1. In this scenario, the pseudo-likelihood $\exp(-\lambda r^\ell(M))$, with $\lambda = n$, aligns precisely with the likelihood of the logistic model. Here,

$$r^\ell(M) = \frac{1}{n}\sum_{i=1}^{N} \text{logit}\left(Y_i(\beta^\top x_i)\right),$$

where $\text{logit}(u) = \log(1 + e^{-u})$ represents the logistic loss. It is important to note that the prior distribution remains consistent with the previous sections. As investigated in Zhang (2004), the logistic loss can function as a convex substitute for the hinge loss, providing an approximation to the 0–1 loss. However, it is essential to highlight that utilizing the logistic loss may result in a slower convergence rate compared to the hinge loss, as discussed in Zhang (2004).

In this investigation, we assess the effectiveness of our suggested approaches using hinge loss, identified as H_{LMC} and H_{MALA} for the LMC and MALA algorithms, respectively. We compare these methods with three other alternatives: (1) $\text{Logit}_{\text{LMC}}$, (2) $\text{Logit}_{\text{MALA}}$, both based on Bayesian logistic regression, and (3) the logistic Lasso, which represents a contemporary and highly regarded method. The logistic Lasso is a frequentist technique, and its implementation is available in the `R` package "`glmnet`" (Friedman et al., 2010).

4.2 | Simulation setup

We examine various scenarios for data generation to evaluate the performance of our approach. Initially, we consider a small-scale setup with dimensions $n = 50, p = 100$. In this initial configuration, the sparsity, or the number of non-zero coefficients in the true parameter β^*, is set as $s_0 = 10$. Subsequently, we explore a larger setup with $n = 200, p = 1{,}000$. In this second configuration, the sparsity of the true parameter β^* is adjusted between $s_0 = 100$ and $s_0 = 10$, with the latter denoting a highly sparse model. The entries of the covariate matrix X are generated from a normal distribution $\mathcal{N}(0, 1)$. In all instances, the non-zero coefficients of β^* are independently and identically drawn from $\mathcal{N}(0, 10^2)$.

Next, we explore the following settings to obtain the responses:

- Setting I:

$$Y = \text{sign}(X\beta^* + N)Z.$$

- Setting II: with $u = X\beta^* + N$, put $p = 1/(1 + e^{-u})$:

$$Y_i \sim \text{Binomial}(p_i)Z.$$

In this context, the variability in the noise term (N, Z) results in distinct scenarios, each contributing to a different setup in every setting. A summary of these variations is provided in Table 1.

The LMC, MALA are run with 30,000 iterations and the first 5,000 steps are discarded as burn-in period. The LMC is initialized at the logistic Lasso while the MALA is initialized at zero-vector. We set the values of tuning parameters λ and τ to 1 for all scenarios. It is important to acknowledge that a better approach could be to tune these parameters using cross validation, which could lead to improved results. The logistic Lasso method is run with default options and that 10-fold cross validation is used to select the tuning parameter.

TABLE 1 Outline of simulation settings.

Setting	Name	Z	N
I.1	Hinge	$Z = 1$	$N = 0$
I.2	Hinge with noise	$Z = 1$	$N \sim \mathcal{N}(0, 1)$
I.3	Hinge with switch	$Z \sim 0.9\delta_1 + 0.1\delta_{-1}$	$N = 0$
I.4	Hinge with switch and noise	$Z \sim 0.9\delta_1 + 0.1\delta_{-1}$	$N \sim \mathcal{N}(0, 1)$
II.1	Logistic	$Z = 1$	$N = 0$
II.2	Logistic with switch	$Z \sim 0.9\delta_1 + 0.1\delta_{-1}$	$N = 0$
II.3	Logistic with noise	$Z = 1$	$N \sim \mathcal{N}(0, 1)$
II.4	Logistic with noise and switch	$Z \sim 0.9\delta_1 + 0.1\delta_{-1}$	$N \sim \mathcal{N}(0, 1)$

Each simulation setting is repeated 100 times and we report the averaged results for the misclassification rate. The results of the simulations study are detailed in Tables 2–4 and the values within parentheses indicate the SDs associated with each misclassification rate percentage.

4.3 | Results from simulations

The exhaustive analysis of results extracted from Tables 2–4 provides a thorough insight into the robust performance of our proposed methods when compared to the Lasso and Bayesian logistic approaches. Significantly, throughout all simulated scenarios, the H_{LMC} method, implemented via the LMC algorithm, consistently showcases the smallest misclassification rate. This substantiates the efficacy of our approach across a diverse array of settings, spanning variations in sample size, sparsity levels, and distinct noise settings.

A noteworthy aspect is the exceptional performance highlighted in Table 2, where H_{LMC} outperforms the Lasso method by nearly 10-fold. This pronounced superiority is particularly evident in scenarios characterized by small sample sizes, reinforcing the robustness of our proposed H_{LMC}

TABLE 2 Misclassification rate.

Setting	$Logit_{LMC}$(%)	H_{LMC}(%)	$Logit_{MALA}$(%)	H_{MALA} (%)	Lasso (%)
I.1	4.30 (2.93)	1.36 (1.63)	6.74 (3.80)	2.92 (2.30)	5.76 (7.38)
I.2	4.36 (3.20)	1.26 (1.52)	6.40 (3.67)	2.38 (2.14)	7.02 (6.75)
I.3	13.9 (5.54)	11.1 (4.59)	15.0 (5.18)	12.0 (4.90)	15.6 (8.09)
I.4	13.7 (5.44)	11.4 (5.17)	15.9 (5.68)	12.0 (4.97)	16.3 (12.1)
II.1	2.16 (2.60)	0.66 (1.14)	5.82 (3.71)	2.42 (2.47)	3.60 (11.1)
II.2	2.08 (2.14)	0.54 (0.98)	6.40 (3.49)	3.08 (2.21)	6.58 (7.00)
II.3	11.2 (4.27)	10.0 (4.09)	14.3 (5.05)	11.6 (4.10)	14.6 (7.90)
II.4	12.0 (4.89)	10.3 (4.17)	15.2 (5.56)	12.5 (4.90)	17.7 (14.5)

Note: $n = 50, p = 100, s_0 = 10$.

TABLE 3 Misclassification rate.

