Journal of Machine Learning Research 26 (2025) 1-54 Submitted 11/23; Revised 1/25; Published 3/25

On Inference for the Support Vector Machine

Jakub Rybak JAKUB.RYBAK18@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
London, SW7 2AZ, U.K.

Heather Battey H.BATTEY@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
London, SW7 2AZ, U.K.

Wen-Xin Zhou WENXINZ@UIC.EDU
Department of Information and Decision Sciences
University of Illinois Chicago
Chicago, IL 60607, USA

Editor: Kenji Fukumizu

Abstract

The linear support vector machine has a parametrised decision boundary. The paper considers inference for the corresponding parameters, which indicate the effects of individual variables on the decision boundary. The proposed inference is via a convolution-smoothed version of the SVM loss function, this having several inferential advantages over the original SVM, whose associated loss function is not everywhere differentiable. Notably, convolution-smoothing comes with non-asymptotic theoretical guarantees, including a distributional approximation to the parameter estimator that scales more favourably with the dimension of the feature vector. The differentiability of the loss function produces other advantages in some settings; for instance, by facilitating the inclusion of penalties or the synthesis of information from a large number of small samples. The paper closes by relating the linear SVM parameters to those of some probability models for binary outcomes.

Keywords: support vector machines, Bahadur representation, convolution smoothing, non-asymptotic statistics

1. Introduction

Owing to their ability to predict complex phenomena across many applications, support vector machines (SVMs) (Cortes and Vapnik, 1995) have become one of the most popular classification algorithms. While the interpretation of its implicit parameters is less direct than from a parametric model-based approach, inference for such parameters is valuable in that it indicates the strength of evidence for variables having a real effect on the decision boundary. Inference for the support vector machine requires minimisation of a non-differentiable hinge loss, leading to a quadratic programming problem. Previous literature has pursued a more computational line of enquiry regarding the scaling of this quadratic

program with dimension (e.g. Joachims, 1998; Hsieh et al., 2008; Fan et al., 2008). Such scaling issues are exacerbated if an ℓ_1 penalty is added to the SVM objective function, as interior point methods then need to be used (Wang et al., 2023).

Several smooth surrogate losses have been proposed in the literature, typically emphasising computation; our concern is with the associated statistical guarantees. We provide a theoretical analysis of a general type of convolution-based smoothing in the context of the support vector machine, which generalises the approach proposed by Lee and Mangasarian (2001). In comparison to alternative smoothing devices, convolution-smoothing comes with non-asymptotic theoretical guarantees for the implicit parameters. Additionally, the distributional approximation for the convolution-smoothed estimator is valid under a weaker requirement on the number of explanatory variables compared to the original SVM and the smoothed version of Wang et al. (2019), the implication being that, with many potential explanatory variables, convolution-smoothing the hinge loss improves statistical properties of the resulting estimator, in addition to any computational advantages more commonly emphasised.

Lee and Mangasarian (2001) proposed smoothing the non-differentiable hinge loss by an integral of a cumulative distribution function, which results in a twice-differentiable surrogate loss. This allowed the authors to replace the quadratic programming problem of standard SVMs by second-order methods (Lee and Mangasarian, 2001). However, theoretical guarantees for this method have not been ascertained. Replacement of the quadratic optimisation by a more feasible alternative is also an objective of Suykens and Vandewalle (1999), who considered ridge regression with binary targets. By adopting Horowitz smoothing (Horowitz, 1998) from the quantile regression literature, Wang et al. (2019) made use of a smooth surrogate for the hinge loss. Although the resulting objective function is twice continuously differentiable, it is also non-convex.

Finite-sample results for variable selection and estimation error of linear SVMs have been established by Zhang (2004). The distributional properties of an unpenalised linear SVM estimator were studied by Koo et al. (2008), who derived a Bahadur representation for linear SVMs. In particular, in the non-separable case and under appropriate regularity conditions, Koo et al. (2008) showed that the population hinge loss is differentiable and locally strongly convex around its minimiser. However, due to the non-differentiability of the hinge loss, the empirical loss does not share these properties. The analysis of Koo et al. (2008) is restricted to an asymptotic setting, but the simulations presented there suggest that the distributional approximation for the SVM estimator is valid only when $n \gg p$, where n is the sample size and p is the number of potential explanatory variables.

We consider convolution smoothing in the context of the SVM. This was previously proposed in the context of quantile regression (Fernandes et al., 2021; He et al., 2023), and its application to SVMs was independently studied in the complementary work by Wang et al. (2023). The resulting empirical loss function is twice continuously differentiable, globally convex, and locally strongly convex. We show that convolution-type smoothing can be seen as a generalisation of the smoothing method of Lee and Mangasarian (2001). We establish non-asymptotic bounds for the estimation error and the Bahadur linearisation error of the convolution-smoothed SVM. Our results indicate that the distributional approximation for

the convolution-smoothed hinge loss is valid under a weaker requirement on the number of variables than for the original SVM and the smoothing of Wang et al. (2019), and is thus more suitable for large-p settings. While a similar result has been established in the quantile regression literature by He et al. (2023), non-asymptotic analysis of SVMs requires appreciable modifications due to a different structure of the error term, a point further elaborated by Wang et al. (2019).

The emphasis for the majority of this paper is on inference for the parameters of the linear SVM, these having an interpretation as outlined above and developed in more detail in Section 9. If, however, the sole purpose is prediction, non-linear classifiers such as kernel SVMs are often advocated for settings with relatively few explanatory variables. Convolution smoothing can be applied in this context too. The extension of the objective function to non-linear SVMs and the resulting quadratic program is given in Section 8.

The paper is organised as follows. Section 3 reviews the existing approaches to smoothing and introduces the convolution-smoothed SVM. In Section 4, finite-sample bounds for the estimation error (Section 4.1) and Bahadur linearisation error (Section 4.2) of the convolution-smoothed estimator are derived. Our finite-sample results rely on local strong convexity of the smoothed hinge loss, which is derived in Appendix A. A comparison of distributional approximations for the convolution-smoothed loss and the original SVM is presented in Section 5, while the comparison with the estimator proposed by Wang et al. (2019) is the subject of Section 6. Bahadur representation derived in Section 4 leads to Wald-type inference. A further advantage of smoothing is that it enables the construction of confidence sets based on an inversion of a score-type test, whose advantages over Wald-based inference in some contexts have been pointed out by Tan et al. (2022) for quantile regression. Score-type inference for convolution-smoothed SVM is briefly outlined in Section 7. In Section 8 we present a non-linear extension of the convolution-smoothed SVM. The paper closes with some conceptual discussion of the interpretation of parameters in the linear support vector machine in Section 9.

2. Notation

Except when it is helpful to be explicit, all constants are represented by C, although their numerical value may change from line to line. For every integer $k \geq 1$, $\mathbb{R}^k$ denotes the k-dimensional Euclidean space. For any vector $v \in \mathbb{R}^k$, let v_{-j} denote a vector obtained from v by omitting the element with index j and for any $u, w \in \mathbb{R}^k$, $\langle u, w \rangle$ denotes the inner product of the two vectors. For $r \geq 0$, define the Euclidean ball and sphere in $\mathbb{R}^k$ as $\mathbb{B}^k(r) = \{u \in \mathbb{R}^k : \|u\|_2 \leq r\}$ and $\mathbb{S}^{k-1}(r) = \partial\mathbb{B}^k(r) = \{u \in \mathbb{R}^k : \|u\|_2 = r\}$, respectively. For unit balls we omit the radius r and write simply $\mathbb{B}$. For a pair of radii $0 < r_1 < r_2$, $\mathbb{B}(r_1, r_2)$ denotes the resulting "doughnut set", i.e. $\mathbb{B}^k(r_1, r_2) \triangleq \{u \in \mathbb{R}^k : r_1 \leq \|u\|_2 \leq r_2\}$. When $k = p + 1$ we omit the superscript and write $\mathbb{B}(r)$ and $\partial\mathbb{B}(r)$. For two sequences of non-negative numbers $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of n such that $a_n \leq Cb_n$, while $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Equality by definition is written $\triangleq$, whereas $\overset{d}{=}$ denotes equality in distribution.

For $x \in \mathbb{R}$, $\{x\}_+$ is the positive part of x, $\text{sgn}(x)$ denotes the sign of x and $\mathbb{1}\{\cdot\}$ is the indicator function.

Unless otherwise stated, upper-case letters X and Y denote random variables, with corresponding lower-case letters their realisations. For a random variable X valued in $\mathbb{R}^p$, let $\mathring{X} \triangleq (1, X^T)^T$. The first element of $\mathring{X}$ is indexed by zero, which implies that the s^{th} element of $\mathring{X}$ coincides with the s^{th} element of X. Where appropriate, we use an analogous notation for deterministic vectors, i.e. $\mathring{v} = (v_0, v)$, $v_0 \in \mathbb{R}$ and $v \in \mathbb{R}^p$. We use the same indexing for all $(p+1)$-dimensional vectors, i.e. for $\alpha \in \mathbb{R}^{p+1}$, $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)$. For a parameter vector $\theta \in \mathbb{R}^{p+1}$, $\theta = (b, w)$, where in the machine learning literature $b \in \mathbb{R}$ is referred to as the bias parameter, and $w \in \mathbb{R}^p$ is the vector orthogonal to the hyperplane given by $\theta^T x = 0$. The conditional density functions of X given $Y = 1$ and $Y = -1$ are $f(\cdot)$ and $g(\cdot)$ respectively. The conditional density of X_s given $X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p$ is denoted by $f(x|X_{-s})$.

3. Smooth SVMs

Let X be a p-dimensional random vector, and Y a binary random variable with values in $\{-1, 1\}$. The linear SVM population optimisation problem solves $\min_{\theta \in \mathbb{R}^{p+1}} L(\theta)$, where $L(\theta) \triangleq \mathbb{E}\{1 - Y\mathring{X}^T\theta\}_+$ is the population hinge loss, with expectation taken with respect to the joint distribution of X and Y, and $\theta = (w, b)$, $w \in \mathbb{R}^p$, $b \in \mathbb{R}$.

Given a sample $(y_i, x_i)_{i=1}^n$, the linear SVM minimises a penalised empirical analogue of L,

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(1 - y_i \mathring{x}_i^T \theta\right)_+ + \frac{\lambda}{2} \|w\|_2^2, \tag{1}$$

where here, and throughout this section, we do not distinguish notationally between estimators and their realisations.

3.1 Existing Approaches

To obtain a smooth objective function, Suykens and Vandewalle (1999) replace the non-differentiable hinge loss by a squared loss. The resulting estimator is equivalent to fitting a ridge regression with a binary response variable. Lee and Mangasarian (2001) proposed replacing the hinge loss with a smooth approximation. In particular, the sub-gradient of the hinge loss is approximated by a cumulative distribution function of a logistic random variable. The empirical hinge loss is thus replaced by an integral of the sigmoid function. The resulting (unpenalised) loss takes the form,

$$\hat{L}_\alpha^{lm}(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} s(\varepsilon_i(\theta), \alpha), \;\; s(x, \alpha) \triangleq x + \frac{1}{\alpha} \log(1 + e^{-\alpha x}), \tag{2}$$

where $\varepsilon_i(\theta) = 1 - \mathring{x}_i^T \theta$ and α is a smoothing constant. As we will see, this method of smoothing is equivalent to convolution-smoothing with a particular choice of kernel. To obtain a strongly convex loss function, Lee and Mangasarian (2001) further modify the

objective function to obtain

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} s(\varepsilon_i(\theta), \alpha)^2 + \frac{\lambda}{2} \|\theta\|_2^2. \tag{3}$$

Unlike the hinge loss, (3) imposes a quadratic penalty on misclassified samples. The solution of (3) no longer converges to the SVM solution as $n \to \infty$.

Adopting the smoothing technique of Horowitz (1998) from the quantile regression literature, Wang et al. (2019) consider smoothing the hinge loss, $(\varepsilon)_+ = \varepsilon \mathbb{1}\{\varepsilon \geq 0\}$, by replacing the indicator function by a smooth alternative. Specifically, consider a function $H : \mathbb{R} \to (0, 1)$, such that,

$$H(x) = \begin{cases} 0 & \text{if } x \leq -1 \\ (0,1) & \text{if } x \in (-1, 1) \\ 1 & \text{if } x \geq 1. \end{cases}$$

The smoothed loss takes the form

$$L^w(\varepsilon, h) \triangleq \varepsilon H\left(\frac{\varepsilon}{h}\right),$$

where h is a smoothing constant. As $h \to \infty$, the smoothed loss approaches the original hinge loss. In practice, $H(\cdot)$ is usually chosen as the integral of a kernel density estimator. Wang et al. (2019) provide a detailed non-asymptotic analysis of the resulting estimator. In particular, as $n \to \infty$, the smoothed estimator converges to the SVM estimator.

3.2 Convolution-smoothed SVM

In the absence of regularisation term ($\lambda = 0$), we can rewrite the sample SVM loss function as

$$\hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i(\theta) \mathbb{1}\{\varepsilon_i(\theta) \geq 0\} = \int_0^{\infty} u \, d\hat{F}(u, \theta),$$

where $\varepsilon_i(\theta) \triangleq 1 - y_i \mathring{x}_i^T \theta$ and $\hat{F}(u, \theta) = (1/n) \sum_{i=1}^{n} \mathbb{1}\{\varepsilon_i(\theta) \leq u\}$ denotes the empirical CDF of $\{\varepsilon_i(\theta)\}_{i=1}^n$. This motivates a smoothing approach in which the discontinuous empirical distribution function $\hat{F}$ is replaced by a continuous alternative. For this, introduce a kernel density function $K : \mathbb{R} \to [0, \infty)$, symmetric around zero, and define a kernel distribution function estimator

$$\hat{F}_h(u, \theta) = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{u} K\left(\frac{t - 1 + y_i \mathring{x}_i^T \theta}{h}\right) dt,$$

with bandwidth $h > 0$ to be chosen. This leads to an empirical smoothed hinge loss,

$$\hat{L}_h(\theta) = \int_0^{\infty} u \, d\hat{F}_h(u, \theta) = \frac{1}{nh} \sum_{i=1}^{n} \int_0^{\infty} uK\left(\frac{u + y_i \mathring{x}_i^T \theta - 1}{h}\right) du.$$

Figure 1 illustrates the smoothed loss for different choices of h. To see why this can be interpreted as convolution-type smoothing, write the convolution operation as $*$ and define $\varphi(u) \triangleq u\,\mathbb{1}(u \geq 0)$, $u \in \mathbb{R}$ and

$$l_h(u) = (\varphi * K_h)(u) = \int_{-\infty}^{\infty} \varphi(t) \frac{1}{h} K\left(\frac{u-t}{h}\right) dt.$$

Hence the smoothed loss function is $\hat{L}_h(\theta) = (1/n)\sum_{i=1}^n l_h(1 - y_i \mathring{x}_i^T \theta)$. The gradient and Hessian of the sample smoothed loss function are

$$\nabla \widehat{L}_h(\theta) = -\frac{1}{n}\sum_{i=1}^n y_i \mathring{x}_i \bar{K}\left(\frac{1 - y_i \mathring{x}_i^T \theta}{h}\right); \quad \nabla^2 \widehat{L}_h(\theta) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{1 - y_i \mathring{x}_i^T \theta}{h}\right) \mathring{x}_i \mathring{x}_i^T, \tag{4}$$

where $\bar{K}(u) = \int_{-\infty}^u K(t)dt$. As long as $K(\cdot)$ is non-negative, the sample smoothed loss function is convex. We thus obtain a twice-differentiable, convex surrogate to the hinge loss. The main difference between the convolution-smoothed loss and the smooth loss of Wang et al. (2019) is that the former is globally convex, while the latter is not, as illustrated in Figure 1.

Lastly, write the estimator obtained by minimising the convolution-smoothed loss as

$$\hat{\theta}_h \triangleq \underset{\theta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \hat{L}_h(\theta),$$

and let $\hat{\Delta} \triangleq \hat{\theta}_h - \theta^*$, i.e. the difference between the $\hat{\theta}_h$ and the minimiser of the population hinge loss. We write the population version of $\hat{L}_h$ as L_h, i.e.

$$L_h(\theta) \triangleq \mathbb{E}[\hat{L}_h(\theta)] = \frac{1}{h}\mathbb{E}\left[\int_0^{\infty} uK\left(\frac{u + Y\mathring{X}^T\theta - 1}{h}\right) du\right]. \tag{5}$$

The gradient and Hessian of the convolution-smoothed population loss will be denoted by $\nabla L_h(\theta)$ and $\nabla^2 L_h(\theta)$ respectively.

The following lemma establishes that the convolution-smoothed SVM can be written in terms of an indefinite integral of the cdf of $K(\cdot)$.

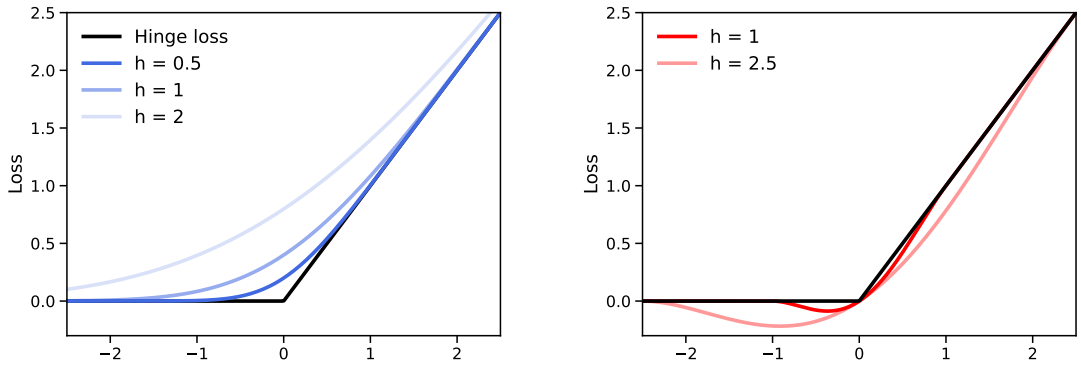
Lemma 1 *Let $K_h : \mathbb{R} \to [0, \infty)$ be a kernel density symmetric around zero with bandwidth h, and $\hat{L}_h$ the corresponding convolution-smoothed hinge loss. Let $\bar{\mathcal{K}}_h$ be an indefinite integral of the cumulative distribution function of K_h. Then,*

$$\underset{\theta}{\operatorname{argmin}}\, \hat{L}_h(\theta) = \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^n \bar{\mathcal{K}}_h(1 - y_i x_i^T \theta).$$

Convolution-smoothing thus approximates the sub-gradient of the hinge loss by a cumulative distribution function of a zero-mean random variable with symmetric density. If the kernel function is a logistic density function with zero mean and scale parameter one, that is, $K(u) = e^{-u}/(1 + e^{-u})^2$, then,

$$\bar{\mathcal{K}}_h(x) = \frac{1}{h}\big\{x + h\log(1 + e^{-x/h})\big\}$$

and we obtain, $\hat{L}_h(\theta) = (1/h)\hat{L}^{lm}_{1/h}(\theta)$, i.e., convolution-smoothing is equivalent to the smoothing of Lee and Mangasarian (2001) in equation (2).



(a) Convolution-smoothing (b) Horowitz smoothing

Figure 1: Hinge loss and its smooth approximations, for different choice of bandwidth (h). Convolution-smoothed loss (plot a) using Gaussian kernel and Horowitz-type smoothing of Wang et al. (2019) using Epanechnikov kernel (plot b).

4. Finite-sample Theoretical Guarantees

In this section we obtain finite-sample bounds on the error of the convolutional smoothed estimator of θ^*, and derive a Bahadur representation for $\hat{\theta}_h$. This leads to a distributional approximation for the convolution-smoothed estimator $\hat{\theta}_h$.

Koo et al. (2008) showed that, under mild conditions, the population minimiser of the SVM hinge loss, θ^*, is unique and non-zero. Assumptions A1—A4 in Koo et al. (2008) are assumed throughout, resulting in Assumptions 2 and 8 below. Additionally, we impose mild assumptions on the kernel density (Assumptions 4 and 7) and on the conditional densities of X (Assumption 5), both of which are satisfied by commonly used density functions. Lastly, the feature vector X is assumed to be sub-Gaussian (Assumption 3). Analogous assumptions are employed by Wang et al. (2019).

Assumption 2 *The minimiser $\theta^* = (b^*, (w^*)^T)^T$ of the SVM population loss function is unique. For some constants $C, C' > 0$, and for some $s \in \{1, \ldots, p\}$, $(\theta^*_s)^{-1} < C'$ and $\|w^*\|_2 < C$.*

Uniqueness of θ^* follows from Lemma 5 in Koo et al. (2008). Existence of an $s \in \{1, \ldots, p\}$ such that $\theta^*_s \neq 0$ is established in Lemma 4 of Koo et al. (2008).

Assumption 3 *The zero-mean random vector $X = (X_1, \ldots, X_p)^T$ is sub-Gaussian with a variance proxy ν_1^2, $\nu_1 > 0$. This implies that for any $u \in \mathbb{S}^{p-1}$, $\mathbb{P}(|u^T X| > t\nu_1) \leq 2e^{-t^2/2}$.*

Let $\mu_k \triangleq \sup_{u \in \mathbb{S}^{p-1}} \mathbb{E}|u^T X|^k$ and let $\mathring{\mu}_k$ denote the corresponding moments of $\mathring{X}$. Let also $\tilde{\mu}_k \triangleq \sup_{u \in \mathbb{S}^{p-1}} \mathbb{E}\left|(u^T X)/\nu_1\right|^k$. Equivalently, for any $u \in \mathbb{S}^{p-1}$, $u^T X$ is a sub-Gaussian random variable with a variance proxy ν_1^2 and for all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda u^T X)] \leq \exp(\lambda^2 \nu_1^2)/2$ (Wainwright, 2019). While μ_k is used throughout most of our derivations, in the proof of Theorem 10 it is helpful to keep track of the variance proxy, which motivates the use of $\tilde{\mu}_k$.