Setting	$Logit_{LMC}$ (%)	H_{LMC}(%)	$Logit_{MALA}$ (%)	H_{MALA} (%)	Lasso (%)
I.1	4.60 (1.99)	1.26 (0.91)	7.52 (2.29)	3.04 (1.22)	8.21 (11.4)
I.2	4.76 (2.20)	1.40 (0.90)	7.93 (2.67)	3.00 (1.18)	12.4 (12.2)
I.3	14.1 (2.84)	10.8 (2.14)	16.2 (3.10)	12.2 (2.23)	21.8 (11.2)
I.4	14.1 (262)	10.9 (2.10)	16.4 (2.50)	12.7 (2.08)	20.5 (15.2)
II.1	4.97 (1.91)	1.27 (0.82)	7.90 (2.68)	3.02 (1.33)	9.26 (10.0)
II.2	5.19 (2.12)	1.26 (0.79)	7.48 (2.44)	2.96 (1.25)	10.2 (11.9)
II.3	13.8 (2.56)	11.1 (2.34)	16.5 (3.01)	12.7 (2.41)	20.3 (11.2)
II.4	14.3 (3.26)	11.3 (2.50)	16.2 (3.63)	12.7 (2.43)	19.7 (10.7)

Note: $n = 200, p = 1{,}000, s_0 = 100$.

TABLE 4 Misclassification rate.

Setting	$\text{Logit}_{\text{LMC}}$ (%)	H_{LMC}(%)	$\text{Logit}_{\text{MALA}}$ (%)	H_{MALA} (%)	Lasso (%)
I.1	4.50 (1.68)	1.22 (0.81)	7.20 (2.21)	2.91 (1.22)	4.25 (3.56)
I.2	4.34 (1.71)	1.20 (0.79)	7.34 (2.25)	3.11 (1.21)	3.84 (3.35)
I.3	14.0 (2.34)	10.7 (2.36)	16.0 (2.98)	11.8 (2.48)	12.7 (4.41)
I.4	14.1 (2.62)	10.9 (2.46)	16.4 (2.97)	12.4 (2.36)	12.8 (4.77)
II.1	4.16 (1.68)	1.04 (0.77)	7.13 (2.36)	2.78 (1.17)	3.75 (3.51)
II.2	4.33 (1.47)	1.10 (0.83)	7.12 (2.02)	2.82 (1.16)	4.04 (3.74)
II.3	14.6 (2.91)	11.1 (2.31)	16.9 (2.59)	12.4 (2.24)	13.2 (4.74)
II.4	14.7 (2.77)	11.2 (2.26)	16.8 (2.71)	12.7 (2.33)	13.1 (4.96)

Note: $n = 200, p = 1{,}000, s_0 = 10$.

method under challenging conditions. This outcome underscores the method's ability to navigate challenges related to limited data, emphasizing its potential applicability in practice where small sample sizes are prevalent.

The second most effective strategy emerges from our proposed method implemented using the MALA, referred to as H_{MALA}. It consistently secures the second-best position across various scenarios. Notably, in almost all cases, except for the logistic model scenario in Table 2, where $\text{Logit}_{\text{LMC}}$ exhibits a slight advantage, H_{MALA} stands out as the runner-up. Even in scenarios where other methods show slight advantages, H_{LMC} maintains dominance. Comparing $\text{Logit}_{\text{MALA}}$ with the logistic Lasso reveals nuanced results. In less sparse situations, exemplified by Table 3, $\text{Logit}_{\text{MALA}}$ demonstrates a slight performance edge. However, in the case of highly sparse models, as exemplified in Table 4, the logistic Lasso emerges as the more efficient choice.

Generally, as observed from Tables 3 and 4, there is a tendency for all considered methods to reduce the misclassification rate as the sparsity level increases. This trend is particularly notable for the logistic Lasso. However, the enhancements in performance for our methods (H_{LMC} and H_{MALA}) are modest, indicating their adaptability to varying levels of sparsity. Similarly, with an increase in both dimension and sample size, as demonstrated from Tables 2 to 4, there are also noticeable performance improvements in our methods (H_{LMC} and H_{MALA}), as well as the logistic Lasso. These findings offer valuable insights into the strengths and limitations of each method, facilitating informed decision-making based on the specific characteristics of the dataset under consideration.

4.4 | An application: Prostate tumor classification with microarray gene expression data

In this section, we evaluate the performance of our proposed methods on a real data.

The "prostate" dataset is accessible through the `R` package "`spls`" (Chung, Chun, & Keles, 2019). It comprises 52 samples corresponding to prostate tumors and 50 samples corresponding to normal tissue. The response variable Y encodes normal and tumor classes as 0 and 1, respectively. The covariates matrix X, with dimensions 102 rows by 6,033 columns, represents gene expression data. Preprocessing steps, including normalization, log transformation, and standardization to achieve zero mean and unit variance across genes, were applied to the

TABLE 5 Misclassification rate for real prostate data.

$\text{Logit}_{\text{LMC}}$ (%)	H_{LMC}(%)	$\text{Logit}_{\text{MALA}}$ (%)	H_{MALA} (%)	Lasso (%)
9.71 (4.41)	9.58 (4.31)	9.74 (4.47)	9.55 (4.42)	9.77 (4.31)

arrays, following the procedures outlined in Dettling (2004) and Dettling and Bühlmann (2002). Additional details can be found in Chung and Keles (2010).

The dataset is randomly partitioned into two subsets: a training set comprising 71 samples and a test set comprising 31 samples, roughly representing 70/30 percent of the total samples. The training data is utilized for running the methods, and their prediction accuracy is assessed based on the test data. This process is repeated 100 times, each instance involving a distinct random partition of the training and test data. The outcomes of this iterative procedure are depicted in Table 5. This strategy enables us to accommodate potential fluctuations in the data and gain a more thorough comprehension of the methods' performance.

The findings presented in Table 5 indicate that all the methods under consideration exhibit effective performance, yielding comparable results. Specifically, H_{MALA} demonstrates an error rate of 9.55%, showcasing proficiency; however, this improvement is marginal when compared to the 9.77% error rate achieved by Lasso. The similarity in the outcomes suggests that, in this context, the performance distinctions between H_{MALA} and Lasso are relatively small.

5 | DISCUSSION AND CONCLUSION

In this work, we present an innovative probabilistic framework designed to address the challenges associated with high-dimensional sparse classification problems. Our approach involves the utilization of exponential weights associated with the empirical hinge loss, leading to the establishment of a pseudo-posterior distribution within a class of sparse linear classifiers. Notably, we introduce a sparsity-inducing prior distribution over this class, utilizing a scaled Student's t-distribution with 3 degrees of freedom.

By employing the PAC-Bayesian bound technique, we derive comprehensive theoretical insights into our proposed methodology, particularly focusing on prediction errors. Specifically, under the low-noise condition, we demonstrate that our approach exhibits a fast rate of convergence of order n^{-1}. Importantly, in the noiseless case, our analysis reveals that the prediction error achieved is minimax-optimal. Furthermore, we establish the contraction rate of our pseudo-posterior, presenting novel findings in the current literature.

Beyond the robust theoretical foundation, our approach facilitates practical implementation insights. We leverage the LMC method, a gradient-based sampling approach, to demonstrate the applicability of our framework. Through extensive simulations and a real data application, our method showcases enhanced robustness across various scenarios, such as varying sample sizes and sparsity levels. Numerical results highlight the superior performance of our approach compared to the logistic Lasso, a widely recognized state-of-the-art method.

Looking ahead, future investigations could delve into the estimation challenges posed by our methodology. Additionally, a crucial aspect not addressed in this paper pertains to variable selection, a topic of paramount importance in practical applications. This opens avenues for further research and exploration in enhancing the versatility and applicability of our proposed approach.

ACKNOWLEDGMENTS

The author would like to thank the Editor, Associate Editor and the reviewer who kindly reviewed the earlier version of this paper and provided valuable suggestions and enlightening comments. The author is indebted to the Associate Editor for the helpful suggestions on references on heavy tailed priors, and to the reviewer for critical comments on the proof. The author acknowledges support from the Norwegian Research Council, grant number 309960 through the Centre for Geophysical Forecasting at NTNU.

CONFLICT OF INTEREST STATEMENT

The author declares no potential conflict of interests.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study. Data and Rcodes are available at https://github.com/tienmt/sparseclassificationEWA/.

ORCID

The Tien Mai https://orcid.org/0000-0002-3514-9636

REFERENCES

Abramovich, F., & Grinshtein, V. (2018). High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory*, *65*(5), 3068–3079.

Abramovich, F., Grinshtein, V., & Pensky, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Annals of Statistics*, *35*(5), 2261.

Alquier, P. (2024). User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, *17*(2), 174–303.

Alquier, P., & Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, *48*(3), 1475–1497.

Alquier, P., Ridgway, J., & Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, *17*(1), 8374–8414.

Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156.

Bickel, P. J., & Levina, E. (2004). Some theory for Fisher's linear discriminant function, naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, *10*(6), 989–1010.

Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(5), 1103–1130.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Berlin, Heidelberg: Springer.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Castillo, I., & Mismer, R. (2018). Empirical Bayes analysis of spike and slab posterior distributions. *Electronic Journal of Statistics*, *12*, 3953–4001.

Castillo, I., & van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, *40*(4), 2069–2101.

Catoni, O. (2003). *A PAC-Bayesian approach to adaptive classification*. Paris: Preprint Laboratoire de Probabilités et Modèles Aléatoires. PMA-840.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*. In J. Picard (Ed.), *Saint-flour Summer School on probability theory 2001 Lecture Notes in Mathematics* (Vol. *1851*). Berlin, Germany: Springer-Verlag.

Catoni, O. (2007). *PAC-Bayesian supervised classification: The thermodynamics of statistical learning IMS Lecture Notes—Monograph Series, 56* (). Beachwood, OH: Institute of Mathematical Statistics.

Chung, D., Chun, H., & Keles, S. (2019). *spls: Sparse partial least squares (SPLS) regression and classification*. R package version 2.2-3.

Chung, D., & Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, *9*(1).

Cottet, V., & Alquier, P. (2018). 1-bit matrix completion: PAC-Bayesian analysis of a variational approximation. *Machine Learning*, *107*(3), 579–603.

Dalalyan, A., & Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, *72*(1-2), 39–61.

Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *3*(79), 651–676.

Dalalyan, A. S. (2020). Exponential weights in multivariate regression and a low-rankness favoring prior. *Annales de l'Institute Henri Poincare, Probabilites et Statistiques*, *56*(2), 1465–1483.

Dalalyan, A. S., Grappin, E., & Paris, Q. (2018). On the exponentially weighted aggregate with the laplace prior. *The Annals of Statistics*, *46*(5), 2452–2478.

Dalalyan, A. S., & Riou-Durand, L. (2020). On sampling from a log-concave density using kinetic langevin diffusions. *Bernoulli*, *26*(3), 1956–1988.

Dalalyan, A. S., & Tsybakov, A. (2012a). Mirror averaging with sparsity priors. *Bernoulli*, *18*(3), 914–944.

Dalalyan, A. S., & Tsybakov, A. B. (2012b). Sparse regression learning by aggregation and langevin monte-carlo. *Journal of Computer and System Sciences*, *78*(5), 1423–1443.

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics*, *20*(18), 3583–3593.

Dettling, M., & Bühlmann, P. (2002). Supervised clustering of genes. *Genome Biology*, *3*, 1–15.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition* (Vol. *31*). New York, NY: Springer.

Durmus, A., & Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, *27*(3), 1551–1587.

Durmus, A., & Moulines, E. (2019). High-dimensional Bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, *25*(4A), 2854–2882.

Ermak, D. L. (1975). A computer simulation of charged particles in solution. I. Technique and equilibrium properties. *The Journal of Chemical Physics*, *62*(10), 4189–4196.

Fan, J., & Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of Statistics*, *36*(6), 2605.

Fan, J., Fan, Y., & Wu, Y. (2010). *High-dimensional classification, chapter high-dimensional data analysis* (Vol. *1*, pp. 3–37). China: World Scientific.

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

Germain, P., Lacasse, A., Laviolette, F., March, M., & Roy, J.-F. (2015). Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, *16*(26), 787–860.

Giraud, C. (2021). *Introduction to high-dimensional statistics*. New York: Chapman and Hall/CRC.

Grünwald, P., & Van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, *12*(4), 1069–1103.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. *2*). New York, NY: Springer.

Herbrich, R., & Graepel, T. (2002). A PAC-Bayesian margin bound for linear classifiers. *IEEE Transactions on Information Theory*, *48*(12), 3140–3150.

Hong, L., & Martin, R. (2020). Model misspecification, Bayesian versus credibility estimation, and Gibbs posteriors. *Scandinavian Actuarial Journal*, *2020*(7), 634–649.

Jewson, J., & Rossell, D. (2022). General Bayesian loss function selection and the use of improper models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(5), 1640–1665.

Johnstone, I. M., & Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, *32*(1), 1594–1649.

Knoblauch, J., Jewson, J., & Damoulas, T. (2022). An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, *23*(132), 1–109.