When the expectation is taken conditional on $Y = 1$, we write μ^f, whereas for $Y = -1$ we write μ^g.

Assumption 4 *The kernel density function, K, is symmetric around zero and uniformly upper-bounded, i.e. for any $u \in \mathbb{R}$, $K(u) \leq \kappa_u$. Moreover, K has finite first and second moments, that is, $\kappa_q \triangleq \int_{\mathbb{R}} |u|^q K(u)du < \infty$ for $q = 1, 2$.*

In particular, Assumption 4 implies that, for a variable U with density function $K(\cdot)$, there exists $u' > 0$ such that for any $u > u' : \mathbb{P}(U > u) \leq Cu^{-\alpha}$, $\alpha \geq 1$, i.e. beyond which the decay in the tails of K is at least linear.

Assumption 5 *Conditional densities f and g are continuous and have finite second moments. Moreover, $\sup_{x\in\mathbb{R}} \max\{f(x|x_{-s}), |x|f(x|x_{-s}), x^2 f(x|x_{-s})\} \leq C$ for some constant $C > 0$. Analogous assumptions are made for $f'(x_s|x_{-s}) \triangleq \frac{df(x_s|x_{-s})}{dx}$ and for the density function $g(\cdot)$.*

Assumption 6 *There exists $x' \in \mathbb{R}$, $x' > 0$, such that for any $x, |x| > x'$, and any α, $|\alpha| \leq 1$, $\exp(\alpha x^2/4\nu_1^2) f(x|x_{-s}) \leq C$ for some $C > 0$.*

Assumption 6 is satisfied by most commonly used distributions and is closely related to sub-Gaussian random variables. To see this, note that for a sub-Gaussian random variable X with a variance proxy ν_1^2, $\mathbb{P}(X \geq t) \leq e\sqrt{8}\mathbb{P}(Z \geq t)$ for $Z \sim N(0, 2\nu_1^2)$ and any $t \geq 0$.

Bahadur representation for the convolution-smoothed SVM relies on the strong convexity of the smoothed loss. The following two assumptions are used to establish the result.

Assumption 7 *There exists $\epsilon > 0$ and $C > 0$ such that $\inf_{|u|\leq\epsilon} K(u) > C$, i.e. the kernel density function is strictly positive in a neighbourhood of zero.*

Assumption 8 *For an orthogonal transformation A_s, $A_s \in \mathbb{R}^{p\times p}$, that maps $w^*/\|w^*\|_2$ to the s-th unit vector e_s for some $1 \leq s \leq d$, there exists $\psi > 0$ and rectangles*

$$\mathcal{D}_*^+(\psi) = \left\{ x \in \mathcal{X} : l_i \leq (A_s x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq s \text{ and } (A_s x)_s \in \frac{1 - b^*}{\|w^*\|_2} + \mathbb{B}(\psi) \right\},$$
$$\mathcal{D}_*^-(\psi) = \left\{ x \in \mathcal{X} : l_i \leq (A_s x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq s \text{ and } (A_s x)_s \in \frac{-1 - b^*}{\|w^*\|_2} + \mathbb{B}(\psi) \right\}$$

such that $f(x) \geq C_1 > 0$ on $\mathcal{D}_*^+$ and $g(x) \geq C_2 > 0$ on $\mathcal{D}_*^-$.

Assumption 8 requires that there exist two (rectangular) subsets of the margins on which conditional densities are bounded away from zero, and is further discussed in Appendix A. Assumption 8 is a direct consequence of Assumption A1 and Assumption A4 in Koo et al. (2008) so Assumption 8 does not impose conditions more restrictive than those needed to establish Bahadur representation of the hinge loss.

4.1 Estimation Error

The first result is an upper bound on $\|\widehat{\theta}_h - \theta^*\|_2$. We henceforth refer to $\|\widehat{\theta}_h - \theta^*\|_2$ as the estimation error and omit the subscript from θ_h.

Theorem 9 (Estimation error) *Under Assumptions 2—5, 7 and 8, for any $t > 0$ and $1 \gtrsim h \gtrsim \sqrt{(p+t)/n}$, $h \leq \min\left\{h_0, \frac{R_1}{C\kappa_2(1+\mathring{\mu}_1)}\right\}$, $h_0 \triangleq \max\left\{C\|w^*\|_2/\epsilon, 1\right\}$, and $n \gtrsim p + t + \log(2)$,*

$$\|\hat{\theta} - \theta^*\|_2 \leq Ch^2\kappa_2(1+\mathring{\mu}_1) + C\sqrt{\frac{p+t+\log(2+2\log(h^{-1}))}{n}}, \tag{6}$$

holds with probability at least $1 - e^{-t}$, where ϵ is defined in Assumption 7.

The exact expression for R_1 can be found in the proof of Proposition 14. Since $h \gtrsim \sqrt{(p+t)/n}$, the leading order in (6) is $\sqrt{(p+t)/n}$. The convergence rate of the $\hat{\theta}_h$, the estimator of θ^* based on the smoothed hinge loss, thus coincides with the rate obtained for the estimator based on the hinge loss (Zhang et al., 2016).

4.2 Wald-type Inference

The following theorem establishes a Bahadur representation for the smoothed estimator $\hat{\theta}_h$.

Theorem 10 (Bahadur representation) *Under Assumptions 2—8, for $n \gtrsim p+t+\log(2)$ and $\sqrt{(2p+t)/n} \lesssim h \lesssim 1$, $h \leq \min\left\{h_0, R_1(C\kappa_2(1+\mathring{\mu}_1))^{-1}\right\}$, $h_0 \triangleq \max\{C\|w^*\|_2/\epsilon, 1\}$, we obtain*

$$\left\|-\nabla\hat{L}_h(\theta^*) - \nabla^2 L_h(\theta^*)(\hat{\theta} - \theta^*)\right\|_2 \leq 6Cr\nu_1^2\sqrt{\frac{2p+t}{nh}} + Cr^2, \tag{7}$$

with probability at least $1 - 2e^{-t}$ for any $t > 0$ and $r \asymp h^2 + \sqrt{(p+t)/n}$.

More precisely, r is the upper-bound on the estimation error from Theorem 9. We can thus restate Theorem 10 in the following more concise form.

Corollary 11 (Bahadur representation) *Under Assumptions 2—8, for $n \gtrsim p + t + \log(2)$ for $\sqrt{(2p+t)/n} \lesssim h \lesssim 1$, $h \leq \min\left\{h_0, R_1(C\kappa_2(1+\mathring{\mu}_1))^{-1}\right\}$, $h_0 \triangleq \max\{C\|w^*\|_2/\epsilon, 1\}$, we obtain*

$$\left\|-\nabla\hat{L}_h(\theta^*) - \nabla^2 L_h(\theta^*)(\hat{\theta} - \theta^*)\right\|_2 \lesssim \sqrt{\frac{p+t}{n}}h^{3/2} + \frac{p+t}{n}\frac{1}{h^{1/2}}. \tag{8}$$

with probability at least $1 - 2e^{-t}$ for any $t > 0$.

Since $\nabla\hat{L}_h(\theta^*)$ is a sum of i.i.d terms (see equation 4) and $\nabla^2 L_h(\theta^*)$ is non-random, the Bahadur representation allows us to approximate $\hat{\theta}_h - \theta^*$ by a sum of i.i.d. terms with high probability, thus enabling us to establish a limiting distribution of the estimator and its functionals.

A key challenge in proving Theorem 9 was establishing that the local strong convexity for the empirical convolution-smoothed loss function holds with high probability in the neighbourhood of θ^*. Since the probability depends on the distribution of X, Assumption 3 is required for Theorems 9 and 10 to hold. A more detailed discussion of strong convexity is provided in Appendix A.

5. Simulations

The distributional approximations of both the SVM estimator derived in Koo et al. (2008) and the smooth SVM estimator presented here hinge on Bahadur representation. The convergence of the distributional approximation is thus governed by the convergence of the Bahadur remainder. We can thus use the non-asymptotic behaviour of the Bahadur remainder to compare the distributional approximation of SVM and smooth SVM in non-asymptotic settings. Since the non-asymptotic bound for the Bahadur remainder has not been derived for SVMs, we resort to simulations. This approach is taken in Section 5.1.

An alternative, used by Koo et al. (2008) for indicating the limitations of SVM in large-p settings, is to compare Type 1 errors as the ratio n/p increases. We pursue this approach in Section 5.2. In Section 5.3, we compare coverage ratios of confidence intervals for the convolution-smoothed estimator and SVM estimator.

As in Koo et al. (2008), we consider Gaussian class-conditional densities with a common covariance matrix throughout all simulations, i.e. $f(X) = N(\mu_f, \Sigma_0)$ and $g(X) = N(\mu_g, \Sigma_0)$, with equal class probabilities. This setting is convenient since θ^* and the Hessian of the population SVM, written $H(\theta^*)$, can be derived analytically (Koo et al., 2008).

5.1 Bahadur Remainder

Since our distributional results, as well as those of Koo et al. (2008), are based on a Bahadur representation, we compare distributional approximations by comparing the L_2 norms of Bahadur remainders of the two methods under a range of n, p settings.

The Bahadur representation of the unpenalised SVM takes the form (Theorem 1 in Koo et al., 2008)

$$\sqrt{n}D(\theta^*)^T(\hat{\theta} - \theta^*) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{1}\{1 - y_i\mathring{x}_i^T\theta^* \geq 0\}y_i\mathring{x}_i + r_{\text{svm}}, \tag{9}$$

where $D(\theta^*)$ is the Hessian matrix of the population hinge loss. Since all population quantities in (9) are known, we can easily calculate the Bahadur remainder r_{svm} for any given n and p. This is not the case for a Bahadur remainder of the smooth SVM (8), as the Hessian depends on a chosen bandwidth and kernel. However, by a triangle inequality, we have an upper bound

$$\|r_{\text{ssvm}}\|_2 \leq \| - \sqrt{n}D(\theta^*)^T(\hat{\theta} - \theta^*) - \sqrt{n}\nabla\hat{L}_h(\theta^*)\|_2. \tag{10}$$

Although this bound may be loose for small n/p, it suffices for current purposes. Equations (9) and (10) enable us to compare the L_2 norms of Bahadur remainders of both SVM and smooth SVM for any n and p.

The usual specification (and implementation) of SVM takes the following form:

$$\hat{L}_{\text{SVM}}(\theta) = C\sum_{i=1}^{n}(1 - y_i\mathring{x}_i^T\theta)_+ + \|w\|_2^2. \tag{11}$$

The theory of Koo et al. (2008) relates to the SVM without regularisation,

$$\hat{L}_{\text{SVM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(1 - y_i \mathring{x}_i^T\theta)_+, \tag{12}$$

solvable by linear programming methods. To this end, introduce slack variables $\xi_1, \ldots, \xi_n$ and solve

$$\min \sum_{i=1}^{n} \xi_i \text{ subject to } \xi_i \geq 0,\ \xi_i \geq 1 - y_i \mathring{x}_i^T\theta,\ i = 1, \ldots, n.$$

We set $\Sigma_0 = c_p I_p$, where I_p denotes a p-dimensional identity matrix, $\mu_1 = (1, \ldots, 1)^T$, a p-dimensional unit vector and $\mu_g = -\mu_f$. The bandwidth is set to the optimal rate from the Bahadur-remainder perspective $h = (p/n)^{1/4}$. The term c_p ensures that as p increases, $\|\theta^*\|_2$ remains unchanged. This ensures that simulation results are not driven by the dependency of Bahadur remainder on $\|\theta^*\|_2$.

Koo et al. (2008) derive the following solutions for population coefficient vector $\theta^* = (b^*, w^*)$,

$$b^* = -\frac{(\mu_f - \mu_g)^T\Sigma^{-1}(\mu_f + \mu_g)}{2a^* d_\Sigma(\mu_f, \mu_g) + d_\Sigma(\mu_f, \mu_g)^2}, \tag{13}$$

$$w^* = \frac{2\Sigma^{-1}(\mu_f - \mu_g)}{2a^* d_\Sigma(\mu_f, \mu_g) + d_\Sigma(\mu_f, \mu_g)^2}, \tag{14}$$

where $d_\Sigma(\mu_f, \mu_g)$ denotes the Mahalanobis distance between the means, i.e., $d_\Sigma(\mu_f, \mu_g) \triangleq [(\mu_f - \mu_g)^T\Sigma^{-1}(\mu_f - \mu_g)]^{1/2}$, $\gamma(x) \triangleq \phi(x)/\Phi(x)$ and $a^* \triangleq \gamma^{-1}(d_\Sigma(\mu_f, \mu_g)/2)$.

From (13) and (14) it follows that c_p is a root of

$$\sqrt{\frac{c_p}{p}}\gamma\left(\frac{2a^* + d_{\Sigma_0}(\mu_f, \mu_g) - \sqrt{p}d_{\Sigma_0}(\mu_f, \mu_g)}{2\sqrt{c_p}}\right) - \frac{d_{\Sigma_0}(\mu_f, \mu_g)}{2}, \tag{15}$$

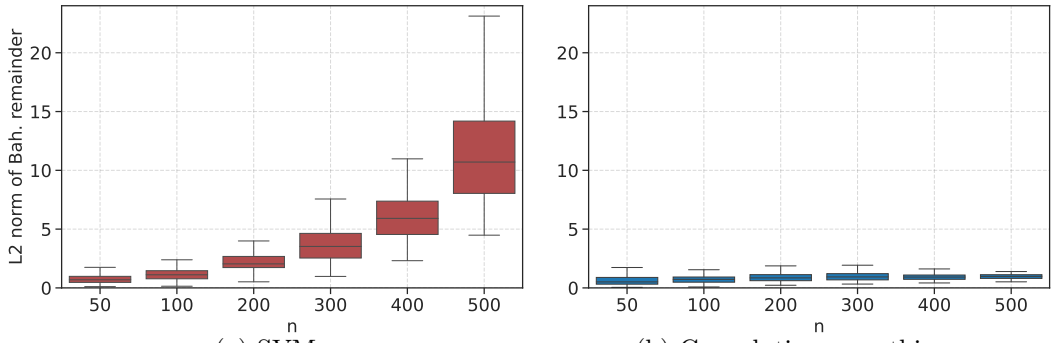
which we solve by Newton-Raphson's method.

Results for a fixed n/p ratio are presented in Figure 2. The L_2 norm of the Bahadur remainder for the smoothed SVM is significantly smaller in both settings, especially for large values of p.

5.2 Type 1 Error Rates

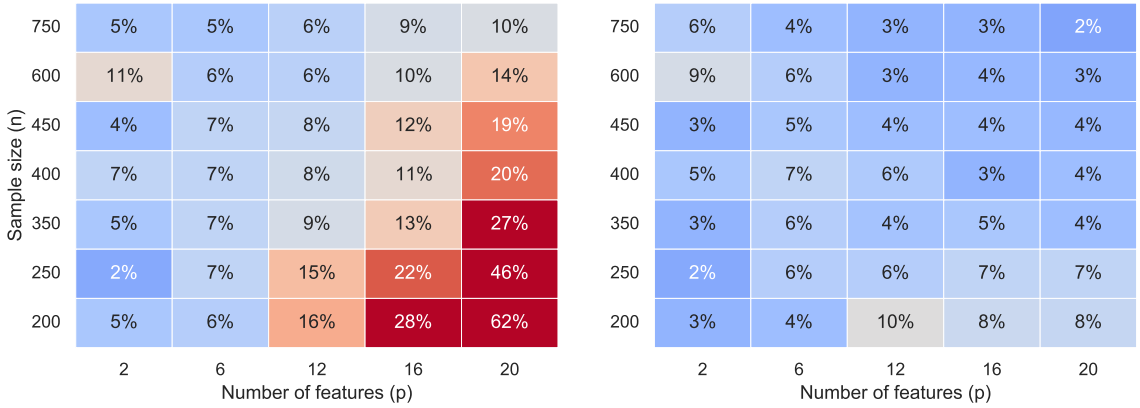
Simulations in Koo et al. (2008) suggest that for large-p settings, the probability of rejecting a hypothesis $H_0 : w_i = 0$ for feature i which has no impact on the response Y exceeds the pre-specified significance level.

We consider the same setting as Koo et al. (2008). Consider p-dimensional mean vectors $\mu_f = (1_{p/2}, 0_{p/2})$, where $1_{p/2}, 0_{p/2}$ denote $p/2$-dimensional vectors of ones and zeros respectively, and $\mu_g = 0$. Then, equation (14) implies that the last $p/2$ coordinates of w^* are zero. For each combination of p and n, we use the asymptotic covariance matrix of the SVM-estimators to calculate Wald statistics, and perform a test of H_0 at the 5% level taking



(a) SVM (b) Convolution-smoothing

Figure 2: The L_2 norm of the Bahadur remainder based on 100 simulations with $n/p = 50$ for SVM (plot a) and Convolution-smoothed SVM (plot b).



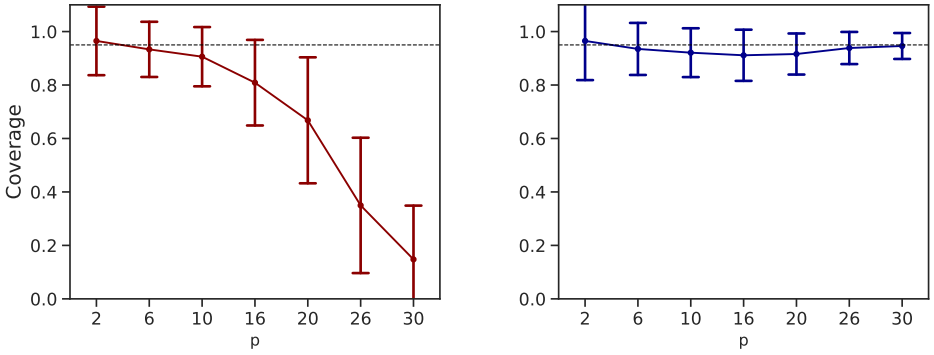
(a) SVM (b) Convolution smoothing

Figure 3: Median value of Type 1 error for testing the significance of a noise variable. Based on 100 simulations for each n, p combination.

$\Phi^{-1}(0.975)$ as the critical value. Over 100 replications we calculate the frequency of false positives for the $p/2$ insignificant variables. A median value of the frequencies is reported for each combination of p and n in Figure 3. The results suggest that Type 1 errors for smooth SVM are substantially closer to the nominal significance level for large-p settings.

5.3 Coverage Ratios

As before, consider p-dimensional mean vectors $\mu_f = (1_{p/2}, 0_{p/2})$ and $\mu_g = 0$. We construct confidence intervals for both zero and non-zero coefficients using a population Hessian derived by Koo et al. (2008). The sample variance of SVM and convolution-smoothed scores



(a) SVM (b) Convolution-smoothing

Figure 4: Median and standard deviation of coverage ratios for SVM (plot a) and convolution-smoothed SVM (plot b). Based on 100 simulations with $n = 500$. The dotted line depicts the target coverage.

are estimated, respectively, by

$$\hat{V}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{1 - y_i x_i^T \theta \geq 0\} x_i x_i^T,$$
$$\hat{V}_h(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \bar{K}^2\left(\frac{1 - y_i x_i^T \theta}{h}\right) x_i x_i^T.$$

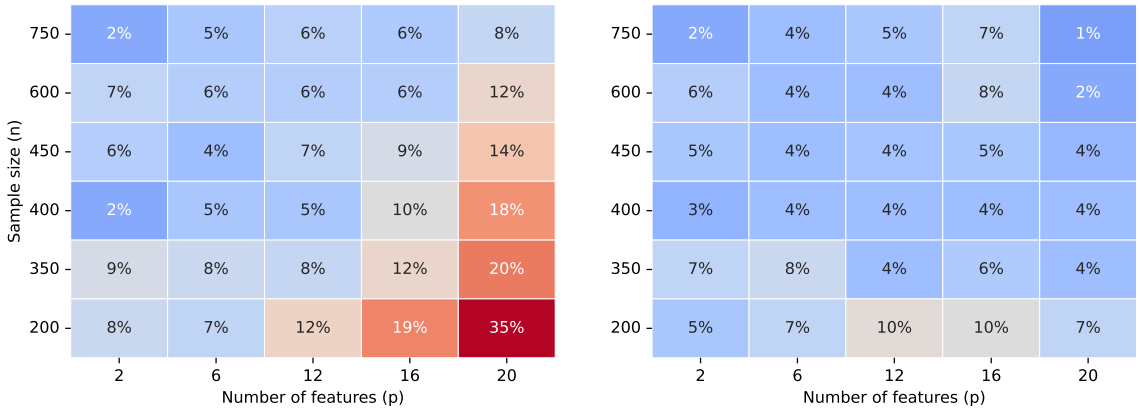
The median and standard deviation of coverage ratios for different choices of p are plotted in Figure 4.

Throughout this section we use Gaussian kernel and set the bandwidth to $h = (p/n)^{1/4}$. Other choices of kernel and bandwidth do not have a significant impact on the results presented in this section, as illustrated in Figure 7 of the Appendix.

6. Comparison to Horowitz-type Smoothing

Using Horowitz-type smoothing of the hinge loss, Wang et al. (2019) establish that as long as the regularisation term is chosen such that $\lambda \asymp \sqrt{p/n}$, the estimation error converges at rate $\sqrt{p/n}$, the same rate as the original hinge loss. As already mentioned, the same rate of convergence is obtained for the convolution-smoothed loss.