Kotte, V. K., Rajavelu, S., & Rajsingh, E. B. (2020). A similarity function for feature pattern clustering and high dimensional text document classification. *Foundations of Science*, *25*(4), 1077–1094.

Langford, J., & Shawe-Taylor, J. (2002). *PAC-Bayes & margins*. In *Proceedings of the 15th international conference on neural information processing systems* (pp. 439–446). Vancouver, BC: MIT Press.

Li, Y., Chai, Y., Zhou, H., & Yin, H. (2021). A novel dimension reduction and dictionary learning framework for high-dimensional data classification. *Pattern Recognition*, *112*, 107793.

Lyddon, S. P., Holmes, C., & Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, *106*(2), 465–478.

Mai, T. T., & Alquier, P. (2015). A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, *9*(1), 823–841.

Mai, T. T., & Alquier, P. (2017). Pseudo-Bayesian quantum tomography with rank-adaptation. *Journal of Statistical Planning and Inference*, *184*, 62–76.

Mammen, E., & Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, *27*(6), 1808–1829.

Massart, P. (2007). In J. Picard (Ed.), *Concentration inequalities and model selection Lecture notes in mathematics* (Vol. *1896*). Berlin, Germany. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003: Springer.

Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Robust generalised bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 997–1022.

Maurer, A. (2004). A note on the PAC Bayesian theorem. *arXiv preprint cs/0411099*.

McAllester, D. (1998). *Some PAC-Bayesian theorems*. In *Proceedings of the eleventh annual conference on computational learning theory* (pp. 230–234). New York: ACM.

Medina, M. A., Olea, J. L. M., Rush, C., & Velez, A. (2022). On the robustness to misspecification of α-posteriors and their variational approximations. *Journal of Machine Learning Research*, *23*(147), 1–51.

Rivoirard, V. (2006). Nonlinear estimation over weak besov spaces and minimax Bayes method. *Bernoulli*, *12*(4), 609–632.

Roberts, G. O., & Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(1), 255–268.

Roberts, G. O., & Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, *4*(4), 337–357.

Roberts, G. O., & Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, *2*(4), 341–363.

Russo, D., & Zou, J. (2019). How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*, *66*(1), 302–323.

Seeger, M. (2002). PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, *3*(Oct), 233–269.

Seeger, M. W. (2008). Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, *9*, 759–813.

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., & Auer, P. (2012). PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, *58*(12), 7086–7093.

Seldin, Y., & Tishby, N. (2010). Pac-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, *11*(12), 3595–3646.

Shawe-Taylor, J., & Williamson, R. (1997). *A PAC analysis of a Bayes estimator*. In *Proceedings of the tenth annual conference on computational learning theory* (pp. 2–9). New York: ACM.

Syring, N., & Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, *106*(2), 479–486.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, *32*(1), 135–166.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Yonekura, S., & Sugasawa, S. (2023). Adaptation of the tuning parameter in general bayesian inference with robust divergence. *Statistics and Computing*, *33*(2), 39.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, *32*(1), 56–85.

Zhang, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, *52*(4), 1307–1321.

How to cite this article: Mai, T. T. (2024). High-dimensional sparse classification using exponential weighting with empirical hinge loss. *Statistica Neerlandica*, 1–28. https://doi.org/10.1111/stan.12342

APPENDIX A. PROOFS

For any $\Theta \subset \mathbb{R}^d$, let $\mathcal{P}(\Theta)$ denote the set of all probability distributions on Θ equipped with the Borel σ-algebra. For $(\mu, \nu) \in \mathcal{P}(\Theta)^2$, $\mathcal{K}(\nu, \mu)$ denotes the Kullback–Leibler divergence. The following Donsker and Varadhan's lemma is an important key to establish our results.

Lemma 1 (Catoni (2007), lemma 1.1.3). *Let $\mu \in \mathcal{P}(\Theta)$. For any measurable, bounded function $h : \Theta \to \mathbb{R}$ we have*:

$$\log \int e^{h(\theta)} \mu(d\theta) = \sup_{\rho \in \mathcal{P}(\Theta)} \left[\int h(\theta) \rho(d\theta) - \mathcal{K}(\rho, \mu) \right].$$

Moreover, the supremum w.r.t ρ in the right-hand side is reached for the Gibbs distribution, $\hat{\rho}(d\theta) \propto \exp(h(\theta))\pi(d\theta)$.

A.1 Proof for slow rate

We remind that $R^* = R(\beta^*), r_n^* = r_n(\beta^*)$. We remind here a version of Hoeffding's inequality for bounded random variables.

Lemma 2. *Let $U_i, i = 1, \dots, n$ be n independent random variables with $a \le U_i \le b$ a.s., and $\mathbb{E}(U_i) = 0$. Then, for any $\lambda > 0$,*

$$\mathbb{E} \exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} U_i \right) \le \exp\left(\frac{\lambda^2 (b-a)^2}{8n} \right).$$

Proof of Theorem 1. *Step 1*:
Put

$$U_i = \mathbb{1}_{Y_i(\beta^\top x_i) \le 0} - \mathbb{1}_{Y_i(\beta^{*\top} x_i) \le 0}.$$

Then, $-1 \le U_i \le 1$ a.s., we apply the Hoeffding's Lemma 2 to get

$$\mathbb{E} \exp\left\{ \lambda [R(\beta) - R^*] - \lambda [r_n(\beta) - r_n^*] \right\} \le \exp\left\{ \frac{\lambda^2}{2n} \right\}.$$

We obtain, for any $\lambda \in (0, n)$,

$$\int \mathbb{E} \exp\left\{ \lambda [R(\beta) - R^*] - \lambda [r_n(\beta) - r_n^*] - \frac{\lambda^2}{2n} \right\} d\pi(\beta) \le 1,$$

and, using the Fubini's theorem, we get that

$$\mathbb{E}\int \exp\left\{\lambda[R(\beta)-R^*]-\lambda[r_n(\beta)-r_n^*]-\frac{\lambda^2}{2n}\right\}\mathrm{d}\pi(\beta)\leq 1, \tag{A1}$$

Consequently, using Lemma 1,

$$\mathbb{E}\exp\left\{\sup_{\rho}\int\left\{\lambda[R(\beta)-R^*]-\lambda[r_n(\beta)-r_n^*]-\frac{\lambda^2}{2n}\right\}\rho(d\beta)-\mathcal{K}(\rho,\pi)\right\}\leq 1.$$

Using Markov's inequality, for $\epsilon\in(0,1)$,

$$\mathbb{P}\left(\sup_{\rho}\int\left\{\lambda[R(\beta)-R^*]-\lambda[r_n(\beta)-r_n^*]-\frac{\lambda^2}{2n}\right\}\rho(d\beta)-\mathcal{K}(\rho,\pi)+\log\epsilon>0\right)\leq\epsilon.$$

Then taking the complementary and we obtain with probability at least $1-\epsilon$ that:

$$\forall\rho,\quad \lambda\int[R(\beta)-R^*]\rho(d\beta)\leq\lambda\int[r_n(\beta)-r_n^*]\rho(d\beta)+\mathcal{K}(\rho,\pi)+\frac{\lambda^2}{2n}+\log\frac{1}{\epsilon}.$$

Now, note that as $r_n^h\geq r_n$ and as it stands for all ρ then the right-hand side can be minimized and, from Lemma 1, the minimizer over $\mathcal{P}(\mathbb{R}^d)$ is $\hat{\rho}_\lambda$. Thus we get, when $\lambda>0$,

$$\int R d\hat{\rho}_\lambda\leq R^*+\inf_{\rho\in\mathcal{P}(\mathbb{R}^d)}\left[\int r_n^h d\rho+\frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right]-r_n^*+\frac{\lambda}{2n}+\frac{1}{\lambda}\log\frac{1}{\epsilon}.$$

Step 2:

First, we have that,

$$\begin{aligned}\int r_n^h(\beta)\rho(d\beta)&=\frac{1}{n}\int\sum_{i=1}^n(1-Y_i(\beta^\top x_i))_+\ \rho(d\beta)\\&\leq\frac{1}{n}\left[\sum_{i=1}^n(1-Y_i(\beta^{*\top}x_i))_+\int\sum_{i=1}^n\left|(\beta-\beta^*)^\top x_i\right|\rho(d\beta)\right]\\&\leq r_n^h(\beta^*)+\frac{1}{n}\sum_{i=1}^n\int\|\beta-\beta^*\|_2\|x_i\|_2\rho(d\beta)\\&\leq r_n^h(\beta^*)+C_{\mathrm{x}}\int\|\beta-\beta^*\|_2\rho(d\beta).\end{aligned} \tag{A2}$$

And for $\rho=p_0$, as in (A6), and using Lemma 5,

$$\int\|\beta-\beta^*\|_2 p_0(d\beta)\leq\left(\int\|\beta-\beta^*\|_2^2 p_0(d\beta)\right)^{1/2}\leq 2\tau\sqrt{d}.$$

From Lemma 6, we have that

$$\mathcal{K}(p_0,\pi)\leq 4s^*\log\left(\frac{C_1}{\tau s^*}\right)+\log(2).$$

From Assumption 2, we have $r_n^h(\beta^*) \le (1+C')r_n^*$, we obtain

$$\int Rd\hat{\rho}_\lambda \le R^* + C'r_n^* + C_x 2\tau\sqrt{d} + \frac{4s^*\log\left(\frac{C_1}{\tau s^*}\right)+\log(2)}{\lambda} + \frac{\lambda}{2n} + \frac{1}{\lambda}\log\left(\frac{1}{\epsilon}\right).$$

Then, we use Lemma 4, with probability at least $1-2\epsilon$, to obtain that

$$\int Rd\hat{\rho}_\lambda \le (1+2C')R^* + C'\frac{1}{n\varsigma}\log\frac{1}{\epsilon} + C_x 2\tau\sqrt{d} + \frac{4s^*\log\left(\frac{C_1}{\tau s^*}\right)+\log(2)}{\lambda} + \frac{\lambda}{2n} + \frac{1}{\lambda}\log\left(\frac{1}{\epsilon}\right). \tag{A3}$$

By taking $\tau = 1/(n\sqrt{d})$, we obtain that

$$\int Rd\hat{\rho}_\lambda \le (1+2C')R^* + C'\frac{1}{n\varsigma}\log\frac{1}{\epsilon} + C_x\frac{4s^*\log\left(\frac{n\sqrt{d}C_1}{s^*}\right)+\log(2)}{\lambda} + \frac{\lambda}{2n} + \frac{1}{\lambda}\log\left(\frac{1}{\epsilon}\right).$$

By taking $\lambda = \sqrt{n\log(nd)}$, we can obtain that

$$\int Rd\hat{\rho}_\lambda \le (1+2C')R^* + C_x\frac{4s^*\log\left(\frac{n\sqrt{d}C_1}{s^*}\right)}{\sqrt{n\log(nd)}} + \frac{\sqrt{\log(nd)}}{2\sqrt{n}} + \left(\frac{1}{\sqrt{n\log(nd)}} + \frac{C'}{n\varsigma}\right)\log(1/\epsilon).$$

Therefore, we can obtain that

$$\int Rd\hat{\rho}_\lambda \le (1+2C')R^* + c\frac{s^*\sqrt{\log\left(n\sqrt{d}/s^*\right)}}{\sqrt{n}} + c\frac{\log(1/\epsilon)}{\sqrt{n\log(nd)}},$$

where c is a universal constant depending only on C', C_1, C_x. The proof is completed. ∎

Proof of Theorem 2. Let ε_n be any sequence in $(0,1)$ such that $\varepsilon_n \to 0$ when $n \to \infty$. From (A1), we have that

$$\mathbb{E}\left[\int \exp\left\{\lambda[R(\beta)-R^*] - \lambda[r_n(\beta)-r_n^*] - \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\beta)\right] - \frac{\lambda^2}{2n} - \log\frac{1}{\varepsilon_n}\right\}\hat{\rho}_\lambda(d\beta)\right] \le \varepsilon_n.$$

We now use the Chernoff's trick, that is, using $\exp(x) \ge 1_{\mathbb{R}_+}(x)$, this yields:

$$\mathbb{E}\left[\mathbb{P}_{\beta\sim\hat{\rho}_\lambda}(\beta \in \Theta_n)\right] \ge 1-\varepsilon_n,$$

where

$$\Theta_n = \left\{\beta : \lambda[R(\beta)-R^*] - \lambda[r_n(\beta)-r_n^*] \le \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\beta)\right] + \frac{\lambda^2}{2n} + \log\frac{2}{\varepsilon_n}\right\}.$$