To compare the distributional approximations for the two estimators, consider $\sqrt{n}(\hat{\theta} - \theta^*)$. As long as the Bahadur remainder is asymptotically negligible, the limiting distribution of the estimator coincides with the approximation term in (8). Corollary 11 establishes that the remainder is of order $p^{1/2}h^{3/2} + pn^{-1/2}h^{-1/2}$. Minimising this expression over h yields the convergence rate of $p^{7/8}/n^{3/8}$. The distributional approximation is thus valid for $p^{7/3} = o(n)$. In comparison, following an analogous derivation, the optimal rate of convergence of the Bahadur remainder for Horowitz-type smoothing is of order $p^{3/4}n^{-1/4}(\log n)^{1/2}$ (Wang et al., 2019). Hence the distributional approximation holds for $p^3 = o(n(\log n)^{-2})$. The



(a) Horowitz smoothing (b) Convolution smoothing

Figure 5: Median value of Type 1 error for testing the significance of a noise variable. Based on 100 simulations for each n, p combination.

distributional approximation for the convolution-smoothed estimator is thus valid under a weaker condition on p, making it more suitable in large-p settings.

To illustrate this, we compare Type 1 errors of Horowitz-smoothed and convolution-smoothed SVM, with the same setup as in Figure 3. Results are presented in Figure 5.

7. Score-based Inference

The Bahadur representation leads naturally to Wald-type tests and the associated confidence intervals. However, the set of parameters consistent with the data need not form an interval, as exemplified by Fieller's problem (Fieller, 1954). An alternative, which does not suffer from this drawback, is a score-type test.

Score-based confidence sets are obtained by evaluating the gradient of the smoothed loss function at the hypothesised parameter values. Let $\theta^0 \triangleq (\theta_0^0, \ldots, \theta_p^0)$, i.e. the vector of parameters under the null hypothesis. Let $\hat{S}(\theta^0)$ denote the gradient of the convolution-smoothed loss function evaluated at θ^0, that is,

$$\hat{S}(\theta^0) = (\hat{S}_0(\theta^0), \ldots, \hat{S}_p(\theta^0))^T \triangleq n\nabla \widehat{L}_h(\theta^0) = -\sum_{i=1}^{n} y_i \mathring{x}_i \bar{K}\left(\frac{1 - y_i \mathring{x}_i^T \theta^0}{h}\right). \tag{16}$$

For a single coordinate $k \in \{0, \ldots, p\}$, $\hat{S}_k = -\sum_{i=1}^n y_i \mathring{x}_{i,k} \xi_i$, where $\xi_i \triangleq \bar{K}(1 - y_i \mathring{x}_i^T \theta^0/h)$ and $x_{i,k}$ denotes k^{th} coordinate of vector x_i. Note that $\hat{S}_k$ is a sum of i.i.d. terms whose mean, the k^{th} element of $\mathbb{E}[\nabla \hat{L}_h(\theta^0)]$, converges to zero at rate h^2 by Lemma 18. Thus, $n^{-1/2}\hat{S}_k$ is expected to be approximately normally distributed. Moreover, as long as $\sqrt{n}h^2 \to 0$, the

score-based t-statistic, defined as $n^{-1/2}\hat{S}_k$ divided by an estimate of the standard deviation,

$$\hat{T}_k \triangleq \frac{\hat{S}_k}{\hat{\sigma}(S_k)} \triangleq \frac{-n^{-1/2}\sum_{i=1}^n y_i \mathring{x}_{i,k}\xi_i}{\sqrt{(n-1)^{-1}\sum_{i=1}^n \left(-y_i\mathring{x}_{i,k}\xi_i + n^{-1}\sum_{i=1}^n y_i\mathring{x}_{i,k}\xi_i\right)^2}} \tag{17}$$

is asymptotically standard normal. Following Efron (1969), we can rewrite the t-statistics in terms of a self-normalised sum,

$$\hat{T}_k = \frac{\hat{S}_k/\hat{V}_k}{\sqrt{\{n - (\hat{S}_k/\hat{V}_k)^2\}/(n-1)}}, \tag{18}$$

where $\hat{V}_k^2 \triangleq \sum_{i=1}^n (\mathring{x}_{i,k}\xi_i)^2$. Thus, by the asymptotic theory of self-normalised processes (de la Peña et al., 2009) we can define the α-level confidence set as

$$\{\theta^0 : \Phi^{-1/2}(\alpha/2) \leq \hat{T}_k(\theta^0) \leq \Phi^{-1/2}(1-\alpha/2)\}. \tag{19}$$

The construction of score-type confidence sets requires considerably more computation than the Wald construction, as $\hat{T}(\theta^0)$ needs to be evaluated over a grid of points θ^0.

8. Non-linear Extension

We now turn briefly to prediction. In large-p settings the feature space is often sufficiently rich that the restriction to linear models is justifiable on the grounds of prediction accuracy (Hsieh et al., 2008). For data sets with fewer explanatory variables, however, non-linear classifiers such as kernel SVMs, often perform better. We thus extend the convolution-smoothed SVM so that the resulting decision boundary is permitted to be non-linear. This comes at the expense of interpretation emphasised elsewhere in the paper. To this end, consider two matrices $A \in \mathbb{R}^{m\times k}$ and $B \in \mathbb{R}^{k\times l}$. Given the kernel function $\kappa(A,B) \in \mathbb{R}^{m\times l}$, Mangasarian (2000) introduce the following quadratic program,

$$\min \frac{1}{n}\sum_{i=1}^n (1 - y_i(\kappa(x_i, \mathbf{X}^T)\mathbf{y}\omega - \gamma))_+ + \frac{1}{2}\lambda u^T Q u,$$

where $Q \in \mathbb{R}^{n\times n}$ is a symmetric, positive-definite matrix, $u \in \mathbb{R}^n$, $\mathbf{X}$ is $n \times p$ with sample x_i in the i^{th} row, and $\mathbf{y} \triangleq \operatorname{diag}(y_1, \ldots, y_n)$. The standard support vector machine is recovered by setting $Q = \mathbf{y}\kappa(\mathbf{X}, \mathbf{X}^T)\mathbf{y}$ and assuming that $\kappa(\mathbf{X}, \mathbf{X}^T)$ is symmetric and positive semidefinite. If $Q = I$, the assumption of symmetry and positive semidefiniteness of $\kappa(\cdot)$ is no longer required for the problem to possess a solution (Mangasarian, 2000).

We can introduce kernel feature space to convolution-smoothed SVM along similar lines, namely by solving

$$\min \hat{L}_h(\theta) = \frac{1}{nh}\sum_{i=1}^n \int_0^\infty uK\left(\frac{u + y_i(\kappa(x_i, \mathbf{X}^T)\mathbf{y}\omega + \gamma) - 1}{h}\right) du. \tag{20}$$

The linear convolution-smoothed SVM can be recovered by setting $\kappa(\mathbf{X}, \mathbf{X}^T) = \mathbf{X}\mathbf{X}^T$ and $w = \mathbf{X}^T\mathbf{y}\omega$.

The modified objective (20) is twice continuously differentiable and convex, with gradient

$$\nabla \widehat{L}_h(\theta) = -\frac{1}{n}\sum_{i=1}^{n} y_i g_i \bar{K}\left(\frac{1 - y_i(\kappa(x_i, \mathbf{X}^T)\mathbf{y}\omega + \gamma)}{h}\right), \tag{21}$$

where $g_i = \big(\kappa(x_i, \mathbf{X}^T)\mathbf{y}, 1\big)$.

9. Interpretation of Parameters in the Linear SVM

Although the support vector machine was conceived with classification as the objective, the decision boundary is described by parameters for which the paper sought to provide reliable inference. This raises questions of whether the parameters have an interpretation.

For a classification rule $\psi : \mathbb{R}^p \to \mathbb{R}$, the SVM population optimisation problem is

$$\min_{\psi:\mathbb{R}^p\to\mathbb{R}} \mathbb{E}\{1 - Y\psi(x)\}_+, \tag{22}$$

where the expectation is with respect to the conditional distribution of Y given $X = x$. Let $\pi(x) \triangleq \mathbb{P}(Y = 1|X = x)$. It can be shown that the SVM population classification rule is

$$\psi_{\text{SVM}}(x) = \text{sgn}\left(\pi(x) - 1/2\right), \tag{23}$$

which is precisely the Bayes classifier. Thus the general form of the SVM aims to estimate the Bayes rule directly, without necessarily assigning an interpretable form to the conditional class probabilities. Once the so-called hypothesis space of decision functions is specified, however, the implicit form of $\pi(x)$ is recovered. In particular, the linear SVM takes $\psi(x) \in \Psi(x) = \{\text{sgn}(\mathring{x}^T\theta) : \theta \in \mathbb{R}^{p+1}\}$. Thus, with θ^* the minimiser of the SVM population loss function (22) over $\Psi(x)$, $\psi_{\text{SVM}}(x) = \text{sgn}(\mathring{x}^T\theta^*)$ so that (23) gives the implicit model

$$\mathbb{P}(Y = y|X = x) = \tfrac{1}{2} + \alpha y \mathring{x}^T\theta^*, \quad y \in \{-1, 1\}, \tag{24}$$

for some $\alpha > 0$. The linear in probability model, discussed by Cox and Wermuth (1992) and particularly by Battey et al. (2019), is

$$\mathbb{P}(Y = y|X = x) = \tfrac{1}{2}(1 + y\mathring{x}^T\theta^{\text{LPM}}), \quad y \in \{-1, 1\}. \tag{25}$$

From (24) we see that the implicit probability model of the linear SVM is the linear in probability model with $\theta^* = \theta^{\text{LPM}}/2\alpha$. The ratios of coefficients thus coincide, $\theta^*_j/\theta^*_k = \theta^{\text{LPM}}_j/\theta^{\text{LPM}}_k$, and so do the corresponding population-level decision boundaries. The ratio θ^*_j/θ^*_k has a substantive interpretation as the increase in probability of $Y = 1$ resulting from a unit increase in x_j relative to a unit increase in x_k. In other words, the ratio specifies by how much x_j needs to change to have the same effect on the probability of a positive outcome as a unit increase in x_k.

This representation also provides insight into the connection to the logistic model. While the coefficients of different models are not directly comparable, their relative values can be

compared. Consider the logistic model $g(\pi) = x^T\beta$ where $g(\pi) = \log\big(\pi(x)/(1-\pi(x))\big)$. On differentiating both sides with respect to the relevant entry of x,

$$\frac{\beta_j}{\beta_k} = \frac{(d/d\pi)g(\pi)\cdot(\partial\pi(x)/\partial x_j)}{(d/d\pi)g(\pi)\cdot(\partial\pi(x)/\partial x_k)} = \frac{\partial\pi(x)/\partial x_j}{\partial\pi(x)/\partial x_k},$$

and the right hand side is $\theta_j^{\mathrm{LPM}}/\theta_k^{\mathrm{LPM}} = \theta_j^*/\theta_k^*$ under the assumption of a linear in probability model. The logistic and linear models cannot hold simultaneously, so the ratios of their coefficients do not coincide exactly. Nevertheless, ratios of estimated coefficients under the two models estimate the same quantity, as described above.

Acknowledgments

This work was supported by funds from the Engineering and Physical Sciences Research Council (EP/T01864X/1) and from the National Science Foundation (NSF DMS-2401268).

Appendix A. Strong Convexity

To establish strong convexity of the sample smoothed loss function (with high probability) we proceed as follows. Lemma 12 shows that the smoothed loss function is Lipschitz continuous. Proposition 13 then establishes that the smoothed population hinge loss is strongly convex at θ^*, with Proposition 14 extending this to a neighbourhood of θ^*. The remaining step is to show that a sample smoothed loss function inherits the strong convexity of its population counterpart with high probability, which leads to Proposition 15. Proposition 15 is further revised for the case when X is a sub-Gaussian random vector by Proposition 16. The final result is summarised by Proposition 17.

Recall that the empirical hinge loss can be written as $\hat{L}(\theta) = (1/n)\sum_{i=1}^{n} \varphi(1 - y_i \mathring{x}_i^T \theta)$, where $\varphi(u) = u\,\mathbb{1}(u \geq 0)$ is 1-Lipschitz. Hence, $\varphi(1 - y_i \mathring{x}_i^T \theta)$ is 1-Lipschitz in $\mathring{x}_i^T \theta$, i.e. for each sample $(\mathring{x}_i, y_i)$ and any θ, $\theta' \in \mathbb{R}$,

$$|\varphi(1 - y_i \mathring{x}_i^T \theta) - \varphi(1 - y_i \mathring{x}_i^T \theta')| \leq |\mathring{x}_i^T \theta - \mathring{x}_i^T \theta'|.$$

The following lemma establishes that the smoothed hinge loss inherits this Lipschitz property.

Lemma 12 *Under Assumption 4 the smoothed hinge loss,* $l_h(1 - y_i \mathring{x}_i^T \theta) = (\varphi * K_h)(1 - y_i \mathring{x}_i^T \theta)$*, is* $(1/2)$*-Lipschitz in* $\mathring{x}_i^T \theta$*.*

To show that the sample smoothed loss function inherits, with high probability, the strong convexity of its population counterpart in a neighbourhood of θ^*, we first show that the population smoothed loss function is strongly convex at θ^*, a property inherited from the hinge loss. Two additional assumptions, Assumption 7 and 8, are used to establish strong convexity of the smoothed hinge loss.

To gain further insight into Assumption 8 and to simplify the notation later on, let $\mathcal{D}$ denote a rectangle along all coordinates except s, i.e.

$$\mathcal{D}(\psi) = \{x_{-s} \in \mathcal{X}_{-s} : l_i \leq (A_s x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq s\}.$$

Note that, for $\psi = 0$, we can rewrite rectangles as

$$\mathcal{D}_*^+(0) = \left\{x \in M^+ : l_i \leq (A_s x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq s\right\},$$
$$\mathcal{D}_*^-(0) = \left\{x \in M^- : l_i \leq (A_s x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq s\right\}$$

where M^+ and M^- are the margins of the SVM, that is

$$M^+ = \left\{x \in \mathcal{X} | b^* + x^T w^* = 1\right\}, \quad M^- = \left\{x \in \mathcal{X} | b^* + x^T w^* = -1\right\}.$$

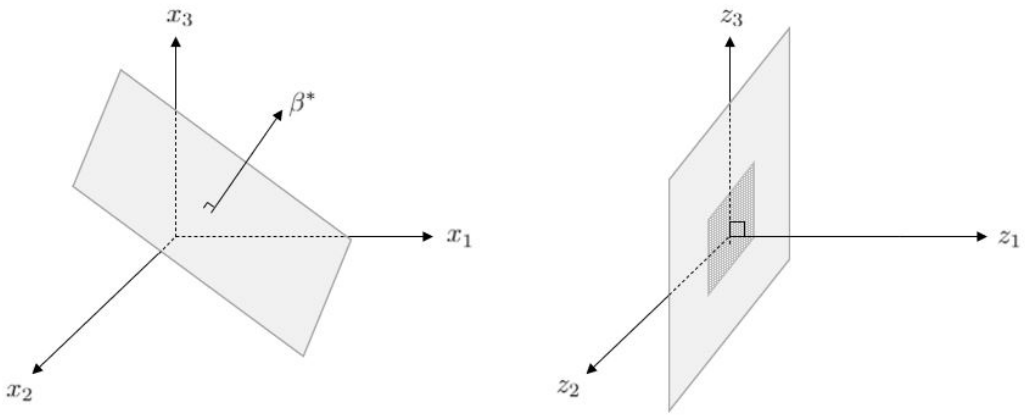
This corresponds to Assumption A4 in Koo et al. (2008). In other words, this assumption requires that there exist two (rectangular) subsets of the margins on which conditional densities are bounded away from zero. Assumption 8 is a direct consequence of Assumption A1 and Assumption A4 in Koo et al. (2008) so Assumption 8 does not impose conditions more restrictive than those needed for the strong convexity of the population hinge loss.

The region for $(A_s x)_s$ in Assumption 8 comes from the following derivation. Let $z = (A_s/\|w^*\|_2)x$, so z becomes the new coordinate system in which w^* becomes e_s. Since w^*

is orthogonal to the separating hyperplane and the margins, the margins lie in a $(p-1)$-dimensional space defined by coordinates z_{-s}.

On the margin, $b^* + w^{*T}x = 1$. On substituting in the expression for z we obtain $b^* + w^{*T}\|w^*\|_2 A_s^T z = 1$. Using the definition of the transformation A_s, this implies that the margin corresponds to a $(p-1)$-dimensional hyperplane that intersects z_s axis at $(1 - b^*)/\|w^*\|_2^2$.

Figure 6 illustrates Assumption 8 for $p = 3$ and $s = 1$.



(A) (B)

Figure 6: Illustration of transformation represented by Assumption 8 for $s = 1$ and $p = 3$. (A) In the case when $p = 3$, the margin is a 2−dimensional hyperplane. Vector w^* is orthogonal to the margin. (B) Transformation A_1 changes the coordinate system such that the margin corresponds to a hyperplane spanned by z_2 and z_3 and the direction along the z_1 axis (corresponding to e_1) is orthogonal to the transformed margin. For $\psi = 0$, the density is assumed to be non-zero for a rectangular subset $\mathcal{D}_*^+(0) = [l_2, v_2] \times [l_3, v_3]$ of the transformed margin, represented by the dark shaded area.

Proposition 13 *Under Assumptions 2, 7 and 8, there exist constants $C, C', C'' > 0$ such that for any $\mathring{\alpha} \in \mathbb{R}^{p+1}$ and for $0 < h \leq \max\{\psi\|w^*\|_2/\epsilon, 1\}$*

$$\mathring{\alpha}^T \nabla^2 L_h(\theta^*)\mathring{\alpha} \geq C\left(\|\mathring{\alpha}\|_2^2 + C'(A_s\alpha)_s^2 h^2\right). \tag{26}$$

For $h \geq \psi\|w^*\|_2/\epsilon$,

$$\mathring{\alpha}^T \nabla^2 L_h(\theta^*)\mathring{\alpha} \geq C''\left(\frac{\|\mathring{\alpha}\|_2^2}{h^2} + C'(A_s\alpha)_s^2\right). \tag{27}$$

Proposition 13 shows that the population smoothed hinge loss has strictly positive definite Hessian at θ^*.

Note that, for $(A_s\alpha)_s = 0$ we recover an inequality for SVM derived in Koo et al. (2008), with identical constant. By the orthogonality of A_s and since $A_s w^*/\|w^*\|_2 = e_s$, the s-th row of A_s is equal to $w^*/\|w^*\|_2$. Hence Proposition 13 establishes a higher lower bound for the curvature of the smooth SVM than that of SVM in all directions, except for the parameter vectors orthogonal to w^*. To bound the approximation error, strong convexity at θ^* is not sufficient, as the estimation error is unlikely to be zero. We thus require strong convexity on the neighbourhood of θ^* instead. Proposition 13 suggests that along directions not orthogonal to θ^* strong convexity holds with a slack, and hence by the continuity of partial derivatives we could argue that along these directions it holds in the neighbourhood of θ^*. Along the directions orthogonal to θ^*, however, the condition for strong convexity might be tight. Proposition 14 shows that despite this, strong convexity in a neighbourhood of θ^* holds.

Proposition 14 *Under Assumptions 2, 7 and 8, there exists* $R_1 > 0$ *such that* $L_h(\theta^* + \Delta)$ *is strongly convex for any* $0 < h \leq \max\{\psi\|w^*\|_2/\epsilon, 1\}$ *and* $\Delta \in \mathbb{B}(hR_1)$. *Specifically, inequalities* (26) *and* (27) *hold for any* $\mathring{\alpha} \in \mathbb{R}^{p+1}$.

Now, we show that in a neighbourhood of θ^* the sample smoothed loss function inherits the strong convexity of the population smoothed loss with high probability. For this, an equivalent characterisation of strong convexity through a first-order Taylor remainder is used.

Let $\mathcal{E}(\Delta)$ be the first-order Taylor series remainder of the sample smoothed loss function,

$$\mathcal{E}(\Delta) = \hat{L}_h(\theta^* + \Delta) - \hat{L}_h(\theta^*) - \langle \nabla \hat{L}_h(\theta^*), \Delta\rangle, \quad \Delta \in \mathbb{R}^{p+1},$$

and let $\bar{\mathcal{E}}(\Delta) = \mathbb{E}\,\mathcal{E}(\Delta)$ be its population counterpart.

Proposition 15 *Under Assumption 4, for any given* $r > 0$ *and* $\delta > 0$,

$$\sup_{\Delta \in \mathbb{B}(r)} |\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| \leq 4r\delta \tag{28}$$

with probability at least $1 - \inf_{\lambda>0} \mathbb{E}[e^{\lambda(\|\bar{X}_n\|_2 - \delta)}]$, *where* $\bar{X}_n \triangleq (1/n)\sum_{i=1}^n \mathring{X}_i$. *Also, for any given* $r_u > r_l > 0$,

$$|\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| \leq 4e^{1/2}\|\Delta\|_2\delta \quad \textit{for all} \quad \Delta \in \mathbb{B}(r_l, r_u), \tag{29}$$

with probability at least $1 - 2\lceil\log(r_u/r_l)\rceil \inf_{\lambda>0} \mathbb{E}[e^{\lambda(\|\bar{X}_n\|_2 - \delta)}]$.

Proposition 14 shows that on $\mathbb{B}(hR_1)$, the population loss function is strongly convex, i.e. $\bar{\mathcal{E}}(\Delta) \geq \kappa\|\Delta\|_2^2$ for some constant $\kappa > 0$. Proposition 15 then shows that, with high probability, the strong convexity also holds for the sample loss function. Since the probability depends on the distribution of X, we now turn to implications of these lemmas for sub-Gaussian variables.

Proposition 16 *Under Assumptions 3 and 4, for any given $r > 0$ and $t > 0$,*

$$\sup_{\Delta \in \mathbb{B}(r)} |\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| \leq Cr\sqrt{\frac{p+t}{n}}$$

with probability at least $1 - e^{-t}$. Also, for any r_l, r_u such that $0 < r_l < r_u$, for any $t > 0$,

$$|\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| \leq C\|\Delta\|_2\sqrt{\frac{p+t}{n}} \text{ for all } \Delta \in \mathbb{B}(r_l, r_u) \tag{30}$$

with probability at least $1 - 2\lceil \log(r_u/r_l) \rceil e^{-t}$.