Using the definition of $\hat{\rho}_\lambda$ and noting that as $r_n \le r_n^h$, for $\beta \in \Theta_n$ we have

$$\begin{aligned}
\lambda[R(\beta) - R^*] &\le \lambda(r(\beta) - r_n^*) + \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\beta)\right] + \frac{\lambda^2}{2n} + \log\frac{2}{\epsilon_n} \\
&\le \lambda\left(r_n^h(\beta) - r_n^*\right) + \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\beta)\right] + \frac{\lambda^2}{2n} + \log\frac{2}{\epsilon_n} \\
&\le -\log\int \exp\left[-\lambda r_n^h(\beta)\right]\pi(\mathrm{d}\beta) - \lambda r_n^* + \frac{\lambda^2}{2n} + \log\frac{2}{\epsilon_n} \\
&= \lambda\left(\int r_n^h(\beta)\hat{\rho}_\lambda(\mathrm{d}\beta) - r_n^*\right) + \mathcal{K}(\hat{\rho}_\lambda, \pi) + \frac{\lambda^2}{2n} + \log\frac{2}{\epsilon_n} \\
&= \inf_\rho\left\{\lambda\left(\int r_n^h(\beta)\rho(\mathrm{d}\beta) - r_n^*\right) + \mathcal{K}(\rho, \pi) + \frac{\lambda^2}{2n} + \log\frac{2}{\epsilon_n}\right\}.
\end{aligned}$$

We upper-bound the right-hand side exactly as Step 2 in the proof of Theorem 1 (with Lemma 4). The result of the theorem is followed. ∎

A.2 Proof for fast rate

We will make use of the following version of the Bernstein's lemma taken from Massart (2007, p. 24).

Lemma 3. *Let $U_1, \ldots, U_n$ be independent real valued random variables. Let us assume that there are two constants v and w such that $\sum_{i=1}^n \mathbb{E}[U_i^2] \le v$ and that for all integers $k \ge 3$, $\sum_{i=1}^n \mathbb{E}\left[(U_i)_+^k\right] \le vk!w^{k-2}/2$. Then, for any $\zeta \in (0, 1/w)$,*

$$\mathbb{E}\exp\left[\zeta\sum_{i=1}^n\left[U_i - \mathbb{E}U_i\right]\right] \le \exp\left(\frac{v\zeta^2}{2(1-w\zeta)}\right).$$

Proof of Theorem 3. *Step 1*:
Fix any β and put

$$U_i = \mathbb{1}_{Y_i(\beta^\top x_i)\le 0} - \mathbb{1}_{Y_i(\beta^{*\top} x_i)\le 0}.$$

Under Assumption 3, we have that $\sum_i \mathbb{E}[U_i^2] \le nC[R(\beta) - R^*]$. Now, for any integer $k \ge 3$, as the 0-1 loss is bounded, we have that

$$\sum_i \mathbb{E}\left[(U_i)_+^k\right] \le \sum_i \mathbb{E}\left[|U_i|^{k-2}|U_i|^2\right] \le \sum_i \mathbb{E}\left[|U_i|^2\right].$$

Thus, we can apply Lemma 3 with $v := nC[R(\beta) - R^*]$, $w := 1$ and $\zeta := \lambda/n$. We obtain, for any $\lambda \in (0, n)$,

$$\mathbb{E}\exp\{\lambda([R(\beta) - R^*] - [r_n(\beta) - r_n^*])\} \le \exp\left\{\frac{C\lambda^2[R(\beta) - R^*]}{2n(1-\lambda/n)}\right\},$$

and

$$\int \mathbb{E}\exp\left\{\lambda[R(\beta) - R^*] - \lambda[r_n(\beta) - r_n^*] - \frac{C\lambda^2[R(\beta) - R^*]}{2n(1-\lambda/n)}\right\}\mathrm{d}\pi(\beta) \le 1.$$

Them, using Fubini's theorem, we get:

$$\mathbb{E}\int \exp\left\{(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\beta)-R^*] - \lambda[r_n(\beta)-r_n^*]\right\}\pi(d\beta) \leq 1. \tag{A4}$$

Consequently, using Lemma 1,

$$\mathbb{E}\exp\left\{\sup_{\rho}\int\left\{(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\beta)-R^*] - \lambda[r_n(\beta)-r_n^*]\right\}\rho(dM) - \mathcal{K}(\rho,\pi)\right\} \leq 1.$$

Using Markov's inequality,

$$\mathbb{P}\left(\sup_{\rho}\int\left\{(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\beta)-R^*] - \lambda[r_n(\beta)-r_n^*]\right\}\rho(d\beta) - \mathcal{K}(\rho,\pi) + \log\epsilon > 0\right) \leq \epsilon.$$

Then taking the complementary and we obtain with probability at least $1-\epsilon$ that:

$$\forall\rho,\quad (\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})\int[R(\beta)-R^*]\rho(d\beta) \leq \lambda\int[r_n(\beta)-r_n^*]\rho(d\beta) + \mathcal{K}(\rho,\pi) + \log\frac{1}{\epsilon}.$$

Now, note that as $r_n^h \geq r_n$,

$$\lambda\left[\int r_n d\rho - r_n^*\right] + \mathcal{K}(\rho,\pi) + \log\frac{1}{\epsilon} \leq \lambda\left[\int r_n^h d\rho + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right] - \lambda r_n^* + \log\frac{1}{\epsilon}.$$

As it stands for all ρ then the right-hand side can be minimized and, from Lemma 1, the minimizer over $\mathcal{P}(\mathbb{R}^d)$ is $\hat{\rho}_\lambda$. Thus we get, when $\lambda < 2n/(C+2)$,

$$\int R d\hat{\rho}_\lambda \leq R^* + \frac{1}{1-\frac{C\lambda}{2n(1-\lambda/n)}}\left\{\inf_{\rho\in\mathcal{P}(\mathbb{R}^d)}\left[\int r_n^h d\rho + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right] - r_n^* + \frac{1}{\lambda}\log\frac{1}{\epsilon}\right\}.$$

Step 2:

From (A2), we have that,

$$\int r_n^h(\beta)\rho(d\beta) \leq r_n^h(\beta^*) + C_x\int\|\beta-\beta^*\|_2\rho(d\beta).$$

And for $\rho = p_0$, as in (A2), and using Lemma 5,

$$\int\|\beta-\beta^*\|_2 p_0(d\beta) \leq 2\tau\sqrt{d}.$$

From Lemma 6, we have that

$$\mathcal{K}(p_0,\pi) \leq 4\|\beta^*\|_0\log\left(\frac{C_1}{\tau\|\beta^*\|_0}\right) + \log(2).$$

From Assumption 2, as $r_n^h(\beta^*) \le (1+C')r_n^*$, we have that

$$\int R d\hat{\rho}_\lambda \le R^* + \frac{1}{1-\frac{C\lambda}{2n(1-\lambda/n)}}\left\{C' r_n^* + C_x 2\tau\sqrt{d} + \frac{4\|\beta^*\|_0 \log\left(\frac{C_1}{\tau\|\beta^*\|_0}\right) + \log(2)}{\lambda} + \frac{1}{\lambda}\log\left(\frac{1}{\epsilon}\right)\right\}.$$

Taking $\lambda = 2n/(3C+2)$, we obtain:

$$\int R d\hat{\rho}_\lambda \le R^* + \frac{3}{2}\left\{C' r_n^* + C_x 2\tau\sqrt{d} + \frac{(3C+2)\left[4\|\beta^*\|_0 \log\left(\frac{C_1}{\tau\|\beta^*\|_0}\right) + \log(2)\right]}{2n} + \frac{(3C+2)\log(1/\epsilon)}{2n}\right\}.$$

Then, we use Lemma 4, with probability at least $1-2\epsilon$, to obtain that

$$\int R d\hat{\rho}_\lambda \le (1+3C')R^* + \frac{3}{2}\left\{C'\frac{1}{n\varsigma}\log\frac{1}{\epsilon} + C_x 2\tau\sqrt{d} + \frac{(3C+2)\left[4\|\beta^*\|_0 \log\left(\frac{C_1}{\tau\|\beta^*\|_0}\right) + \log(2)\right]}{2n} + \frac{(3C+2)\log(1/\epsilon)}{2n}\right\}. \tag{A5}$$

By taking $\tau = 1/(n\sqrt{d})$, we can obtain that

$$\int R d\hat{\rho}_\lambda \le (1+3C')R^* + C_{C,C',C_x}\frac{\|\beta^*\|_0 \log\left(\frac{n\sqrt{d}C_1}{\|\beta^*\|_0}\right) + \log(1/\epsilon)}{n},$$

where C_{C,C',C_x} is a universal constant depending only on C, C', C_1, C_x. The proof is completed. ∎

Proof of Theorem 4. Let ε_n be any sequence in $(0,1)$ such that $\varepsilon_n \to 0$ when $n \to \infty$. From (A4), we have that

$$\mathbb{E}\left[\int \exp\left\{(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\beta) - R^*] - \lambda[r_n(\beta) - r_n^*] - \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\beta)\right] - \log\frac{1}{\varepsilon_n}\right\}\hat{\rho}_\lambda(d\beta)\right] \le \varepsilon_n.$$

Using Chernoff's trick, that is, using $\exp(x) \ge 1_{\mathbb{R}_+}(x)$, this gives:

$$\mathbb{E}\left[\mathbb{P}_{\beta\sim\hat{\rho}_\lambda}(\beta \in \Omega_n)\right] \ge 1 - \varepsilon_n,$$

where

$$\Omega_n = \left\{\beta : (\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\beta) - R^*] - \lambda[r_n(\beta) - r_n^*] \le \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\beta)\right] + \log\frac{2}{\varepsilon_n}\right\}.$$

Using the definition of $\hat{\rho}_\lambda$ and noting that $r_n \le r_n^h$, for $\beta \in \Omega_n$ we have

$$\begin{aligned}\left(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)}\right)\left[R(\beta) - R^*\right] &\le \lambda(r_n(\beta) - r_n^*) + \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\beta)\right] + \log\frac{2}{\epsilon_n} \\ &\le \lambda\left(r_n^h(\beta) - r_n^*\right) + \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\beta)\right] + \log\frac{2}{\epsilon_n} \\ &\le -\log\int \exp\left[-\lambda r_n^h(\beta)\right]\pi(\mathrm{d}\beta) - \lambda r_n^* + \log\frac{2}{\epsilon_n} \\ &= \lambda\left(\int r_n^h(\beta)\hat{\rho}_\lambda(\mathrm{d}\beta) - r_n^*\right) + \mathcal{K}(\hat{\rho}_\lambda, \pi) + \log\frac{2}{\epsilon_n} \\ &= \inf_\rho\left\{\lambda\left(\int r_n^h(\beta)\rho(\mathrm{d}\beta) - r_n^*\right) + \mathcal{K}(\rho, \pi) + \log\frac{2}{\epsilon_n}\right\}.\end{aligned}$$

We upper-bound the right-hand side exactly as Step 2 in the proof of Theorem 3 (with Lemma 4). The result of the theorem is followed. ∎

A.3 Proof of auxiliary lemmas

Lemma 4. *For $\epsilon \in (0,1)$, with probability at least $1-\epsilon$, we have, for every $\varsigma \in (0,1)$, that*

$$r_n^* \le (1+\varsigma)R^* + \frac{1}{n\varsigma}\log\frac{1}{\epsilon}.$$

or we can have $r_n^* \le 2R^* + \frac{1}{n\varsigma}\log\frac{1}{\epsilon}$.

Proof. Let $\varsigma \in (0,1)$, we have that

$$\begin{aligned}\mathbb{E}(\exp[\varsigma n r_n^*]) &= \prod_{i=1}^n \mathbb{E}\left(\exp\left[\varsigma \mathbb{1}_{(Y_i(\beta^{*\top}x_i)<0)}\right]\right) \\ &= \prod_{i=1}^n \mathbb{E}\left\{\exp\left[\varsigma \mathbb{1}_{(Y_i(\beta^{*\top}x_i)<0)} + 0(1 - \mathbb{1}_{(Y_i(\beta^{*\top}x_i)<0)})\right]\right\} \\ &\le \prod_{i=1}^n \left\{e^\varsigma \mathbb{E}\left[\mathbb{1}_{(Y_i(\beta^{*\top}x_i)<0)}\right] + \left(1 - \mathbb{E}\left[\mathbb{1}_{(Y_i(\beta^{*\top}x_i)<0)}\right]\right)\right\} \\ &\le \prod_{i=1}^n (e^\varsigma R^* + 1 - R^*) = \prod_{i=1}^n (R^*(e^\varsigma - 1) + 1) \\ &\le \prod_{i=1}^n \exp(R^*(e^\varsigma - 1)) = \exp(nR^*(e^\varsigma - 1)).\end{aligned}$$

Thus we obtain, for $\epsilon \in (0,1)$:

$$\mathbb{E}\left[\exp\left(\varsigma n r_n^* - nR^*(e^\varsigma - 1) - \log\frac{1}{\epsilon}\right)\right] \le \epsilon.$$