Finally, for sub-Gaussian random vectors, the following lemma provides a lower-bound on first-order Taylor error.

Proposition 17 (Strong convexity) *Suppose Assumptions 2—4, 7 and 8 hold and $h \leq \max\{\psi\|w^*\|_2/\epsilon, 1\}$. Then, there exists $R_1 > 0$ such that, for any r, $hR_1 > r > 0$, it holds with probability at least $1 - e^{-t}$ that*

$$[\nabla L_h(\theta^* + \Delta) - \nabla L_h(\theta^*)]^T\Delta \geq C\|\Delta\|_2^2 - C'r\sqrt{\frac{p+t}{n}}, \text{ for all } \Delta \in \mathbb{B}(r),$$

for any $t > 0$. Also, for any $0 < r_l < r_u < hR_1$, and $t > 0$,

$$[\nabla L_h(\theta^* + \Delta) - \nabla L_h(\theta^*)]^T\Delta \geq C\|\Delta\|_2^2 - C'\|\Delta\|_2\sqrt{\frac{p+t}{n}}, \text{ for all } \Delta \in \mathbb{B}(r_l, r_u)$$

with probability at least $1 - 2\lceil \log(r_u/r_l) \rceil e^{-t}$.

The exact expression for R_1 can be found in the proof of Proposition 14.

Appendix B. Proofs of Main Results

Proof of Lemma 1

The SVM loss can be written as

$$\hat{L}(\theta) = \mathbb{E}_{\varepsilon \sim \hat{F}(\theta)}[\varepsilon \mathbb{1}\{\varepsilon \geq 0\}],$$

where the random variable ε has distribution function $\hat{F}(\varepsilon, \theta)$. Similarly, convolution-smoothed SVM takes the form

$$\hat{L}_h(\theta) = \mathbb{E}_{\varepsilon \sim \hat{F}_h(\theta)}[\varepsilon \mathbb{1}\{\varepsilon \geq 0\}].$$

The integrated tail probability expectation formula (see e.g. Lo, 2018) yields

$$\mathbb{E}_{\varepsilon \sim \hat{F}_h(\varepsilon, \theta)}[\varepsilon \mathbb{1}\{\varepsilon \geq 0\}] = \int_0^\infty (1 - \hat{F}_h(t))dt.$$

Using the definition of $\hat{F}_h(u)$ and using a repeated substitution of variables ($v = w-1+yx\theta$, $u = t - 1 + yx\theta$),

$$\begin{aligned}
\mathbb{E}_{\varepsilon\sim\hat{F}_h(\varepsilon,\theta)}[\varepsilon\,\mathbb{1}\{\varepsilon \geq 0\}] &= \int_0^\infty \left[\frac{1}{nh}\sum_{i=1}^n \int_t^\infty K\left(\frac{w-1+y_i x_i^T\theta}{h}\right) dw\right] dt \\
&= \int_0^\infty \left[\frac{1}{nh}\sum_{i=1}^n \int_{t-1+yx\theta}^\infty K\left(\frac{v}{h}\right) dv\right] dt \\
&= \frac{1}{nh}\sum_{i=1}^n \int_{-1+y_i x_i\theta}^\infty \left[\int_u^\infty K\left(\frac{v}{h}\right) dv\right] du \\
&= \frac{1}{nh}\sum_{i=1}^n \int_{-\infty}^{1-y_i x_i\theta} \left[\int_{-\infty}^z K\left(\frac{v}{h}\right) dv\right] dz \\
&= \frac{1}{n}\sum_{i=1}^n \int_{-\infty}^{1-y_i x_i^T\theta} \bar{K}_h(z) dz \\
&= \frac{1}{n}\sum_{i=1}^n \bar{\mathcal{K}}_h(1-y_i x_i^T\theta),
\end{aligned}$$

where

$$\bar{K}_h(z) \triangleq \int_{-\infty}^z K\left(\frac{t}{h}\right) dt$$

is the corresponding cumulative distribution function and $\bar{\mathcal{K}}_h$ denotes its indefinite integral. ■

B.1 Proof of Theorem 9 (Estimation error)

Let $\mathbb{B}(t) = \{u \in \mathbb{R}^p : \|u\|_2 \leq t\}$. For some $r_0 > 0$ let

$$\begin{aligned}
\eta &= \sup_\lambda \left\{\lambda \in [0,1] : \lambda\big(\widehat{\theta}_h - \theta^*\big) \in \mathbb{B}(r_0)\right\}, \\
\widetilde{\theta} &= \theta^* + \eta\big(\widehat{\theta}_h - \theta^*\big).
\end{aligned}$$

By the definition of η, $\widetilde{\theta} \in \theta^* + \mathbb{B}(r_0)$. Thus, η is the largest number from $[0,1]$ such that a linear combination of $\widehat{\theta}_h$ and θ^*, with weights η and $1-\eta$ respectively, falls into a neighbourhood of θ^*.

From above it is clear that if $\widehat{\theta}_h \in \theta^* + \mathbb{B}(r_0)$, $\eta = 1$ and $\widetilde{\theta} \in \theta^* + \mathbb{B}(r_0)$, whereas if $\widehat{\theta}_h \notin \theta^* + \mathbb{B}(r_0)$, $\eta < 1$ and $\widetilde{\theta} \in \theta^* + \partial\mathbb{B}(r_0)$.

As in Tan et al. (2022), we first derive an upper bound for $\|\widetilde{\theta} - \theta^*\|_2$ and then, by an appropriate selection of constants argue that this bound is smaller than r_0, i.e. that $\widetilde{\theta}$ lies in the interior of $\theta^* + \mathbb{B}(r_0)$. From the discussion above this implies that $\eta = 1$, so, by the definition of $\widetilde{\theta}$, $\widetilde{\theta} = \widehat{\theta}_h$. We thus have an upper bound for $\|\widehat{\theta}_h - \theta^*\|_2$. To simplify the notation, we omit the subscript h from $\hat{\theta}$ from now onwards.

Define Bregman divergence for a (convex) function $\psi : \mathbb{R}^k \to \mathbb{R}$ as:

$$D_\psi(w, w') = \psi(w') - \psi(w) - \langle \nabla\psi(w), w' - w\rangle$$

and the symmetrised divergence as

$$D^S_\psi(w, w') = D_\psi(w, w') + D_\psi(w', w) = \langle \nabla\psi(w') - \nabla\psi(w), w' - w\rangle. \tag{31}$$

By Lemma 2 (in section C.1) of Sun et al. (2020), for $\psi = \hat{L}_h$,

$$D^S_{\hat{L}_h}(\widetilde{\theta}, \theta^*) \leq \eta D^S_{\hat{L}_h}(\widehat{\theta}, \theta^*). \tag{32}$$

From equation (32) and the definition of symmetrised divergence (31),

$$\langle \nabla\widehat{L}_h(\widetilde{\theta}) - \nabla\widehat{L}_h(\theta^*), \widetilde{\theta} - \theta^*\rangle \leq \eta\langle \nabla\widehat{L}_h(\widehat{\theta}) - \nabla\widehat{L}_h(\theta^*), \widehat{\theta} - \theta^*\rangle. \tag{33}$$

By the definition of $\widehat{\theta}$ we have

$$D^S_{\hat{L}_h}(\widehat{\theta}, \theta^*) = \langle \nabla\widehat{L}_h(\widehat{\theta}) - \nabla\widehat{L}_h(\theta^*), \widehat{\theta} - \theta^*\rangle = -\langle \nabla\widehat{L}_h(\theta^*), \widehat{\theta} - \theta^*\rangle. \tag{34}$$

Combining equations (33) and (34) and using the Cauchy-Schwartz inequality,

$$\|\widetilde{\theta} - \theta^*\|_2^2 \frac{D^S_{\hat{L}_h}(\widetilde{\theta}, \theta^*)}{\|\widetilde{\theta} - \theta^*\|_2^2} \leq -\eta\langle \nabla\widehat{L}_h(\theta^*), \widehat{\theta} - \theta^*\rangle \leq \big\|\nabla\widehat{L}_h(\theta^*)\big\|_2 \big\|\widetilde{\theta} - \theta^*\big\|_2. \tag{35}$$

Thus,

$$\|\widetilde{\theta} - \theta^*\|_2 \leq \frac{\|\nabla\widehat{L}_h(\theta^*)\|_2}{\inf_{\theta \in \theta^* + \mathbb{B}(r_0)} D^S_{\hat{L}_h}(\theta, \theta^*)/\|\theta - \theta^*\|_2^2}. \tag{36}$$

By the triangle inequality,

$$\|\nabla\widehat{L}_h(\theta^*)\|_2 \leq \big\|\nabla\widehat{L}_h(\theta^*) - \mathbb{E}\big[\nabla\widehat{L}_h(\theta^*)\big]\big\|_2 + \big\|\mathbb{E}\big[\nabla\widehat{L}_h(\theta^*)\big]\big\|_2,$$

which upper bounds the numerator.

Lemma 18 (Smoothing bias term) *Under Assumptions 2, 4 and-5 for bandwidth $h > 0$,*

$$\big\|\mathbb{E}\big[\nabla\widehat{L}_h(\theta^*)\big]\big\|_2 \leq \frac{Ch^2\kappa_2(1 + \mathring{\mu}_1)}{2\theta_s^{*2}}.$$

In other words, as bandwidth goes to zero, the smoothed loss better approximates the hinge loss, and hence the population gradient of smoothed loss at θ^* goes to zero.

Lemma 19 (Centred score function) *Under Assumptions 2-5 for any $t > 0$, and $1 \gtrsim h \gtrsim \frac{p+t}{n}$, $n \gtrsim p + t$,*

$$\mathbb{P}\left[\big\|\nabla\widehat{L}_h(\theta^*) - \mathbb{E}\big[\nabla\widehat{L}_h(\theta^*)\big]\big\|_2 \leq C\sqrt{\frac{p+t}{n}}\right] \geq 1 - 2e^{-t}.$$

Thus, using Lemmas 18 and 19, for any $t > 0$,

$$\|\nabla \widehat{L}_h(\theta^*)\|_2 \leq Ch^2\kappa_2(1+\mathring{\mu}_1) + C\sqrt{\frac{p+t}{n}} \tag{37}$$

with probability at least $1 - 2e^{-t}$.

From the definition of the symmetrised divergence (31) it is clear that the denominator in (36) is closely related to the convexity of the sample smoothed loss function. Strong convexity on a neighbourhood $\theta^* + \mathbb{B}(r_0)$ requires that, for some $\kappa > 0$,

$$\forall \theta^1, \theta^2 \in \theta^* + \mathbb{B}(r_0) : \langle \nabla \hat{L}_h(\theta^1) - \nabla \hat{L}_h(\theta^2), \theta^1 - \theta^2 \rangle \geq \kappa \|\theta^1 - \theta^2\|_2^2. \tag{38}$$

Establishing strong convexity over the neighbourhood of θ^* thus amounts to providing a lower-bound for the denominator on the right-hand side of (36).

To this end, let $\mathcal{E}(\Delta)$ be a first order Taylor error of the sample smoothed loss function,

$$\mathcal{E}(\Delta) \triangleq \hat{L}_h(\theta^* + \Delta) - \hat{L}_h(\theta^*) - \langle \nabla \hat{L}_h(\theta^*), \Delta \rangle$$

and $\bar{\mathcal{E}}$ be a Taylor error of the population smoothed loss. Koo et al. (2008) established strong convexity of the population hinge loss function at θ^*. Establishing strong convexity of a sample smoothed hinge loss in the neigborhood of θ^* consists of three steps. Firstly, we show that the smoothed hinge loss is also strongly convex at θ^*. Secondly, we show that this also holds for the neighbourhood of θ^*. Lastly, using Lipschitz continuity of the loss function we show that, with high probability, this property is inherited by the sample smoothed hinge loss. This results in Proposition 17, which we re-state below for ease of reference.

Proposition 20 *Suppose Assumptions 2, 3, 4, 7 and 8 hold and $h \leq \max\{\psi\|w^*\|_2/\epsilon, 1\}$. Then, there exists $R_1 > 0$ such that, for any r, $hR_1 > r > 0$, it holds with probability at least $1 - e^{-t}$ that*

$$[\nabla L_h(\theta^* + \Delta) - \nabla L_h(\theta^*)]^T \Delta \geq C\|\Delta\|_2^2 - C' r\sqrt{\frac{p+t}{n}}, \text{ for all } \Delta \in \mathbb{B}(r),$$

for any $t > 0$. Also, for any $0 < r_l < r_u < hR_1$, and $t > 0$,

$$[\nabla L_h(\theta^* + \Delta) - \nabla L_h(\theta^*)]^T \Delta \geq C\|\Delta\|_2^2 - C'\|\Delta\|_2\sqrt{\frac{p+t}{n}}, \text{ for all } \Delta \in \mathbb{B}(r_l, r_u)$$

with probability at least $1 - 2\lceil \log(r_u/r_l) \rceil e^{-t}$.

The exact expression for R_1 can be found in the proof of Proposition 14. The restriction to the neighbourhood of R_1 arises since only on the ball with this radius, the sample smoothed loss function is (locally) strongly convex.

Let $r_0 = C'h\kappa_2(1+\mathring{\mu}_1)$, and assume that $r_0 \leq hR_1$. Let $r_l = Chr_0$, $r_l < r_0$. Then $\tilde{\Delta} \triangleq \tilde{\theta} - \theta^* \in \mathbb{B}(r_l)$, $\|\tilde{\theta} - \theta^*\|_2 < r_0$, implying $\hat{\theta} = \tilde{\theta}$ and the claimed bound follows. On the other hand, for $\tilde{\Delta} \in \mathbb{B}(r_l, r_0)$ we have, by Proposition 17,

$$\inf_{\Delta \in \mathbb{B}(r_l, r_0)} \frac{D^S(\theta^* + \Delta, \theta^*)}{\|\Delta\|_2^2} = \inf_{\Delta \in \mathbb{B}(r_l, r_0)} \frac{\langle \nabla \hat{L}_h(\theta^* + \Delta) - \nabla \hat{L}_h(\theta^*), \Delta \rangle}{\|\Delta\|_2^2} \geq C - C\frac{1}{\|\tilde{\Delta}\|_2}\sqrt{\frac{p+t}{n}},$$

with probability at least $1 - 2\left\lceil \log(h^{-1}) \right\rceil e^{-t}$. Combining this bound with (35) and (37), we obtain,

$$\|\tilde{\Delta}\|_2 \leq Ch^2\kappa_2(1+\mathring{\mu}_1) + C\sqrt{\frac{p+t}{n}},$$

with probability at least $1-2\left\lceil \log(h^{-1}) \right\rceil e^{-t}-2e^{-t}$ for any $t>0$. Now, we choose bandwidth h such that $\tilde{\theta}$ falls to an interior of $\mathbb{B}(r_0)$, i.e. $\|\tilde{\Delta}\|_2 < r_0$. For this to hold we require

$$Ch^2\kappa_2(1+\mathring{\mu}_1) + C\sqrt{\frac{p+t}{n}} < C'h\kappa_2(1+\mathring{\mu}_1),$$

which holds for $1 \gtrsim h \gtrsim \sqrt{(p+t)/n}$. Then, $\tilde{\Delta}$ falls into the interior of $\mathbb{B}(r_0)$ and $\tilde{\theta} = \hat{\theta}$. We thus obtain,

$$\|\hat{\Delta}\|_2 \leq Ch^2\kappa_2(1+\mathring{\mu}_1) + C\sqrt{\frac{p+t}{n}},$$

with probability at least $1 - 2\left\lceil \log(h^{-1}) \right\rceil e^{-t} - 2e^{-t}$ for any $t>0$. The result follows by setting $t = t' + \log 2(1+\log(h^{-1}))$ for $t'>0$. ■

B.2 Proof of Theorem 10 (Bahadur representation)

Consider $r \asymp h^2 + \sqrt{\frac{p+t}{n}}$. Then by Theorem 9, $\hat{\theta} \in \theta^* + \mathbb{B}(r)$ with probability at least $1-e^{-t}$ for any $t>0$. Conditioning on such an event leads us to consider $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathbb{B}(r)$. Consider

$$\Lambda(\hat{\Delta}) \triangleq \nabla\hat{L}_h(\theta^* + \hat{\Delta}) - \nabla\hat{L}_h(\theta^*) - \nabla^2 L_h(\theta^*)\hat{\Delta}. \tag{39}$$

By the optimality of $\hat{\theta}$ the first term is zero. As $\hat{\theta}$ belongs to a neighbourhood $\mathbb{B}(r)$ of θ^* with high probability, it suffices to bound $\sup_{\Delta\in\mathbb{B}(r)} \|\Lambda(\Delta)\|_2$. By the triangle inequality,

$$\sup_{\Delta\in\mathbb{B}(r)} \|\Lambda(\Delta)\|_2 \leq \sup_{\Delta\in\mathbb{B}(r)} \|\mathbb{E}\Lambda(\Delta)\|_2 + \sup_{\Delta\in\mathbb{B}(r)} \|\Lambda(\Delta) - \mathbb{E}\Lambda(\Delta)\|_2.$$

The first term, $\sup_{\Delta\in\mathbb{B}(r)} \|\mathbb{E}\Lambda(\Delta)\|_2$, can be upper-bounded using the Taylor remainder from expansion of $\nabla\hat{L}_h(\theta^*+\Delta)$ around $\nabla\hat{L}_h(\theta^*)$. Specifically,

$$\mathbb{E}\left[\Lambda(\Delta)\right] = \int_0^1 \nabla^2 L_h(\theta^* + v\Delta)\Delta - \nabla^2 L_h(\theta^*)\Delta dv.$$

Hence

$$\|\mathbb{E}\left[\Lambda(\Delta)\right]\|_2 \leq \int_0^1 \|\nabla^2 L_h(\theta^*+v\Delta) - \nabla^2 L_h(\theta^*)\|_2 \, \|\Delta\|_2 dv,$$

for any Δ in the local neighbourhood of zero. The following lemma provides an upper-bound for $\sup_{\Delta\in\mathbb{B}(r)} \|\nabla^2 L_h(\theta^*+\Delta) - \nabla^2 L_h(\theta^*)\|_2$.

Lemma 21 *Under Assumptions 2-5 and 7-8 for any $t > 0$, for $1 \gtrsim h \gtrsim \sqrt{(p+t)/n}$, $h \leq \min\left\{h_0, R_1(C\kappa_2(1+\mathring{\mu}_1))^{-1}\right\}$, $h_0 \triangleq \max\{C\|w^*\|_2/\epsilon, 1\}$ and $n \gtrsim p + t + \log(2)$, and $\Delta \in \mathbb{B}(r)$,*

$$\|\nabla^2 L_h(\theta^* + \Delta) - \nabla^2 L_h(\theta^*)\|_2 \leq Cr,$$

where C depends on the moments of X, κ_1 and $\|\theta^*\|_2$.

Using Lemma 21,

$$\sup_{\Delta \in \mathbb{B}(r)} \|\mathbb{E}\Lambda(\Delta)\|_2 \leq Cr^2.$$

Consider now the term $\sup_{\Delta\in\mathbb{B}(r)} \|\Lambda(\Delta) - \mathbb{E}\Lambda(\Delta)\|_2$. Introduce a zero-mean gradient process $G(\theta) \triangleq \nabla\hat{L}_h(\theta) - \nabla L_h(\theta)$ and observe that

$$\begin{aligned}\Lambda(\Delta) - \mathbb{E}\Lambda(\Delta) &= \left[\nabla\hat{L}_h(\theta^* + \Delta) - \nabla L_h(\theta^* + \Delta)\right] - \left[\nabla\hat{L}_h(\theta^*) - \nabla L_h(\theta^*)\right] \\ &= G(\theta^* + \Delta) - G(\theta^*).\end{aligned}$$

Hence

$$\sup_{\Delta\in\mathbb{B}(r)} \|\Lambda(\Delta) - \mathbb{E}\Lambda(\Delta)\|_2 = \sup_{\Delta\in\mathbb{B}(r)} \|G(\theta^* + \Delta) - G(\theta^*)\|_2 \triangleq \sup_{\Delta\in\mathbb{B}(r)} \|\Lambda_0(\Delta)\|_2.$$

We now use Theorem 3.2 in Spokoiny (2013) to bound the term above.

Taking the gradient

$$\begin{aligned}\nabla\Lambda_0(\Delta) &= \nabla^2\hat{L}_h(\theta^* + \Delta) - \nabla^2 L_h(\theta^* + \Delta) \\ &= \sum_{i=1}^n \frac{1}{nh}\mathring{X}_i\mathring{X}_i^T K\left(\frac{1 - Y_i\mathring{X}_i^T(\theta^* + \Delta)}{h}\right) - \mathbb{E}\left[\frac{1}{h}\mathring{X}\mathring{X}^T K\left(\frac{1 - Y\mathring{X}^T\theta^* - Y\mathring{X}^T\Delta}{h}\right)\right] \\ &\triangleq \sum_{i=1}^n \Lambda_0^i(\Delta). \end{aligned} \tag{40}$$

In order to verify condition (A.4) in Spokoiny (2013), we need, for any α, $\alpha' \in \mathbb{S}^p$, an upper bound

$$\sup_{\Delta\in\mathbb{B}(r)} \log\mathbb{E}\exp\left[\lambda\alpha^T\nabla\Lambda_0(\Delta)\alpha'\right] \leq \frac{\nu_0^2\lambda^2}{2}, \;\; \lambda^2 \leq 2g^2.$$

Hence consider

$$\mathbb{E}\exp\left[n^{1/2}\lambda\alpha^T\nabla\Lambda_0(\Delta)\alpha'/\nu_1^2\right] = \prod_{i=1}^n \mathbb{E}\exp\left[n^{1/2}\lambda\alpha^T\nabla\Lambda_0^i(\Delta)\alpha'/\nu_1^2\right], \tag{41}$$

where the equality follows since $\Delta \in \mathbb{B}(r)$ and does not depend on X. Since $e^u \leq 1 + u + \frac{u^2 e^{|u|}}{2}$ we obtain,

$$\begin{aligned}&\mathbb{E}\exp\left[n^{1/2}\lambda\alpha^T\nabla\Lambda_0^i(\Delta)\alpha'/\nu_1^2\right] \\ &\leq \mathbb{E}\left\{1 + \lambda[n^{1/2}\alpha^T\nabla\Lambda_0^i(\Delta)\alpha'/\nu_1^2] + \frac{\lambda^2}{2\nu_1^4}[n^{1/2}\alpha^T\nabla\Lambda_0^i(\Delta)\alpha']^2\exp\left(\frac{|\lambda|}{\nu_1^2}|n^{1/2}\alpha^T\nabla\Lambda_0^i(\Delta)\alpha'|\right)\right\},\end{aligned}$$

where the second term is zero by the definition of $\Lambda(\Delta)$. Hence,

$$\begin{aligned}
&\mathbb{E}\exp\left[\lambda n^{1/2}\alpha^T\nabla\Lambda_0^i(\Delta)\alpha/\nu_1^2\right]\\
&\quad\le \mathbb{E}\left\{1+\frac{\lambda^2}{\nu_1^4}\frac{1}{2}\left[n^{1/2}\alpha^T\nabla\Lambda_0^i(\Delta)\alpha\right]^2\exp\left(\left|\frac{\lambda}{\nu_1^2}n^{1/2}\alpha^T\nabla\Lambda_0^i(\Delta)\alpha\right|\right)\right\}\\
&\quad\le 1+\frac{\lambda^2}{2n\nu_1^4h^2}\mathbb{E}\big(\alpha^T\mathring{X}_i\alpha'^T\mathring{X}_iK_i-\mathbb{E}\big[\alpha^T\mathring{X}\alpha'^T\mathring{X}K\big]\big)^2e^{\frac{|\lambda|}{hn^{1/2}\nu_1^2}(|\alpha^T\mathring{X}_i\alpha'^T\mathring{X}_iK_i|+|\mathbb{E}(\alpha^T\mathring{X}\alpha'^T\mathring{X}K)|}
\end{aligned}$$

where we used the notation $K_i \triangleq K\left((1-Y_i\mathring{X}_i^T(\theta^*+\Delta))/h\right)$ and $K \triangleq K\left((1-Y\mathring{X}^T(\theta^*+\Delta))/h\right)$. The last term can be simplified using a uniform upper-bound on the kernel density (Assumption 4),

$$\exp\mathbb{E}\left[\frac{|\lambda|}{n^{1/2}\nu_1^2}\frac{1}{h}\alpha^T\mathring{X}\alpha'^T\mathring{X}K\left(\frac{1-Y\mathring{X}^T\theta^*-\mathring{X}^T\Delta}{h}\right)\right]\le\exp\left(\frac{|\lambda|\kappa_u}{hn^{1/2}\nu_1^2}\mathbb{E}\left[\alpha^T\mathring{X}\alpha'^T\mathring{X}\right]\right). \tag{42}$$

By the Cauchy-Schwartz inequality,

$$\mathbb{E}\left[\frac{\alpha^T\mathring{X}\alpha'^T\mathring{X}}{\nu_1^2}\right]\le\left\{\mathbb{E}\left[\left(\alpha^T\mathring{X}\right)^2/\nu_1^2\right]\mathbb{E}\left[\left(\alpha'^T\mathring{X}\right)^2/\nu_1^2\right]\right\}^{1/2}\le\tilde{\mu}_2, \tag{43}$$

we obtain

$$\exp\mathbb{E}\left[\frac{|\lambda|}{\nu_1^2n^{1/2}}\frac{1}{h}\alpha^T\mathring{X}\alpha'^T\mathring{X}K\left(\frac{1-Y\mathring{X}^T\theta^*-\mathring{X}^T\Delta}{h}\right)\right]\le\exp\left(\frac{|\lambda|\kappa_u}{hn^{1/2}}\tilde{\mu}_2\right).$$

On letting $|\lambda|\le h\sqrt{n}/\kappa_u$, we obtain

$$\exp\mathbb{E}\left[\frac{|\lambda|}{\nu_1^2n^{1/2}}\frac{1}{h}\alpha^T\mathring{X}\alpha'^T\mathring{X}K\left(\frac{1-Y\mathring{X}^T\theta^*-\mathring{X}^T\Delta}{h}\right)\right]\le e^{\tilde{\mu}_2}. \tag{44}$$

Hence, using $-2ab\le a^2+b^2$,

$$\begin{aligned}
&\mathbb{E}\exp\left[\lambda n^{1/2}\alpha^T\nabla\Lambda_0^i(\Delta)\alpha/\nu_1^2\right]\\
&\qquad\le 1+\frac{\lambda^2}{n\nu_1^4}e^{\tilde{\mu}_2}\mathbb{E}\left[\left(\frac{1}{h}\alpha^T\mathring{X}_i\alpha'^T\mathring{X}_iK_i\right)^2\exp\left(\frac{|\lambda|}{n^{1/2}\nu_1^2}\left|\frac{1}{h}\alpha^T\mathring{X}_i\alpha'^T\mathring{X}_iK_i\right|\right)\right]\\
&\qquad+\frac{\lambda^2}{n\nu_1^4}e^{\tilde{\mu}_2}\left(\mathbb{E}\left[\frac{1}{h}\alpha^T\mathring{X}\alpha'^T\mathring{X}K\right]\right)^2\mathbb{E}\left[\exp\left(\frac{|\lambda|}{n^{1/2}\nu_1^2}\left|\frac{1}{h}\alpha^T\mathring{X}_i\alpha'^T\mathring{X}_iK_i\right|\right)\right].
\end{aligned} \tag{45}$$

Note that for any $u\in\mathbb{S}^p$, $\mathbb{E}[\exp(\langle u,\mathring{X}\rangle^2)/(8\nu_1^2)]\le\mathbb{E}[\exp(u_0^2/(4\nu_1^2)+\langle u,x\rangle^2/(4\nu_1^2))]$. Let $Z\triangleq\langle u,x\rangle^2/4\nu_1^2$. Then $\mathbb{P}(Z\ge t)\le 2e^{-2t}$ for any $t>0$ and $\mathbb{E}(e^Z)\le 3$ and $\mathbb{E}(Z^2e^Z)\le 8$. Thus,

$$\mathbb{E}\left[\exp(\langle u,\mathring{X}\rangle^2/(8\nu_1^2))\right]\le 3\exp(1/(4\nu_1^2)).$$

Consider now the third term on the right-hand side of (45). By the Cauchy-Schwartz inequality and by letting $|\lambda| \leq \frac{1}{8}\frac{h\sqrt{n}}{\kappa_u}$,

$$\mathbb{E}\exp\left(\frac{\kappa_u|\lambda|}{h\sqrt{n}}\frac{8}{8\nu_1^2}|\alpha^T\mathring{X}\alpha'^T\mathring{X}|\right) \leq \left(\mathbb{E}\left[e^{\frac{(\alpha^T\mathring{X})^2}{8\nu_1^2}}\right]\mathbb{E}\left[e^{\frac{(\alpha'^T\mathring{X})^2}{8\nu_1^2}}\right]\right)^{1/2} \leq 3e^{1/(4\nu_1^2)}. \tag{46}$$

On the other hand using Assumptions 2 - 5, we can easily obtain,

$$\mathbb{E}\left[\frac{1}{h\nu_1^2}K|\langle\alpha,\mathring{X}\rangle\langle\alpha',\mathring{X}\rangle|\right] \leq \frac{C}{\nu_1^2}\left[1+2\bar{\mu}_1+\bar{\mu}_2\right], \tag{47}$$

where $\bar{\mu}_k \triangleq \max\{\mu_k^f, \mu_k^g\}$.

Combining (46) and (47), for $|\lambda| \leq \frac{1}{8}\frac{h\sqrt{n}}{\kappa_u}$,

$$\left(\mathbb{E}\left[\frac{1}{\nu_1^2 h}\alpha^T\mathring{X}\alpha'^T\mathring{X}K\right]\right)^2\mathbb{E}\exp\left(\frac{|\lambda|}{n^{1/2}\nu_1^2}\left|\frac{1}{h}\alpha^T\mathring{X}_i\alpha'^T\mathring{X}_iK_i\right|\right) \leq 3e^{1/(4\nu_1^2)}\frac{C}{\nu_1^4}\left[1+2\bar{\mu}_1+\bar{\mu}_2\right]^2. \tag{48}$$

Turning now to the second term in (45), note that for any $z \geq 0$, and any $t > 0$, $m \geq 0$, $z^m \leq (m/(te))^m e^{tz}$. Thus (use $t = 1/16$, $m = 2$),

$$\left(\frac{\langle\alpha,\mathring{X}_i\rangle\langle\alpha',\mathring{X}_i\rangle}{\nu_1^2}\right)^2 \leq (32/e)^2\exp\left(\frac{|\langle\alpha,\mathring{X}_i\rangle\langle\alpha',\mathring{X}_i\rangle|}{16\nu_1^2}\right). \tag{49}$$

For $|\lambda| \leq \frac{1}{16}\frac{\sqrt{n}h}{\kappa_u}$, using Assumption 4 (first inequality), equation (49) and Assumption 4 (second inequality), $ab \leq a^2/2 + b^2/2$ (third inequality), the Cauchy-Schwartz inequality (penultimate inequality), and Assumptions 3, 4 and 6 (last inequality) we obtain,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{h\nu_1^2}\langle\alpha,\mathring{X}_i\rangle\langle\alpha',\mathring{X}_i\rangle K_i\right)^2\exp\left(\frac{|\lambda|}{n^{1/2}\nu_1^2}\frac{1}{h}\left|\langle\alpha,\mathring{X}_i\rangle\langle\alpha',\mathring{X}_i\rangle K_i\right|\right)\right] \\
&\leq \mathbb{E}\left[\left(\frac{1}{h\nu_1^2}\langle\alpha,\mathring{X}_i\rangle\langle\alpha',\mathring{X}_i\rangle K_i\right)^2\exp\left(\frac{1}{16\nu_1^2}\left|\langle\alpha,\mathring{X}_i\rangle\langle\alpha',\mathring{X}_i\rangle\right|\right)\right] \\
&\leq \frac{\kappa_u}{h^2}\left(\frac{32}{e}\right)^2\mathbb{E}\left[K_i\exp\left(\frac{1}{8\nu_1^2}\left|\langle\alpha,\mathring{X}_i\rangle\langle\alpha',\mathring{X}_i\rangle\right|\right)\right] \\
&\leq \left(\frac{32}{e}\right)^2\frac{\kappa_u}{h^2}\mathbb{E}\left[K_i\exp\left(\frac{1}{16\nu_1^2}\left(\langle\alpha,\mathring{X}_i\rangle^2+\langle\alpha',\mathring{X}_i\rangle^2\right)\right)\right] \\
&\leq \left(\frac{32}{e}\right)^2\frac{\kappa_u}{h^2}\mathbb{E}\left[K_i\exp\left(\frac{1}{8\nu_1^2}\langle\alpha,\mathring{X}_i\rangle^2\right)\right]^{1/2}\mathbb{E}\left[K_i\exp\left(\frac{1}{8\nu_1^2}\langle\alpha',\mathring{X}_i\rangle^2\right)\right]^{1/2} \\
&\leq Ch^{-1}.
\end{aligned} \tag{50}$$

Combining (45), (48) and (50), and using $h \lesssim 1$,

$$\mathbb{E}\exp\left[\lambda n^{1/2}\alpha^T\nabla\Lambda_0^i(\Delta)\alpha/\nu_1^2\right] \leq 1+\frac{C^2\lambda^2}{2nh} \leq \exp\left(\frac{C^2\lambda^2}{2nh}\right).$$

Substituting into equation (41) yields

$$\mathbb{E}\exp\left[n^{1/2}\lambda\alpha^T\nabla\Lambda_0(\Delta)\alpha'/\nu_1^2\right] \leq \exp\left(\frac{C^2\lambda^2}{2h}\right),$$

which verifies the condition (A.4) in Spokoiny (2013), with $\nu_0 = \frac{C}{h^{1/2}}$ and $\lambda \leq \frac{1}{16}\frac{h\sqrt{n}}{\kappa_u}$. Let $g = \frac{h}{\kappa_u}\sqrt{\frac{n}{2}}$. By the requirement $4p + 2t \leq g^2$ in Spokoiny (2013), we obtain the restriction $h \geq 2\kappa_u\sqrt{\frac{2p+t}{n}}$. Hence,

$$\mathbb{P}\left\{\sup_{\Delta\in\mathbb{B}(r)} \|\Lambda(\Delta) - \mathbb{E}\Lambda(\Delta)\|_2 \geq 6Cr\nu_1^2\sqrt{\frac{2p+t}{nh}}\right\} \leq e^{-t}.$$

Overall we obtain,

$$\sup_{\Delta\in\mathbb{B}(r)} \|\Lambda(\Delta)\|_2 \leq 6Cr\nu_1^2\sqrt{\frac{2p+t}{nh}} + Cr^2,$$

with probability at least $1 - 2e^{-t}$. ■

B.3 Proof of Strong convexity

B.3.1 Proof of Lemma 12

For any $y \in \{\pm 1\}$, $\mathring{x} \in \mathbb{R}^{p+1}$ and $\theta, \theta' \in \mathbb{R}^{p+1}$,

$$\begin{aligned}
&|l_h(1 - y\mathring{x}^T\theta) - l_h(1 - y\mathring{x}^T\theta')| \\
&= \frac{1}{h}\left|\int_0^\infty uK\left(\frac{u + y\langle\mathring{x},\theta\rangle - 1}{h}\right)du - \int_0^\infty uK\left(\frac{u + y\langle\mathring{x},\theta'\rangle - 1}{h}\right)du\right| \\
&= \left|\int_0^\infty (wh + 1 - y\langle\mathring{x},\theta\rangle)K(w)dw - \int_0^\infty (w'h + 1 - y\langle\mathring{x},\theta'\rangle)K(w')du\right| \\
&= \left|\int_0^\infty y\langle\mathring{x},\theta - \theta'\rangle K(w)dw\right| \\
&\leq |\mathring{x}^T(\theta - \theta')| \int_0^\infty K(w)\,dw \\
&= \frac{1}{2}|\mathring{x}^T(\theta - \theta')|,
\end{aligned}$$

where the last equality uses the fact that $K(\cdot)$ is symmetric around zero and integrates to one. ■

B.3.2 Proof of Proposition 13

This proof is based on a similar argument to that of Lemma 5 in Koo et al. (2008), with modifications to account for the form of the smoothed loss.

Without loss of generality we can assume that $s = 1$. Let $\mathring{\alpha} = (\alpha_0, \alpha^T)^T \in \mathbb{R}^{p+1}$ with $\alpha \in \mathbb{R}^p$ and $\theta^* = (b^*, w^{*T})^T \in \mathbb{R}^{p+1}$. Then,

$$\begin{aligned}\mathring{\alpha}^T \nabla^2 L_h(\theta^*) \mathring{\alpha} &= \mathring{\alpha}^T \, \mathbb{E}\left[\frac{1}{h} K\left(\frac{1 - Y\langle \mathring{X}, \theta^* \rangle}{h}\right) \mathring{X} \mathring{X}^T\right] \mathring{\alpha} \\ &= \bar{\pi}_+ \int_{\mathcal{X}} \frac{1}{h} K\left(\frac{1 - \langle \mathring{x}, \theta^* \rangle}{h}\right) \langle \mathring{\alpha}, \mathring{x} \rangle^2 f(x) dx \\ &+ \bar{\pi}_- \int_{\mathcal{X}} \frac{1}{h} K\left(\frac{1 + \langle \mathring{x}, \theta^* \rangle}{h}\right) \langle \mathring{\alpha}, \mathring{x} \rangle^2 g(x) dx.\end{aligned}$$

Let $Z = A_1 X / \|w^*\|_2$ and introduce $a \triangleq A_1 \alpha \in \mathbb{R}^p$. Then $X = \|w^*\|_2 A_1^T Z$,

$$\mathring{\alpha}^T \mathring{X} = \alpha_0 + \|w^*\|_2 Z^T A_1 \alpha = \alpha_0 + \|w^*\|_2 Z^T a$$

and

$$1 - \mathring{X}^T \theta^* = 1 - b^* - \|w^*\|_2 Z^T A_1 w^* = 1 - b^* - Z^T e_1 \|w^*\|_2^2.$$

The Jacobian of the transformation $x \mapsto z = A_1 x / \|w^*\|_2$ is $\|w^*\|_2^p$. Hence,

$$\begin{aligned}&\int_{\mathcal{X}} \frac{1}{h} K\left(\frac{1 - \mathring{x}^T \theta^*}{h}\right) \langle \mathring{\alpha}, \mathring{x} \rangle^2 f(x) dx \\ &= \int_{\mathcal{Z}} \frac{1}{h} K\left(\frac{1 - b^* - \|w^*\|_2^2 z_1}{h}\right) \left(\alpha_0 + \|w^*\|_2 z^T a\right)^2 f(\|w^*\|_2 A_1^T z) \|w^*\|_2^p dz \\ &= \|w^*\|_2^{p-2} \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} K(q) \left(\alpha_0 + \|w^*\|_2 a_1 \frac{1 - b^* + qh}{\|w^*\|_2^2} + \|w^*\|_2 a_{-1}^T z_{-1}\right)^2 \\ &\quad f\left(\|w^*\|_2 A_1^T \left(\frac{1 - b^* + qh}{\|w^*\|_2^2}, z_{-1}^T\right)^T\right) dq dz_{-1} \\ &= \frac{1}{\|w^*\|_2} \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} K(q) \left(\alpha_0 + a_1 \frac{1 - b^* + qh}{\|w^*\|_2} + \sum_{j=2}^p u_j a_j\right)^2 f\left(A_1^T \left(\frac{1 - b^* + qh}{\|w^*\|_2}, u^T\right)^T\right) dq du.\end{aligned}$$

The second equality follows from a change of variables, $q = -(1 - b^* - \|w^*\|_2^2 z_1)/h$, which implies $z_1 = (1 - b^* + qh)/\|w^*\|_2^2$ and $dz_1 = (h/\|w^*\|_2^2) dq$. The third equality uses another change of variables, namely $u_j = \|w^*\|_2 z_j$, for $j = 2, \ldots, p$. The Jacobian of this transformation is $\|w^*\|_2^{-(p-1)}$.

For any $\psi > 0$ let $\mathcal{D}_u^+ = \left\{u : A_1^T \left(\frac{1-b^*+qh}{\|w^*\|_2}, u^T\right)^T \in \mathcal{D}_*^+(\psi)\right\}$. By the orthogonality of A_1 this is just an inverse transformation back to the original coordinate system x. Using

Assumptions 7 and 8,

$$\begin{aligned}
&\int_{\mathcal{X}} \frac{1}{h} K\left(\frac{1-\mathring{x}^T\theta^*}{h}\right)\langle\mathring{\alpha},x\rangle^2 f(x)dx \\
&\quad\geq \frac{1}{\|w^*\|_2} C \int_{\mathcal{D}_u^+}\int_{-\frac{\psi\|w^*\|_2}{h}}^{\frac{\psi\|w^*\|_2}{h}} K(q)\left[\alpha_0 + a_1\frac{1-b^*+qh}{\|w^*\|_2} + \sum_{j=2}^p u_j a_j\right]^2 dq\, du \\
&\quad\geq \frac{1}{\|w^*\|_2} C \int_{\mathcal{D}_u^+}\int_{-\bar{\epsilon}}^{\bar{\epsilon}}\left[\alpha_0 + a_1\frac{1-b^*+qh}{\|w^*\|_2} + \sum_{j=2}^p u_j a_j\right]^2 dq\, du,
\end{aligned}$$

where $\bar{\epsilon} = \min\{\psi\|w^*\|_2/h, \epsilon\}$. Note that $\bar{\epsilon}$ depends on h. This dependence disappears for $h \leq \psi\|w^*\|_2/\epsilon$, as this implies $\bar{\epsilon} = \epsilon$ and for $h \leq 1$, in which case we can take $\bar{\epsilon} = \min\{\psi\|w^*\|_2, \epsilon\}$. Consider $h \leq \max\{\psi\|w^*\|_2/\epsilon, 1\}$. We return to the case of large h later on.

Introduce independent uniform random variables $U_2, \ldots, U_p$ and Q, where $U_j \stackrel{d}{=} \text{Unif}(l_j, v_j)$, $j = 2, \ldots, p$, i.e. $(U_2, \ldots, U_p)^T \in \mathcal{D}_u^+$, and $Q \stackrel{d}{=} \text{Unif}(-\bar{\epsilon}, \bar{\epsilon})$. Then,

$$\begin{aligned}
&\int_{\mathcal{X}} \frac{1}{h} K\left(\frac{1-\mathring{x}^T\theta^*}{h}\right)\langle\mathring{\alpha},\mathring{x}\rangle^2 f(x)dx \geq \frac{C\text{vol}(\mathcal{D}_u^+)2\bar{\epsilon}}{\|w^*\|_2}\mathop{\mathbb{E}}_{Q,U}\left[\alpha_0 + a_1\frac{1-b^*+Qh}{\|w^*\|_2} + \sum_{j=2}^p U_j a_j\right]^2 \\
&= \frac{2C\bar{\epsilon}\text{vol}(\mathcal{D}_u^+)}{\|w^*\|_2}\left\{\left[\alpha_0 + a_1\frac{1-b^*}{\|w^*\|_2} + \mathop{\mathbb{E}}_{Q,U}\left(a_1\frac{Qh}{\|w^*\|_2} + \sum_{j=2}^p U_j a_j\right)\right]^2 + \text{Var}\left[\sum_{j=2}^p U_j a_j + \frac{a_1 Qh}{\|w^*\|_2}\right]\right\},
\end{aligned}$$

where the equality follows since $\mathbb{E}(X^2) = \mathbb{E}(X)^2 + \text{Var}(X)$.