Now, using Markov's inequality that $\mathbb{P}(W > 0) \le \mathbb{E}[e^W]$ for any W, we get that

$$\varsigma n r_n^* - (e^\varsigma - 1)nR^* - \log\frac{1}{\epsilon} \le 0,$$

with probability at least $1-\epsilon$. Thus, the result of the lemma is obtained by noting that $e^{\varsigma} \leq 1+\varsigma+\varsigma^2, \varsigma \in (0,1)$. ∎

Definition 1. We define the following distribution as a translation of the prior π,

$$p_0(\beta) \propto \pi(\beta-\beta^*)\mathbb{1}_{B_1(2d\tau)}(\beta-\beta^*). \tag{A6}$$

It is worth highlighting that given $\|\beta^*\|_1 \leq C_1 - 2d\tau$, when the condition $\beta-\beta^* \in B_1(2d\tau)$ holds, it implies that $\beta \in B_1(C_1)$. Consequently, the distribution p_0 is absolutely continuous with respect to the prior distribution π.

Lemma 5. *Let p_0 be the probability measure defined by (A6). If $d \geq 2$ then*

$$\int_\Lambda \|\beta-\beta^*\|^2 p_0(d\beta) \leq 4d\tau^2.$$

Proof. First, we have that

$$\int_\Lambda \|\beta-\beta^*\|^2 p_0(d\beta) = d\int_\Lambda (\beta_1-\beta_1^*)^2 p_0(d\beta).$$

Using lemma 2 from Dalalyan and Tsybakov (2012a) we get

$$\int_\Lambda (\beta_1-\beta_1^*)^2 p_0(d\beta) \leq 4\tau^2,$$

and the desired inequality follows. ∎

Lemma 6. *Let p_0 be the probability measure defined by (A6). Then*

$$KL(p_0,\pi) \leq 4s^* \log\left(\frac{C_1}{\tau s^*}\right) + \log(2).$$

Proof. From lemma 3 in Dalalyan and Tsybakov (2012a), we have that

$$KL(p_0,\pi) \leq 4\sum_{j=1}^{d} \log(1+|\beta_j^*|/\tau) + \log(2).$$

Then, from corollary 1 in Dalalyan and Tsybakov (2012a) (that using Jensen's inequality), we have that

$$\frac{1}{s^*}\sum_{j=1}^{d} \log(1+|\beta_j^*|/\tau) \leq \log\left(1+\frac{\|\beta^*\|_1}{\tau s^*}\right).$$

As $\|\beta^*\|_1 \leq C_1 - 2d\tau$, we then have that

$$\log\left(1+\frac{\|\beta^*\|_1}{\tau s^*}\right) \leq \log\left(1+\frac{C_1-2d\tau}{\tau s^*}\right) \leq \log\left(\frac{C_1}{\tau s^*}-1\right) \leq \log\left(\frac{C_1}{\tau s^*}\right),$$

by noting that $\|\beta^*\|_0 \leq s^* \leq n < d$. Thus, we obtain the desired result. ∎

A.4 Proofs for Section 3.3

Proof of Proposition 1. From the proof of Theorem 1, inequality (A3), with probability at least $1-2\epsilon$, we have that

$$\int Rd\hat{\rho}_\lambda \leq (1+2C')R^* + C'\frac{1}{n\varsigma}\log\frac{1}{\epsilon} + C_{\mathrm{x}}2\tau\sqrt{d} + \frac{4s^*\log\left(\frac{C_1}{\tau s^*}\right)+\log(2)}{\lambda} + \frac{\lambda}{2n} + \frac{1}{\lambda}\log\left(\frac{1}{\epsilon}\right).$$

By taking now $\tau = s^*/(n\sqrt{d})$ and $\lambda = \sqrt{ns^*\log(de/s^*)}$, we obtain that

$$\int Rd\hat{\rho}_\lambda \leq$$
$$(1+2C')R^* + C_{\mathrm{x}}\frac{2s^*}{n} + \frac{4s^*\log\left(\frac{n\sqrt{d}C_1}{s^*s^*}\right)}{\sqrt{ns^*\log(de/s^*)}} + \frac{\sqrt{s^*\log(de/s^*)}}{2\sqrt{n}}$$
$$+\left(\frac{1}{\sqrt{ns^*\log(de/s^*)}} + \frac{C'}{n\varsigma}\right)\log\left(\epsilon^{-1}\right).$$

By noting that

$$\frac{n\sqrt{d}}{s^*s^*} = \frac{n}{s^*e\sqrt{d}}\frac{de}{s^*} \leq \left(\frac{de}{s^*}\right)^2. \tag{A7}$$

Therefore, we can obtain that

$$\int Rd\hat{\rho}_\lambda \leq (1+2C')R^* + c\frac{\sqrt{s^*\log(de/s^*)}}{\sqrt{n}} + c\frac{\log(1/\epsilon)}{\sqrt{ns^*\log(de/s^*)}},$$

where c is a universal constant depending only on C_1, C', C_{x}. The proof is completed. ∎

Proof of Proposition 2. From the proof of Theorem 3, inequality (A5), with probability at least $1-2\epsilon$, we have that

$$\int Rd\hat{\rho}_\lambda \leq (1+3C')R^* +$$
$$\frac{3}{2}\left\{\frac{C'}{n\varsigma}\log\frac{1}{\epsilon} + C_{\mathrm{x}}2\tau\sqrt{d} + \frac{(3C+2)\left[4s^*\log\left(\frac{C_1}{\tau s^*}\right)+\log(2)\right]}{2n} + \frac{(3C+2)\log(1/\epsilon)}{2n}\right\}.$$

By taking now $\tau = s^*/(n\sqrt{d})$, we obtain that

$$\int Rd\hat{\rho}_\lambda \leq (1+3C')R^* + C_{C,C',C_{\mathrm{x}}}\frac{s^*\log\left(\frac{n\sqrt{d}C_1}{s^*s^*}\right)+\log(1/\epsilon)}{n},$$

and using the notice in (A7), we obtain

$$\int R d\hat{\rho}_{\lambda} \leq (1+3C')R^{*} + C_{C,C',C_{\mathrm{x}}} \frac{s^{*}\log\left(\frac{de}{s^{*}}\right) + \log(1/\epsilon)}{n},$$

where $C_{C,C',C_{\mathrm{x}}}$ is a universal constant depending only on $C, C_1, C', C_{\mathrm{x}}$. The proof is completed. ■