Letting $m_j = \frac{l_j+v_j}{2}$, for $j = 2, \ldots p$, we get $\mathbb{E}[U_j] = m_j$ and

$$\begin{aligned}
\text{Var}\left(\sum_{j=2}^p a_j U_j + a_1\frac{Qh}{\|w^*\|_2}\right) &\geq \min_{2\leq j\leq p}\text{Var}(U_j)\sum_{j=2}^p a_j^2 + \frac{a_1^2h^2}{\|w^*\|_2}\text{Var}(Q) \\
&= \min_{2\leq j\leq p}\text{Var}(U_j)\sum_{j=2}^p a_j^2 + \frac{\bar{\epsilon}^2 a_1^2 h^2}{3\|w^*\|_2}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\bar{\pi}_+\int_{\mathcal{X}} \frac{1}{h} K\left(\frac{1-\mathring{x}^T\theta^*}{h}\right)\langle\mathring{\alpha},\mathring{x}\rangle^2 f(x)dx \\
&\quad\geq \frac{C\bar{\pi}_+\text{vol}(\mathcal{D}_u^+)2\bar{\epsilon}}{\|w^*\|_2}\left\{\left[\alpha_0 + a_1\frac{1-b^*}{\|w^*\|_2} + \sum_{j=2}^p a_j m_j\right]^2 + \min_{2\leq j\leq p}\text{Var}(U_j)\sum_{j=2}^p a_j^2 + \frac{\bar{\epsilon}^2 a_1^2 h^2}{3\|w^*\|_2}\right\}
\end{aligned}$$

$$\begin{aligned}
&\bar{\pi}_-\int_{\mathcal{X}} \frac{1}{h} K\left(\frac{1+\mathring{x}^T\theta^*}{h}\right)\langle\mathring{\alpha},\mathring{x}\rangle^2 f(x)dx \\
&\quad\geq \frac{C\bar{\pi}_-\text{vol}(\mathcal{D}_u^-)2\bar{\epsilon}}{\|w^*\|_2}\left\{\left[\alpha_0 - a_1\frac{1+b^*}{\|w^*\|_2} + \sum_{j=2}^p a_j m_j\right]^2 + \min_{2\leq j\leq p}\text{Var}(U_j)\sum_{j=2}^p a_j^2 + \frac{\bar{\epsilon}^2 a_1^2 h^2}{3\|w^*\|_2}\right\}.
\end{aligned}$$

Then,

$$\begin{aligned}\mathring{\alpha}^T\nabla^2 L_h(\theta^*)\mathring{\alpha} &\geq C\bigg(\alpha_0 - a_1\frac{1+b^*}{\|w^*\|_2} + \sum_{j=2}^{p} a_j m_j\bigg)^2 + C\bigg(\alpha_0 + a_1\frac{1-b^*}{\|w^*\|_2} + \sum_{j=2}^{p} a_j m_j\bigg)^2 \\ &\quad + 2C\min_{2\leq j\leq p}\mathrm{Var}(U_j)\sum_{j=2}^{p} a_j^2 + 2C\frac{\bar{\epsilon}^2 a_1^2 h^2}{3\|w^*\|_2} \\ &\geq C\bigg(\alpha_0 - a_1\frac{1+b^*}{\|w^*\|_2} + \sum_{j=2}^{p} a_j m_j\bigg)^2 + C\bigg(\alpha_0 + a_1\frac{1-b^*}{\|w^*\|_2} + \sum_{j=2}^{p} a_j m_j\bigg)^2 \\ &\quad + 2C\sum_{j=2}^{p} a_j^2 + 2C\frac{\bar{\epsilon}^2 a_1^2 h^2}{3\|w^*\|_2}.\end{aligned}$$

Following the argument in Lemma 5 of Koo et al. (2008), the first three terms represent a positive-definite quadratic form $\mathcal{Q} = \mathcal{Q}(b^*, a_1, \ldots, a_p)$. Let $\upsilon_1 > 0$ be the smallest eigenvalue of the matrix corresponding to $\mathcal{Q}$. Then,

$$\mathring{\alpha}^T\nabla^2 L_h(\theta^*)\mathring{\alpha} \geq C\bigg(\upsilon_1\|\mathring{\alpha}\|_2^2 + C' a_1^2 h^2\bigg) = C\bigg(\upsilon_1\|\mathring{\alpha}\|_2^2 + C'(A_1\alpha)_1^2 h^2\bigg),$$

where

$$C' \triangleq 2\frac{\bar{\epsilon}^2}{3\|w^*\|_2},$$

from which the result follows.

Now consider the case of $h \geq \frac{\psi\|w^*\|_2}{\epsilon}$. For any such h, $\bar{\epsilon} = \frac{\psi\|w^*\|_2}{h}$. Following an analogous derivation to the one above, we obtain,

$$\mathring{\alpha}^T\nabla^2\,\mathbb{E}\left[L_h(\theta^*)\right]\mathring{\alpha} \geq C\bigg(\frac{\upsilon_1\|\mathring{\alpha}\|_2^2}{h^2} + C'(A_1\alpha)_1^2\bigg).$$

■

B.3.3 Proof of Proposition 14

Let $\Delta \in \mathbb{R}^{p+1}$ and assume that $s = 1$. We write

$$\Delta \triangleq (\Delta_b, \Delta_w), \quad \Delta_b \in \mathbb{R},\ \Delta_w \in \mathbb{R}^p,$$

in line with our notation $\theta^* = (b^*, w^*)$. For any $\mathring{\alpha} \in \mathbb{R}^{p+1}$, $\mathring{\alpha} \triangleq (\alpha_0, \alpha)$,

$$\mathring{\alpha}^T\nabla^2\,\mathbb{E}[\hat{L}_h(\theta^* + \Delta)]\mathring{\alpha} = \int_{\mathcal{X}}\frac{1}{h}K\bigg(\frac{1 - \mathring{x}^T\theta^* - \mathring{x}^T\Delta}{h}\bigg)\langle\mathring{\alpha}, \mathring{x}\rangle^2 f(x)dx.$$

Since Δ appears only inside the kernel, we can use derivation analogous to those in the proof of Proposition 13, as long as $K\big((1 - \mathring{x}^T\theta^* - \mathring{x}^T\Delta)/h\big)$ can be lower-bounded by a constant.

To derive the lower bound in Proposition 13, Assumptions 7 and 8 were imposed, which require a non-zero conditional density of X and non-zero kernel density over a subset of the (true) margin. This region can be described by the following three conditions.

Condition 22 (Assumption 8) *Let* $z \triangleq \frac{A_1 x}{\|w^*\|_2}$. *For the first coordinate,*

$$(A_1 x)_1 \in \frac{1 - b^*}{\|w^*\|_2} + \mathbb{B}(\psi).$$

Let $z_1^* \triangleq (1 - b^*)/\|w^*\|_2^2$. *Then,* $z_1 \in z_1^* + \mathbb{B}\big(\psi/\|w^*\|_2\big)$.

Condition 23 (Assumption 8) *For* $i \in \{2, \dots p\}$: $l_i \leq (A_1 x)_i \leq v_i$, *and hence*

$$\frac{l_i}{\|w^*\|_2} \leq z_i \leq \frac{v_i}{\|w^*\|_2}.$$

To simplify the notation, let $z^* \in \mathbb{R}^p$, $z_i^* = \frac{l_i + v_i}{2\|w^*\|_2}$, *i.e. the center of the subset of the margin over which density is assumed to be positive. Let* $R_z \triangleq \min_{i \in \{2,\dots,p\}} \{(v_i - l_i)/2\|w^*\|_2\}$. *Then* $z^* + \mathbb{B}(R_z) \subset \mathcal{D}_*^+$ *and conditional density of* X *is thus strictly positive over this ball.*

Condition 24 (Assumption 7) *The kernel density is bounded from below on* $\mathbb{B}(\epsilon)$, *i.e.* $K(x) > C$ *for* $x \in \mathbb{B}(\epsilon)$.

Since $z = \frac{A_1 x}{\|w^*\|_2}$,

$$\begin{aligned}\frac{1 - \mathring{x}^T\theta^* - \mathring{x}^T\Delta}{h} &= \frac{1 - \|w^*\|_2\langle w^*, A_1^T z\rangle - b^* - \|w^*\|_2\langle \Delta_w, A_1^T z\rangle - \Delta_b}{h} \\ &= q - \frac{\|w^*\|_2\langle A_1\Delta_w, z\rangle + \Delta_b}{h},\end{aligned}$$

where $q = \frac{1 - b^* - \|w^*\|_2^2 z_1}{h}$. Note that $A_1\Delta_w$ is a transformation of the error Δ_w into the coordinate system z. We can separate error in the direction of w^* and direction orthogonal to θ^* by considering $(A_1\Delta_w)_1$ and $(A_1\Delta_w)_{-1}$ separately. The goal is to understand for which Δ is Condition 24 satisfied subject to conditions 22 and 23 above. This is equivalent to saying that the estimation error Δ must small enough for the margin based on $\hat{\theta}$ to intersect the region within the support of conditional densities and the kernel.

Hence we must have,

$$\left|\frac{\Delta_b + \|w^*\|_2\langle A_1\Delta_w, z\rangle}{h}\right| \leq \epsilon.$$

Then,

$$\begin{aligned}\left|\frac{\|w^*\|_2\langle (A_1\Delta_w), z\rangle}{h}\right| &= \frac{1}{h}\|w^*\|_2 \left|\big\langle A_1\Delta_w, \big(z_1 - z_1^* + z_1^*, z_{-1}^* + (z_{-1} - z_{-1}^*)\big)\big\rangle\right| \\ &\leq \frac{1}{h}\|w^*\|_2 \|A_1\Delta_w\|_2 \left(|z_1^*| + |z_1 - z_1^*| + \big\|z_{-1}^*\big\|_2 + \big\|z_{-1} - z_{-1}^*\big\|_2\right) \\ &\leq \frac{1}{h}\|w^*\|_2 \|\Delta_w\|_2 \left(|z_1^*| + \big\|z_{-1}^*\big\|_2 + R_z + \frac{\psi}{\|w^*\|_2}\right),\end{aligned}$$

where z_{-1} denotes vector $(z_2, \dots, z_p)$, the second inequality follows from the Cauchy-Schwartz inequality and the last inequality uses the definition of z_{-1}^* and orthogonality of A_1. Overall, Condition 24 holds for any Δ satisfying

$$|\Delta_b| + \|w^*\|_2 \|\Delta_w\|_2 \left(|z_1^*| + \big\|z_{-1}^*\big\|_2 + R_z + \frac{\psi}{\|w^*\|_2}\right) \leq h\epsilon.$$

Thus, for any $\Delta \in \mathbb{B}(hR_1)$

$$R_1 \triangleq \frac{\epsilon}{1 + \left(\|w^*\|_2 \left\|z^*_{-1}\right\|_2 + \|w^*\|_2 R_z + \psi + |1 - b^*|/\|w^*\|_2\right)}, \tag{51}$$

the result analogous to that of Proposition 13 can be obtained using the same argument. ■

B.3.4 Proof of Proposition 15

Before proving Proposition 15, recall that a function $\varphi : \mathbb{R} \to \mathbb{R}$ is called a contraction if, for all $s, t \in \mathbb{R}$, $|\varphi(s) - \varphi(t)| \leq |s - t|$. If, in addition, $\varphi(0) = 0$, we say that φ is a centred contraction.

The proof of Proposition 15 uses the following contraction inequality for Rademacher complexity.

Lemma 25 (Ledoux and Talagrand, 2013) *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be convex and increasing, and let $\varphi_i : \mathbb{R} \to \mathbb{R}$, $i \leq n$ be centred contractions. Then, for any bounded subset $T \subset \mathbb{R}^n$,*

$$\mathbb{E}\left[f\left(\sup_{t\in T}\left|\sum_{i=1}^n \varepsilon_i \varphi_i(t_i)\right|\right)\right] \leq \mathbb{E}\left[f\left(2\sup_{t\in T}\left|\sum_{i=1}^n \varepsilon_i t_i\right|\right)\right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher variables.

We first re-write $\mathcal{E}(\Delta)$ as $\mathcal{E}(\Delta) = (1/n)\sum_{i=1}^n \mathcal{E}_i(\Delta)$, where

$$\mathcal{E}_i(\Delta) = l_h(1 - Y_i \mathring{X}_i^T(\theta^* + \Delta)) - l_h(1 - Y_i \mathring{X}_i^T \theta^*) + l'_h(1 - Y_i \mathring{X}_i^T \theta^*) Y_i \mathring{X}_i^T \Delta$$

satisfies $\mathcal{E}_i(0) = 0$. Lemma 12 implies that $\mathcal{E}_i(\Delta)$ is 1-Lipschitz in $\mathring{x}_i^T \Delta$. Define the random quantity

$$A = \frac{1}{4r} \sup_{\Delta \in \mathbb{B}(r)} |\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)|.$$

We aim to control the probability of $A \geq \delta$ for a given $\delta > 0$ and do so by controlling its moment generating function. Let $\varepsilon_1, \ldots \varepsilon_n$ be independent Rademacher variables. Using the Rademacher symmetrisation (see, e.g. Proposition 4.11 in Wainwright, 2019) and convexity of the exponential function, we can work with a symmetrised version of A instead. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}[\exp(\lambda A)] &\leq \mathbb{E}\left[\exp\left(\frac{\lambda}{2r} \sup_{\Delta \in \mathbb{B}(r)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathcal{E}_i(\Delta)\right|\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{\lambda}{r} \sup_{\Delta \in \mathbb{B}(r)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \Delta, \mathring{X}_i\rangle\right|\right)\right],
\end{aligned}$$

where the second inequality follows from Lemma 25 combined with the Lipschitz continuity of $\mathcal{E}_i(\Delta)$. By Hölder's inequality we see that for any $\Delta \in \mathbb{B}(r)$,

$$\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \Delta, \mathring{X}_i\rangle\right| = \left|\Delta^T\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathring{X}_i\right)\right| \leq \|\Delta\|_2 \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathring{X}_i\right\|_2 \leq r\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathring{X}_i\right\|_2.$$

Hence,

$$\mathbb{E}\left[e^{\lambda(A-\delta)}\right] \leq \mathbb{E}\left[\exp\left(\lambda\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathring{X}_i\right\|_2 - \lambda\delta\right)\right], \quad \lambda \in \mathbb{R}.$$

By Markov's inequality, it holds for any $\delta > 0$ that

$$\mathbb{P}\big[A \geq \delta\big] \leq \inf_{\lambda>0}\mathbb{E}\left[e^{\lambda(A-\delta)}\right] \leq \inf_{\lambda>0}\mathbb{E}\left[\exp\left(\lambda\left\|\frac{1}{n}\sum_{i=1}^{n}\mathring{X}_i\right\|_2 - \lambda\delta\right)\right].$$

This establishes inequality (28). Now we turn to proving the uniform bound that holds for all $\Delta \in \Theta\big(r_l, r_u\big)$ for $0 < r_l < r_u$. For $m \in \mathbb{N}$, define the sets

$$\mathcal{S}_m \triangleq \big\{\Delta \in \mathbb{R}^p : \gamma^{m-1}r_l \leq \|\Delta\|_2 \leq \gamma^m r_l\big\},$$

for some $\gamma > 1$. Then $\Theta(r_l, r_u) \subseteq \cup_{i=1}^{i=N}\mathcal{S}_i$, where $N = \lceil \frac{\log(r_u/r_l)}{\log\gamma} \rceil$. It follows that

$$\begin{aligned}
&\mathbb{P}\big\{\exists\Delta \in \Theta(r_l, r_u) : |\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| > 4\|\Delta\|_2\gamma\delta\big\} \\
&\quad\leq \mathbb{P}\big\{\exists\Delta \in \cup_{i=1}^{i=N}\mathcal{S}_i : |\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| > 4\|\Delta\|_2\gamma\delta\big\} \\
&\quad\leq \sum_{i=1}^{N}\mathbb{P}\big\{\exists\Delta \in \mathcal{S}_i : |\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| > 4\|\Delta\|_2\gamma\delta\big\} \\
&\quad\leq \sum_{i=1}^{N}\mathbb{P}\big\{\exists\Delta \in \mathcal{S}_i : |\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| > 4\gamma^{i-1}\gamma r_l\delta\big\} \\
&\quad\leq \sum_{i=1}^{N}\mathbb{P}\left\{\sup_{\Delta\in\mathbb{B}(\gamma^i r_l)}|\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| > 4\gamma^i r_l\delta\right\} \\
&\quad\leq \sum_{i=1}^{N}\inf_{\lambda>0}\mathbb{E}\left[e^{\lambda(\|\bar{X}_n\|_2-\delta)}\right] = \left\lceil\frac{\log(r_u/r_l)}{\log\gamma}\right\rceil\inf_{\lambda>0}\mathbb{E}\left[e^{\lambda(\|\bar{X}_n\|_2-\delta)}\right],
\end{aligned}$$

where the second inequality follows from the union bound and the third inequality uses equation (28). Setting $\gamma = e^{1/2}$, we obtain inequality (29). ■

B.3.5 Proof of Proposition 16

Note that,

$$\mathbb{E}\left[\exp(\lambda\|\bar{X}_n\|_2)\right] = \mathbb{E}\left[\exp(\lambda\sup_{v\in\mathbb{S}^p}v^T\bar{X}_n)\right] = \mathbb{E}\left[\max_{v\in\mathbb{S}^p}\exp(\lambda v^T\bar{X}_n)\right].$$

By Assumption 3, $(X_i)_{i=1}^n$ are independent sub-Gaussian random vectors, thus $(v^TX_i)_{i=1}^n$ are independent sub-Gaussian random variables and $v^T\bar{X}_n$, being an average of such, satisfies for $v \in \mathbb{S}^p$,

$$\mathbb{E}\left[\exp(\lambda v^T\bar{X}_n)\right] = \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{1}{n}v^T\mathring{X}_i\right)\right] \leq \exp\left(1 + \frac{\lambda^2\nu_1^2}{2n}\right).$$

For any $\gamma \in (0,1)$ there exists a γ-net N_γ of the unit sphere, such that

$$|N_\gamma| \leq \left(\frac{2}{\gamma}+1\right)^{p+1}.$$

Let $\mathcal{N}$ be a $\gamma-$net of a unit ball $\mathcal{B} \triangleq \mathbb{B}^{p+1}(1)$. Then for any $u \in \mathbb{B}^{p+1}(1)$, there exists $v \in \mathcal{N}$ and $w \in \mathbb{R}^{p+1}$, such that $\|w\|_2 \leq \gamma$ and $u = v + w$. Then, for any sub-Gaussian random variable X with variance proxy ν_1,

$$\max_{u\in\mathcal{B}} u^T X \leq \max_{v\in\mathcal{N}} v^T X + \max_{w\in\gamma\mathcal{B}} w^T X.$$

As a result,

$$\max_{u\in\mathcal{B}} u^T X \leq \frac{1}{1-\gamma}\max_{v\in\mathcal{N}} v^T X.$$

For any $\lambda > 0$ and taking $X = \bar{X}_n$ we thus have

$$\begin{aligned}\mathbb{E}[\exp(\max_{u\in\mathcal{B}} \lambda u^T \bar{X}_n)] &\leq \mathbb{E}\left[\exp\left(\frac{1}{1-\gamma}\max_{v\in\mathcal{N}} \lambda v^T \bar{X}_n\right)\right] \leq \mathbb{E}\sum_{i=1}^{|\mathcal{N}|}\left[\exp\left(\frac{\lambda}{1-\gamma} v_i^T \bar{X}_n\right)\right] \\ &\leq |\mathcal{N}|\exp\left(\frac{\lambda^2\nu_1^2}{2(1-\gamma)^2 n}+1\right) \leq \left(\frac{2}{\delta'}+1\right)^{p+1}\exp\left(\frac{\lambda^2\nu_1^2}{2(1-\gamma)^2 n}+1\right).\end{aligned}$$

Let $\gamma = \frac{2}{e-1}$ to obtain

$$\inf_{\lambda>0}\mathbb{E}[\exp(\max_{u\in\mathcal{B}} \lambda u^T \bar{X}_n) - \lambda\delta] \leq \inf_{\lambda>0}\exp\left(20\frac{\lambda^2\nu_1^2}{n} + p + 2 - \lambda\delta\right).$$

The expression is minimised by $\lambda = \frac{\delta n}{40\nu_1^2}$. Hence, for any $\delta > 0$,

$$\inf_{\lambda>0}\mathbb{E}[\exp(\max_{u\in\mathcal{B}} \lambda u^T \bar{X}_n) - \lambda\delta] \leq \exp\left(p + 2 - \frac{\delta^2 n}{80\nu_1^2}\right).$$

Take $\delta = \sqrt{\frac{80\nu_1^2}{n}(p+2+\tau)}$, $\tau > 0$. Then, by Proposition 15, for any given $r > 0$ and any $\tau > 0$, using equation (28),

$$\sup_{\Delta\in\mathbb{B}(r)} |\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| \leq Cr\sqrt{\frac{p+2+\tau}{n}}$$

with probability at least $1 - e^{-\tau}$. Moreover, from inequality (29), we have, for any given r_l, r_u such that $0 < r_l < r_u$, for any $\tau > 0$,

$$|\mathcal{E}(\Delta) - \bar{\mathcal{E}}(\Delta)| \leq C\|\Delta\|_2\sqrt{\frac{p+2+\tau}{n}} \quad \text{for all } \Delta \in \mathbb{B}(r_l, r_u) \tag{52}$$

with probability at least $1 - 2\lceil\log(r_u/r_l)\rceil e^{-\tau}$. ■

B.3.6 Proof of Proposition 17

By Proposition 14 for any $\mathring{\alpha} \in \mathbb{R}^{p+1}$, $\mathring{\alpha} \triangleq (\alpha_0, \alpha)$, and for $h \leq \max\left\{\frac{\psi\|w^*\|_2}{\epsilon}, 1\right\}$,

$$\mathring{\alpha}^T \nabla^2 \mathbb{E}\left[\hat{L}_h(\theta^*)\right] \mathring{\alpha} \geq C\bigg(\|\mathring{\alpha}\|_2^2 + C'(A_1\alpha)_1^2 h^2\bigg),$$

for any $\Delta \in \mathbb{B}(hR_1)$. Hence, for any such Δ,

$$\bar{\mathcal{E}}(\Delta) \triangleq \mathbb{E}[\hat{L}_h(\theta^* + \Delta)] - \mathbb{E}[\hat{L}_h(\theta^*)] - \langle \nabla\, \mathbb{E}[\hat{L}_h(\theta^*)], \Delta\rangle \geq \|\Delta\|_2^2.$$

For $r \in (0, hR_1)$, we thus have, by Proposition 16, with probability at least $1 - e^{-t}$,

$$\mathcal{E}(\Delta) \geq C\|\Delta\|_2^2 - C' r\sqrt{\frac{p+t}{n}}, \tag{53}$$

for all $\Delta \in \mathbb{B}(r)$. Similarly, for $0 < r_l < r_u < hR_1$, with probability at least $1 - 2\lceil \log(r_u/r_l)\rceil e^{-t}$, for all $\Delta \in \mathbb{B}(r_l, r_u)$

$$\mathcal{E}(\Delta) \geq C\|\Delta\|_2^2 - C'\|\Delta\|_2\sqrt{\frac{p+t}{n}}. \tag{54}$$

Note that the equations (53) and (54) can be respectively written as

$$\langle \nabla L_h(\theta^* + \Delta) - \nabla L_h(\theta^*), \Delta\rangle \geq C\|\Delta\|_2^2 - C' r\sqrt{\frac{p+t}{n}}, \text{ for all } \Delta \in \mathbb{B}(r)$$
$$\langle \nabla L_h(\theta^* + \Delta) - \nabla L_h(\theta^*), \Delta\rangle \geq C\|\Delta\|_2^2 - C'\|\Delta\|_2\sqrt{\frac{p+t}{n}}, \text{ for all } \Delta \in \mathbb{B}(r_l, r_u).$$

■

Appendix C. Proofs of Auxiliary Results

C.1 Proof of Lemma 18 (Smoothing Bias)

Let $\mathring{v} \in \mathbb{S}^p$. We use the notation $\mathring{v} = (v_0, v)$, where $v_0 \in \mathbb{R}$ and $v \in \mathbb{R}^p$. Then,

$$\begin{aligned} v^T\, \mathbb{E}\left[\nabla \widehat{L}_h(\theta^*)\right] &= \mathbb{E}\left[Y \mathring{v}^T \mathring{X} \bar{K}\bigg(\frac{1 - Y\mathring{X}^T\theta^*}{h}\bigg)\right] \\ &= \bar{\pi}_+ \int_{\mathcal{X}} \mathring{v}^T \mathring{x} \bar{K}\bigg(\frac{1 - \mathring{x}^T\theta^*}{h}\bigg) f(x)dx - \bar{\pi}_- \int_{\mathcal{X}} \mathring{v}^T \mathring{x} \bar{K}\bigg(\frac{1 + \mathring{x}^T\theta^*}{h}\bigg) g(x)dx. \end{aligned}$$

We focus on the conditional expectation given $Y = 1$, i.e.

$$\begin{aligned} &\bar{\pi}_+ \int_{\mathcal{X}} \mathring{v}^T \mathring{x} \bar{K}\bigg(\frac{1 - \mathring{x}^T\theta^*}{h}\bigg) f(x)dx \\ &\quad = \bar{\pi}_+ \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} \mathring{v}^T \mathring{x} \bar{K}\bigg(\frac{1 - \mathring{x}^T\theta^*}{h}\bigg) f(x_s|x_{-s}) f(x_{-s}) dx_s dx_{-s}. \end{aligned}$$

Analogous derivations hold for $Y=-1$.

Since under Assumption 2, $\theta^*_s \neq 0$ for some $s \in \{1,\dots,p\}$, let $u = \frac{1-\mathring{x}^T\theta^*}{h}$, so $x_s = \frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}$. By a change of variables,

$$\begin{aligned}
&\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}} \mathring{v}^T\mathring{x}\bar{K}\bigg(\frac{1-\mathring{x}^T\theta^*}{h}\bigg)f(x_s|x_{-s})f(x_{-s})dx_s dx_{-s}\\
&= -\frac{h}{\theta^*_s}\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}}\bigg(\frac{v_s(1-\mathring{x}^T_{-s}\theta^*_{-s}-hu)}{\theta^*_s}\bigg)\bar{K}(u)f\bigg(\frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}\bigg|x_{-s}\bigg)f(x_{-s})dudx_{-s}\\
&\quad -\frac{h}{\theta^*_s}\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}}\bigg(v_0+v^T_{-s}x_{-s}\bigg)\bar{K}(u)f\bigg(\frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}\bigg|x_{-s}\bigg)f(x_{-s})dudx_{-s}
\end{aligned}$$

Let

$$I_1 \triangleq -\frac{h}{\theta^*_s}\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}} v_s\frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}\bar{K}(u)f\bigg(\frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}\bigg|x_{-s}\bigg)f(x_{-s})dudx_{-s} \quad (55)$$

$$I_{-1} \triangleq -\frac{h}{\theta^*_s}\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}} \mathring{v}^T_{-s}\mathring{x}_{-s}\bar{K}(u)f\bigg(\frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}\bigg|x_{-s}\bigg)f(x_{-s})dudx_{-s}. \quad (56)$$

Lemma 26 *Under Assumptions 2, 4 and 5,*

$$\begin{aligned}
&-\frac{h}{\theta^*_s}\int_{\mathbb{R}^{p-s}}\int_{\mathbb{R}} v_s\frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}\bar{K}(u)f\bigg(\frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}\bigg|x_{-s}\bigg)f(x_{-s})dudx_{-s}\\
&\qquad = -v_s\mathbb{E}_{X|Y=1}\Big[Y\mathring{X}_s\mathbb{1}\Big\{1-Y\mathring{X}^T\theta^*\geq 0\Big\}\Big]+\tilde{I}_1,
\end{aligned}$$

where $|\tilde{I}_1| \leq \kappa_2|v_s|C\frac{h^2}{2\theta^{*2}_s}$.

Lemma 27 *Under Assumptions 2, 4 and 5,*

$$\begin{aligned}
&-\frac{h}{\theta^*_s}\int_{\mathbb{R}} \mathring{v}^T_{-s}\mathring{x}_{-s}\bar{K}(u)f\bigg(\frac{1-\mathring{x}^T_{-s}\theta^*_{-s}-hu}{\theta^*_s}\bigg|x_{-s}\bigg)du\\
&\qquad = -\mathring{v}^T_{-s}\mathbb{E}_{X|Y=1}\Big[Y\mathring{X}_{-s}\mathbb{1}\Big\{1-Y\mathring{X}^T\theta^*\geq 0\Big\}\Big]+\tilde{I}_{-1},
\end{aligned}$$

where $|\tilde{I}_{-1}| \leq C\frac{h^2\kappa_2}{2\theta^{*2}_s}\mathring{\mu}_1$.

By Lemma 26,

$$I_1 = -v_s\mathbb{E}_{X|Y=1}\Big[Y\mathring{X}_s\mathbb{1}\Big\{1-Y\mathring{X}^T\theta^*\geq 0\Big\}\Big]+\tilde{I}_1 \quad (57)$$

and by Lemma 27

$$I_{-1} = -\mathring{v}^T_{-s}\mathbb{E}_{X|Y=1}\Big[Y\mathring{X}_{-s}\mathbb{1}\Big\{1-Y\mathring{X}^T\theta^*\geq 0\Big\}\Big]+\tilde{I}_{-1}. \quad (58)$$

Thus,

$$I_1+I_{-1} = -\mathring{v}^T\mathbb{E}_{X|Y=1}\Big[Y\mathring{X}\mathbb{1}\Big\{1-Y\mathring{X}^T\theta^*\geq 0\Big\}\Big]+\tilde{I}_1+\tilde{I}_{-1}.$$

Using the triangle inequality,

$$\left|\int_{\mathbb{R}^p} \mathring{v}^T \mathring{x} \bar{K}\left(\frac{1-\mathring{x}^T\theta^*}{h}\right) f(x)dx + \mathring{v}^T \mathbb{E}_{X|Y=1}\left[Y\mathring{X}\,\mathbb{1}\left\{1 - Y\mathring{X}^T\theta^* \geq 0\right\}\right]\right| \leq |\tilde{I}_1| + |\tilde{I}_{-1}|. \tag{59}$$

Thus, using the bounds for $\tilde{I}_1$ and $\tilde{I}_{-1}$ in Lemmas 26 and 27,

$$\left|\int_{\mathbb{R}^p} \mathring{v}^T \mathring{x} \bar{K}\left(\frac{1-\mathring{x}^T\theta^*}{h}\right) f(x)dx + \mathring{v}^T \mathbb{E}_{X|Y=1}\left[Y\mathring{X}\,\mathbb{1}\left\{1 - Y\mathring{X}^T\theta^* \geq 0\right\}\right]\right| \leq \frac{Ch^2\kappa_2}{2\theta_s^{*2}}(1+\mathring{\mu}_1).$$

By analogous derivations for $Y=-1$ and since $\mathbb{E}\left[Y\mathring{X}\,\mathbb{1}\left\{1 - Y\mathring{X}^T\theta^* \geq 0\right\}\right] = 0$, we obtain, for any $\mathring{v} \in \mathbb{S}^p$,

$$\left|\mathring{v}^T\,\mathbb{E}\left[\nabla \widehat{L}_h(\theta^*)\right]\right| \leq \frac{Ch^2\kappa_2}{2\theta_s^{*2}}(1+\mathring{\mu}_1).$$

■

C.1.1 Proof of Lemma 26

Write

$$I_1 \triangleq -\frac{h}{\theta_s^*}\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}} v_s \frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\bar{K}(u) f\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\middle| x_{-s}\right) f(x_{-s})dudx_{-s}. \tag{60}$$

Let $G(t) = \int_{-\infty}^t x_s f(x_s|x_{-s})dx_s$. By the Leibnitz rule,

$$dG(t) = tf(t|x_{-s}) \tag{61}$$

and

$$\begin{aligned} dG\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\right) &\triangleq \frac{\partial}{\partial u} G\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\right) \\ &= -\frac{h}{\theta_s^*}\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\right) f\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\middle| x_{-s}\right). \end{aligned}$$

On substituting into the expression for I_1,

$$\begin{aligned} &-\frac{h}{\theta^*}\int_{\mathbb{R}}\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\right)\bar{K}(u) f\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\middle| x_{-s}\right)du \\ &\quad = \int_{\mathbb{R}} \bar{K}(u) dG\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\right)du \\ &\quad = -\int_{\mathbb{R}} K(u) G\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\right)du, \end{aligned}$$

where the last equality was obtained using integration by parts and from Assumptions 3 and 5. By a first-order Taylor expansion around $(1-\mathring{x}_{-s}^T\theta_{-s}^*)/\theta_s^*$,

$$G\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^* - hu}{\theta_s^*}\right) = G\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}\right) + dG\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}\right)\left(\frac{-hu}{\theta_s^*}\right) + \mathcal{R}, \tag{62}$$

where the remainder, $\mathcal{R}$, has the following form

$$\mathcal{R} = \int_0^1 \left[dG\Big(\frac{1 - \theta^*_{-s} \mathring{x}_{-s}}{\theta^*_s} + t \frac{-hu}{\theta^*_s} \Big) - dG\Big(\frac{1 - \theta^*_{-s} \mathring{x}_{-s}}{\theta^*_s} \Big) \right] \Big(- \frac{hu}{\theta^*_s} \Big) dt.$$

By a change of variables ($w \triangleq thu$),

$$\mathcal{R} = -\frac{1}{\theta^*_s} \int_0^{hu} \left[dG\Big(\frac{1 - \theta^*_{-s} \mathring{x}_{-s}}{\theta^*_s} - \frac{w}{\theta^*_s} \Big) - dG\Big(\frac{1 - \theta^*_{-s} \mathring{x}_{-s}}{\theta^*_s} \Big) \right] dw.$$

Under Assumption 5, for any $a, b \in \mathbb{R}$, $|af(a|x_{-s}) - bf(b|x_{-s})| \leq C|a-b|$. Hence,

$$\begin{aligned} |\mathcal{R}| &\leq \frac{1}{\theta^*_s} \int_0^{hu} \left| dG\Big(\frac{1 - \theta^*_{-s} \mathring{x}_{-s}}{\theta^*_s} - \frac{w}{\theta^*_s} \Big) - dG\Big(\frac{1 - \theta^*_{-s} \mathring{x}_{-s}}{\theta^*_s} \Big) \right| dw \\ &\leq C \int_0^{hu} \frac{|w|}{\theta^{*2}_s} dw \leq C \frac{(hu)^2}{2\theta^{*2}_s}. \end{aligned} \tag{63}$$

Using (62),

$$\begin{aligned} I_1 &= -v_s \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} G\Big(\frac{1 - \mathring{x}^T_{-s} \theta^*_{-s}}{\theta^*_s} \Big) K(u) f(x_{-s}) du dx_{-s} \\ &\quad + v_s \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} dG\Big(\frac{1 - \mathring{x}^T_{-s} \theta^*_{-s}}{\theta^*_s} \Big) \frac{hu}{\theta^*_s} K(u) f(x_{-s}) du dx_{-s} \\ &\quad - v_s \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} \mathcal{R} K(u) f(x_{-s}) du dx_{-s} \\ &\triangleq I_1^A + I_1^B + I_1^C, \end{aligned}$$

where integrals I_1^A, I_1^B, I_1^C correspond to the respective terms in the expression.

We can further simplify I_1^A by noting that,

$$\int_{\mathbb{R}} G\Big(\frac{1 - \mathring{x}^T_{-s} \theta^*_{-s}}{\theta^*_s} \Big) K(u) du = G\Big(\frac{1 - \mathring{x}^T_{-s} \theta^*_{-s}}{\theta^*_s} \Big)$$

and

$$\begin{aligned} &\int_{\mathbb{R}^{p-1}} G\Big(\frac{1 - \mathring{x}^T_{-s} \theta^*_{-s} - hu}{\theta^*_s} \Big) f(x_{-s}) dx_{-s} \\ &\qquad = \int_{\mathbb{R}^{p-1}} \int_{-\infty}^{\frac{1 - \mathring{x}^T_{-s} \theta^*_{-s}}{\theta^*_s}} x_s f(x_s|x_{-s}) f(x_{-s}) dx_s dx_{-s} \\ &\qquad = \int_{\mathbb{R}^p} \mathbb{1}\left\{ \frac{1 - \mathring{x}^T_{-s} \theta^*_{-s}}{\theta^*_s} \geq x_s \right\} x_s f(x_s|x_{-s}) f(x_{-s}) dx \\ &\qquad = \mathbb{E}_{X|Y=1} \left[Y X_s \mathbb{1}\left\{ 1 - Y \mathring{X}^T \theta^* \geq 0 \right\} \right]. \end{aligned}$$

Hence,

$$I_1^A = -v_s \mathbb{E}_{X|Y=1} \left[Y \mathring{X}_s \mathbb{1}\left\{ 1 - Y \mathring{X}^T \theta^* \geq 0 \right\} \right]. \tag{64}$$

Using the expression (61) for dG, we can write I_1^B as,

$$I_1^B = v_s \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} \frac{hu}{\theta_s^*} \left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} \right) f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} \middle| x_{-s} \right) K(u) f(x_{-s}) du dx_{-s}. \tag{65}$$

By the symmetry of kernel density $\int_{\mathbb{R}} uK(u)du = 0$. Hence $I_1^B = 0$.

Finally, by Assumption 4 and equation (63),

$$\begin{aligned} |I_1^C| &\leq \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} |v_s \mathcal{R} K(u) f(x_{-s})| \, du dx_{-s} \\ &\leq |v_s| C \frac{h^2}{2\theta_s^{*2}} \int_{\mathbb{R}^{p-1}} \left[\int_{\mathbb{R}} u^2 K(u) du \right] f(x_{-s}) dx_{-s} \\ &\leq \kappa_2 C \frac{h^2}{2\theta_s^{*2}}. \end{aligned} \tag{66}$$

■

C.1.2 Proof of Lemma 27

On integrating by parts,

$$\begin{aligned} &-\frac{h}{\theta_s^*} \int_{\mathbb{R}} v_{-s}^T \mathring{x}_{-s} \bar{K}(u) f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^* - hu}{\theta_s^*} \middle| x_{-s} \right) du \\ &\quad = -\int_{\mathbb{R}} v_{-s}^T \mathring{x}_{-s} K(u) F\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^* - hu}{\theta_s^*} \middle| x_{-s} \right) du. \end{aligned}$$

By a Taylor expansion of $F\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^* - hu}{\theta_s^*} \middle| x_{-s} \right)$ around $\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*}$ (i.e. around $u = 0$),

$$\begin{aligned} F\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^* - hu}{\theta_s^*} \middle| x_{-s} \right) &= F\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} \middle| x_{-s} \right) + f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} \middle| x_{-s} \right) \left(\frac{-uh}{\theta_s^*} \right) \\ &+ \int_0^1 \left[f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} - t \frac{uh}{\theta_s^*} \middle| x_{-s} \right) - f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} \middle| x_{-s} \right) \right] \left(\frac{-uh}{\theta_s^*} \right) dt, \end{aligned}$$

where again, by a change of variables ($z = tuh$),

$$\begin{aligned} \mathcal{R}_2 &\triangleq \int_0^1 \left[f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} - t \frac{uh}{\theta_s^*} \middle| x_{-s} \right) - f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} \middle| x_{-s} \right) \right] \left(-\frac{uh}{\theta_s^*} \right) dt \\ &= -\frac{1}{\theta_s^*} \int_0^{uh} f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} - \frac{z}{\theta_s^*} \middle| x_{-s} \right) - f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} \middle| x_{-s} \right) dz. \end{aligned}$$

By Assumption 5, $f(x_s|x_{-s})$ is Lipschitz-continuous with constant, say, C, which implies

$$\begin{aligned} \left| \frac{1}{\theta_s^*} \int_0^{uh} f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} - \frac{z}{\theta_s^*} \middle| x_{-s} \right) - f\left(\frac{1 - \mathring{x}_{-s}^T \theta_{-s}^*}{\theta_s^*} \middle| x_{-s} \right) dz \right| &\leq \frac{1}{\theta_s^{*2}} \int_0^{uh} C z dz \\ &= C \frac{(uh)^2}{2\theta_s^{*2}}. \end{aligned} \tag{67}$$

Hence we have

$$
\begin{aligned}
I_{-1} &\triangleq -\frac{h}{\theta_s^*}\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}} \mathring{v}_{-s}^T\mathring{x}_{-s}\bar{K}(u)f\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*-hu}{\theta_s^*}\middle|x_{-s}\right)f(x_{-s})dudx_{-s}\\
&= -\int_{\mathbb{R}^{p-1}}\mathring{v}_{-s}^T\mathring{x}_{-s}\int_{\mathbb{R}}F\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}\middle|x_{-s}\right)K(u)duf(x_{-s})dx_{-s}+\\
&-\int_{\mathbb{R}^{p-1}}\mathring{v}_{-s}^T\mathring{x}_{-s}\int_{\mathbb{R}}f\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}\middle|x_{-s}\right)\left(-\frac{hu}{\theta_s^*}\right)K(u)duf(x_{-s})dx_{-s}+\\
&-\int_{\mathbb{R}^{p-1}}\mathring{v}_{-s}^T\mathring{x}_{-s}\int_{\mathbb{R}}\mathcal{R}_2K(u)duf(x_{-s})dx_{-s}\\
&= I_{-1}^A+I_{-1}^B+I_{-1}^C. \qquad (68)
\end{aligned}
$$

Since the kernel density is symmetric around zero, $I_{-1}^B=0$, and since the density $K(\cdot)$ integrates to 1,

$$
\begin{aligned}
I_{-1}^A &= -\int_{\mathbb{R}^{p-1}}\mathring{v}_{-s}^T\mathring{x}_{-s}F\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}\middle|x_{-s}\right)f(x_{-s})dx_{-s}\\
&= -\int_{\mathbb{R}^{p-1}}\mathring{v}_{-s}^T\mathring{x}_{-s}\int_{-\infty}^{\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}}f(x_s|x_{-s})dx_sf(x_{-s})dx_{-s}\\
&= -\int_{\mathbb{R}^p}\mathring{v}_{-s}^T\mathring{x}_{-s}\mathbb{1}\left\{\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}\geq x_s\right\}f(x)dx\\
&= -\mathring{v}_{-s}^T\mathbb{E}_{X|Y=1}\left[Y\mathring{X}_{-s}\mathbb{1}\left\{1-Y\mathring{X}^T\theta^*\geq 0\right\}\right]. \qquad (69)
\end{aligned}
$$

Also, by (67),

$$
\begin{aligned}
|I_{-1}^C| &= \left|\int_{\mathbb{R}^{p-1}}\mathring{v}_{-s}^T\mathring{x}_{-s}\int_{\mathbb{R}}\mathcal{R}_2K(u)duf(x_{-s})dx_{-s}\right|\\
&\leq \int_{\mathbb{R}^{p-1}}\left|\mathring{v}_{-s}^T\mathring{x}_{-s}\right|\int_{\mathbb{R}}\left|\mathcal{R}_2\right|K(u)duf(x_{-s})dx_{-s}\\
&\leq C\frac{h^2}{2\theta_s^{*2}}\int_{\mathbb{R}^{p-1}}\left|\mathring{v}_{-s}^T\mathring{x}_{-s}\right|\left[\int_{\mathbb{R}}u^2K(u)du\right]f(x_{-s})dx_{-s}\\
&\leq C\frac{h^2\kappa_2}{2\theta_s^{*2}}\int_{\mathbb{R}^{p-1}}\left|\mathring{v}_{-s}^T\mathring{x}_{-s}\right|f(x_{-s})dx_{-s}\\
&\leq C\frac{h^2\kappa_2}{2\theta_s^{*2}}\mathring{\mu}_1,
\end{aligned}
$$

where the last two inequalities use Assumption 4 and the notation introduced in the Assumption 3. ■

C.2 Proof of Lemma 19 (Centred Score Function)

The proof uses the following result, due to Cai and Liu (2011).

Lemma 28 (Cai and Liu, 2011) *Let $Z_1, \ldots, Z_n$ be independent random variables with zero mean. Suppose that there exists some $t > 0$ and Γ_n such that $\sum_{i=1}^n \mathbb{E}\, Z_i^2 e^{t|Z_i|} \leq \Gamma_n^2$. Then for $0 \leq \eta \leq \Gamma_n$,*

$$\mathbb{P}\left[\sum_{i=1}^n Z_i \geq C_t \Gamma_n \eta\right] \leq \exp(-\eta^2),$$

where $C_t = t + t^{-1}$.

By the triangle inequality

$$\left\|\nabla \widehat{L}_h(\theta^*) - \mathbb{E}\big[\nabla \widehat{L}_h(\theta^*)\big]\right\|_2 \leq \left\|\nabla \widehat{L}_h(\theta^*) - \nabla \widehat{L}(\theta^*) - \mathbb{E}\big[\nabla \widehat{L}_h(\theta^*)\big]\right\|_2 + \left\|\nabla \widehat{L}(\theta^*)\right\|_2. \quad (70)$$

We use Lemma 28 to bound both terms on the right-hand side. The bounds are provided by the following auxiliary lemmas.

Lemma 29 *Under Assumption 3, for some constant $C > 0$ and for any t and $n \gtrsim p + t$,*

$$\mathbb{P}\left[\left\|\nabla \widehat{L}(\theta^*)\right\|_2 \leq C\sqrt{\frac{p+t}{n}}\right] \geq 1 - e^{-t},$$

for any $t > 0$.

Lemma 30 *Under Assumptions 2, 3, 4 and 5, for some constant $C > 0$, for any $t > 0$, as long as $1 \gtrsim h \gtrsim \frac{p+t}{n}$,*

$$\mathbb{P}\left[\left\|\nabla \widehat{L}_h(\theta^*) - \mathbb{E}\left[\nabla \widehat{L}_h(\theta^*)\right]\right\|_2 \leq C\sqrt{\frac{h(p+t)}{n}}\right] \geq 1 - e^{-t}.$$

Hence, using Lemmas 29 and 30,

$$\mathbb{P}\left[\left\|\nabla \widehat{L}_h(\theta^*) - \mathbb{E}\big[\nabla \widehat{L}_h(\theta^*)\big]\right\|_2 \leq C\frac{(1+\sqrt{h})}{\sqrt{n}}\sqrt{p+t}\right] \geq 1 - 2e^{-\tau},$$

for any $t > 0$, such that $t < \min\{Cnh - p, Cn - p\}$. ■

C.2.1 Proof of Lemma 29

For any $\delta \in (0, 1)$ there exists a δ-net N_δ of the unit sphere, such that

$$|N_\delta| \leq \left(\frac{2}{\delta} + 1\right)^{p+1}.$$

Let $v_s, \ldots, v_{N_\delta}$ be unit vectors constituting the δ-net. Then,

$$\left\|\nabla \widehat{L}(\theta^*)\right\|_2 \leq \left(\frac{1}{1-\delta}\right) \sup_{j \in \{1,\ldots,N_\delta\}} v_j^T \nabla \widehat{L}(\theta^*).$$

Let $v \in N_\delta$ and note that

$$v^T \nabla \widehat{L}(\theta^*) = v^T \big(\nabla \widehat{L}(\theta^*) - \mathbb{E}\big[\nabla \widehat{L}(\theta^*)\big] \big),$$

We use the Lemma 28 to bound this term. First note that for any non-negative random variable X, any $t > 0$ and $m \geq 0$,

$$X^m \leq \left(\frac{m}{te}\right)^m e^{tX}. \tag{71}$$

Then, the condition required by Lemma 28 can be easily verified since for any $t_0 > 0$ we have

$$\begin{aligned}
&\mathbb{E}\left(Y \mathring{v}^T \mathring{X} \mathbb{1}\left\{ 1 - Y \mathring{X}^T \theta^* \geq 0 \right\} \right)^2 \exp\left(\frac{t_0}{2} \left| Y \mathring{v}^T \mathring{X} \mathbb{1}\left\{ 1 - Y \mathring{X}_i^T \theta^* \geq 0 \right\} \right| \right) \\
&\quad \leq \mathbb{E}(\mathring{v}^T \mathring{X})^2 \exp\left(\frac{t_0}{2} \left| \mathring{v}^T \mathring{X} \right| \right) \leq C\, \mathbb{E} \exp\left(t_0 \left| \mathring{v}^T \mathring{X} \right| \right).
\end{aligned}$$

Note that, by Assumption 3, $\mathbb{E} \exp\big(t_0 \big| \mathring{v}^T \mathring{X} \big|\big) \leq C \exp(t_0^2 \nu_1^2 / 2)$. Thus, taking $Z_i' = Y_i \mathring{v}^T \mathring{X}_i \mathbb{1}\big\{1 - Y_i \mathring{X}_i^T \theta^* \geq 0\big\}$ we have $\sum_{i=1}^n \mathbb{E}\, Z_i'^2 e^{t_0 |Z_i'|/2} \leq nC \exp(t_0^2 \nu_1^2 / 2)$.

Let $\Gamma_n = \sqrt{nC}$ and $\eta = \sqrt{\gamma p \log n}$. By definition, $\mathbb{E}\big[\nabla \widehat{L}(\theta^*)\big] = 0$, and using Lemma 28 we get, for any γ such that $C \frac{n}{p \log n} > \gamma > 0$,

$$\mathbb{P}\left[\mathring{v}^T \nabla \hat{L}(\theta^*) \geq \sqrt{\frac{C \gamma p \log n}{n}} \right] \leq \exp(-\gamma p \log n) = n^{-\gamma p}. \tag{72}$$

Finally, applying a union bound over all vectors v in the net $\mathcal{N}_\delta$, $|\mathcal{N}_\delta| \leq (1 + 2/\delta)^p$, from equation (72) we obtain,

$$\mathbb{P}\left[\left\| \nabla \widehat{L}(\theta^*) \right\|_2 \leq \sqrt{\frac{C \gamma p \log n}{n}} \right] \geq 1 - \left(1 + \frac{2}{\delta}\right)^p n^{-\gamma p} = 1 - \exp\left[p \log\left(1 + \frac{2}{\delta}\right) - \gamma p \log n \right].$$

Taking $\delta = 2/(e-1)$ yields probability $1 - \exp\left[p - \gamma p \log n\right]$.Taking $t = \gamma p \log n$,

$$\mathbb{P}\left[\left\| \nabla \widehat{L}(\theta^*) \right\|_2 \leq \sqrt{\frac{Ct}{n}} \right] \geq 1 - \exp\left[p - t\right].$$

Setting $t = p + t'$, $t' > 0$ we obtain, for any $t' > 0$,

$$\mathbb{P}\left[\left\| \nabla \widehat{L}(\theta^*) \right\|_2 \leq \sqrt{\frac{C(p + t')}{n}} \right] \geq 1 - e^{-t'}.$$

Lastly, conditions of the Lemma 28 require $\gamma p \log n < Cn$ and hence $n \gtrsim p + t'$. ■

C.2.2 Proof of Lemma 30

Let $E(\theta^*) \triangleq 1 - Y\mathring{X}^T\theta^*$. To use Lemma 28 again, consider, for some $t > 0$,

$$\begin{aligned}
&\mathbb{E}\left[Y\mathring{v}^T\mathring{X}\bar{K}\left(\frac{E(\theta^*)}{h}\right) - Y\mathring{v}^T\mathring{X}\,\mathbb{1}\{E(\theta^*) \geq 0\}\right]^2 \exp\left[t|\mathring{v}^T\mathring{X}Y|\left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right|\right] \\
&= \mathbb{E}(\mathring{v}^T\mathring{X})^2\left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\left\{E(\theta^*) \geq 0\right\}\right]^2 \exp\left[t|\mathring{v}^T\mathring{X}|\left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right|\right] \\
&\leq 2\,\mathbb{E}(v_s\mathring{X}_s)^2\left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right]^2 \exp\left[t|\mathring{v}^T\mathring{X}|\left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right|\right] \\
&+ 2\,\mathbb{E}(\mathring{v}_{-s}^T\mathring{X}_{-s})^2\left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\left\{E(\theta^*) \geq 0\right\}\right]^2 \exp\left[t|\mathring{v}^T\mathring{X}|\left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right|\right],
\end{aligned} \tag{73}$$

where the inequality uses $a^2+b^2 \geq -2ab$. As before, we condition on Y and take expectation with respect to X first. Similarly, we use the law of iterated expectation to condition on X_{-s}.

We start by considering the first term in the sum in (73) and conditioning on $Y = 1$. Using a change of variables $u = \frac{1-x^T\theta^*}{h}$ and Assumption 5 (first inequality), triangle inequality and the fact that $\bar{K}(u) \leq 1$ for any $u \in \mathbb{R}$ (second inequality) and the assumption $K(\cdot)$ is symmetric around zero (last equality), we obtain, letting $e \triangleq 1 - \mathring{x}^T\theta^*$,

$$\begin{aligned}
&\int_{\mathbb{R}}(v_s\mathring{x}_s)^2\left[\bar{K}\left(\frac{e}{h}\right) - \mathbb{1}\{e \geq 0\}\right]^2 \exp\left[t(|v_s\mathring{x}_s| + |v_{-s}^T\mathring{x}_{-s}|)\left|\bar{K}\left(\frac{e}{h}\right) - \mathbb{1}\{e \geq 0\}\right|\right] f(x_s|x_{-s})dx_s \\
&\leq -\frac{hC}{\theta_s^*}v_s^2\int_{\mathbb{R}}(\bar{K}(u) - \mathbb{1}(u))^2 \exp\left[t\left(\left|v_s\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s} - hu}{\theta_s^*}\right| + t|v_{-s}^T\mathring{x}_{-s}|\right)(\mathbb{1}(u) - \bar{K}(u))\right] du \\
&\leq -\frac{hC}{\theta_s^*}v_s^2 \exp\left(t|v_{-s}^T\mathring{x}_{-s}| + |v_s|\left|\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}}{\theta_s^*}\right|\right)\int_0^\infty (1-\bar{K}(u))^2 \exp\left[\frac{t|v_s|hu}{\theta_s^*}(1-\bar{K}(u))\right] du \\
&- \frac{hC}{\theta_s^*}v_s^2 \exp\left(t|v_{-s}^T\mathring{x}_{-s}| + |v_s|\left|\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}}{\theta_s^*}\right|\right)\int_{-\infty}^0 \bar{K}(u)^2 \exp\left[t\left(|v_s|\,\frac{h}{\theta_s^*}(-u)\right)\bar{K}(u)\right] du \\
&= -\frac{v_s^2hC}{\theta_s^*} \exp\left(t|v_{-s}^T\mathring{x}_{-s}| + |v_s|\left|\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}}{\theta_s^*}\right|\right)\int_0^\infty (1-\bar{K}(u))^2 \exp\left[\frac{t|v_s|hu}{\theta_s^*}(1-\bar{K}(u))\right] du.
\end{aligned} \tag{74}$$

We start by upper-bounding the integral in (74). By definition $1 - \bar{K}(u) = \mathbb{P}(U \geq u)$, where random variable U has density K. Hence the behaviour of the integral in equation (74) is governed by the tail behaviour of the kernel density. By the discussion following Assumption 4,

$$\begin{aligned}
&\int_0^\infty (1-\bar{K}(u))^2 \exp\left[t\left(|v_s|\,\frac{hu}{\theta_s^*}\right)(1-\bar{K}(u))\right] du \\
&\qquad = \int_0^\infty \mathbb{P}(U \geq u)^2 \exp\left[t\left(|v_s|\,\frac{hu}{\theta_s^*}\right)\mathbb{P}(U \geq u)\right] du \\
&\qquad \leq u' \exp\left[t\,|v_s|\,\frac{hu'}{\theta_s^*}\right] + \int_{u'}^\infty Cu^{-2\alpha} \exp\left[t\left(|v_s|\,\frac{hu}{\theta_s^*}\right)Cu^{-\alpha}\right] du.
\end{aligned} \tag{75}$$

Since $\alpha \geq 1$, $\exp(\cdot)$ is a decreasing function of u and hence, for $u \geq u'$,

$$\exp\left[t\left(|v_s|\,\frac{hu}{\theta^*_s}\right)Cu^{-\alpha}\right] \leq \exp\left[t\,|v_s|\,\frac{h}{\theta^*_s}Cu'^{1-\alpha}\right],$$

which is a constant. Hence

$$\begin{aligned}\int_0^\infty (1-\bar{K}(u))^2 \exp\left[t\left(|v_s|\,\frac{hu}{\theta^*_s}\right)(1-\bar{K}(u))\right] du \\ \leq u' \exp\left[t\,|v_s|\,\frac{hu'}{\theta^*_s}\right] + C\exp\left[t\,|v_s|\,\frac{h}{\theta^*_s}Cu'^{1-\alpha}\right]\frac{1}{2\alpha-1}u'^{1-2\alpha}. \end{aligned} \tag{76}$$

Noting that the right hand side is independent of U or X_{-s} we can now turn to the term outside of integral in equation (74). By Assumption 3,

$$\mathbb{E}_{X|Y=1}\exp\left(t|\mathring{v}_{-s}\mathring{X}_{-s}| + t|v_s|\left|\frac{1-\theta^*_{-s}\mathring{X}_{-s}}{\theta^*_s}\right|\right) \leq C\exp\left[\frac{\nu_1^2t^2}{2}\left(1+\frac{\|\theta^*_{-s}\|_2}{\theta^*_s}\right)^2\right]. \tag{77}$$

Overall, using (74), (76) and (77), we have shown that, for any $t > 0$,

$$\begin{aligned}\mathbb{E}_{X|Y=1}(v_sX_s)^2\left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\left\{E(\theta^*)\geq 0\right\}\right]^2 \exp\left[t|\mathring{v}^T\mathring{X}|\left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\left\{E(\theta^*)\geq 0\right\}\right|\right] \\ \leq \frac{hC}{\theta^*_s}v_s^2\exp\left[\frac{tCh}{\theta^*_s} + \frac{\nu_1^2t^2}{2}\left(1+\frac{\|\theta^*_{-s}\|_2}{\theta^*_s}\right)^2\right]. \end{aligned} \tag{78}$$

Note that, theoretically the right-hand side of the inequality is unbounded for $h \to \infty$. However, as we are interested in $h \to 0$ we can take, say, $h \leq C'$ and define C appropriately to simplify the expression to

$$\begin{aligned}&\mathbb{E}_{X|Y=1}(v_sx_s)^2\left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\left\{E(\theta^*)\geq 0\right\}\right]^2 \exp\left[t|\mathring{v}^T\mathring{X}|\left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\left\{E(\theta^*)\geq 0\right\}\right|\right] \\ &\leq C\frac{h}{\theta^*_s}\exp\left[\frac{tC}{\theta^*_s} + \frac{\nu_1^2t^2}{2}\left(1+\frac{\|\theta^*_{-s}\|_2}{\theta^*_s}\right)^2\right] \\ &= C\frac{h}{\theta^*_s}\exp\big[Ct(t+1)\big]. \end{aligned}$$

We now address the second term in the equation (73). Again, condition on $Y = 1$ and X_{-s} first. Using equations (71) and (76), and Assumptions 5 and 3, we obtain, for any $t > 0$ and $t' > 0$,

$$\begin{aligned}&\mathbb{E}_{X|Y=1}(\mathring{v}^T_{-s}\mathring{X}_{-s})^2\left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*)\geq 0\}\right]^2\exp\left[t|\mathring{v}^T\mathring{X}|\left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*)\geq 0\}\right|\right] \\ &\leq C\frac{h}{\theta^*_s}\left(\frac{2}{t'e}\right)^2\mathbb{E}\exp\left[t'|\mathring{v}^T_{-s}\mathring{X}_{-s}| + t|\mathring{v}^T_{-s}\mathring{X}_{-s}| + t|v_s|\frac{|1-\mathring{X}^T_{-s}\theta^*_{-s}|}{\theta^*_s}\right] \\ &\leq C\frac{h}{\theta^*_s}\exp\left(\frac{1}{2}\nu_1^2\left[t' + t\left(1+\frac{\|\theta^*_{-s}\|_2}{\theta^*_s}\right)\right]^2\right). \end{aligned}$$

Letting $t' = t$,

$$\mathbb{E}_{X|Y=1}(\mathring{v}_{-s}^T \mathring{X}_{-s})^2 \left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right]^2 \exp\left[t|\mathring{v}^T \mathring{X}| \left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right|\right] \leq C\frac{h}{\theta_s^*} \exp\left[Ct(1+t)\right]. \tag{79}$$

Overall, from equations (78) and (79) we have, for any $t > 0$, conditional on $Y = 1$,

$$\mathbb{E}_{X|Y=1}(\mathring{v}^T \mathring{X})^2 \left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right]^2 \exp\left[t|\mathring{v}^T \mathring{X}Y| \left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right|\right] \leq Ch \exp[Ct(t+1)].$$

Analogous derivations hold for $Y = -1$, which leads us to the following result. There exists a constant $C > 0$ such that for any $t > 0$,

$$\mathbb{E}(Y\mathring{v}^T \mathring{X})^2 \left[\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\left\{E(\theta^*) \geq 0\right\}\right]^2 \exp\left[t|\mathring{v}^T \mathring{X}Y| \left|\bar{K}\left(\frac{E(\theta^*)}{h}\right) - \mathbb{1}\{E(\theta^*) \geq 0\}\right|\right] \leq Ch \exp[Ct(t+1)]. \tag{80}$$

Let

$$Z_i = Y_i \mathring{v}^T \mathring{X}_i \bar{K}\left(\frac{1 - Y_i \mathring{X}_i^T \theta^*}{h}\right) - Y_i \mathring{v}^T \mathring{X}_i \mathbb{1}\left\{1 - \mathring{X}_i^T \theta^* Y_i \geq 0\right\}.$$

Then equation (80) for $t = 1$ implies that, for some constant $C > 0$,

$$\sum_{i=1}^n \mathbb{E} Z_i^2 e^{|Z_i|} \leq Chn. \tag{81}$$

Setting $\Gamma_n = \sqrt{Cnh}$ and $\eta = \sqrt{\gamma p \log n}$ we obtain, using Lemma 28,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n Z_i \geq \sqrt{\frac{Ch\gamma p \log n}{n}}\right] \leq \exp(-\gamma p \log n) = n^{-\gamma p}, \tag{82}$$

as long as $\gamma < \frac{Cnh}{p \log n}$.

Finally, applying a union bound over all vectors v in the net $\mathcal{N}_\delta$, $|\mathcal{N}_\delta| \leq (1 + 2/\delta)^p$, with $\delta = 2/(e-1)$,

$$\mathbb{P}\left[\left\|\nabla \widehat{L}_h(\theta^*) - \mathbb{E}\left[\nabla \widehat{L}_h(\theta^*)\right]\right\|_2 \leq C\sqrt{\frac{\gamma h p \log n}{n}}\right] \geq 1 - \exp(p - \gamma p \log n).$$

Let $t'' = \gamma p \log n$. Then,

$$\mathbb{P}\left[\left\|\nabla \widehat{L}_h(\theta^*) - \mathbb{E}\left[\nabla \widehat{L}_h(\theta^*)\right]\right\|_2 \leq C\sqrt{\frac{ht''}{n}}\right] \geq 1 - e^{p - t''}.$$

We can thus let $t'' = p + \tilde{t}$, $\tilde{t} > 0$ to obtain the final bound,

$$\mathbb{P}\left[\left\|\nabla \widehat{L}_h(\theta^*) - \mathbb{E}\left[\nabla \widehat{L}_h(\theta^*)\right]\right\|_2 \leq C\sqrt{\frac{h(\tilde{t}+p)}{n}}\right] \geq 1 - e^{-\tilde{t}}.$$

Lastly, conditions of the Lemma 28 require $\gamma p \log n < Cnh$ and hence $\tilde{t} < Cnh - p$. The result thus holds for bandwidth $h \gtrsim \frac{p+\tilde{t}}{n}$. ■

C.3 Proof of Lemma 21

Firstly, recall that $\theta^*_s > C$ for some $C > 0$ by Assumption 2. Thus, by Theorem 9, $\hat{\theta}_s > C'$ for some $C' > 0$. By definition,

$$\begin{aligned}
&\|\nabla^2 L_h(\theta^* + \Delta) - \nabla^2 L_h(\theta^*)\|_2 \\
&= \left\|\mathbb{E}\left[\frac{\mathring{X}\mathring{X}^T}{h} K\left(\frac{1 - Y\mathring{X}^T\theta^* - \mathring{X}^T\Delta}{h}\right)\right] - \mathbb{E}\left[\frac{\mathring{X}\mathring{X}^T}{h} K\left(\frac{1 - Y\mathring{X}^T\theta^*}{h}\right)\right]\right\|_2.
\end{aligned}$$

Let $\mathring{v} \in \mathbb{S}^p$. Considering the first term and conditioning on $Y = 1$,

$$\begin{aligned}
&\int_{\mathcal{X}} \frac{(\mathring{v}^T\mathring{x})^2}{h} K\left(\frac{1 - \mathring{x}^T\theta^* - \mathring{x}^T\Delta}{h}\right) f(x)dx \\
&= \int_{\mathcal{X}_{-s}} \int_{\mathbb{R}} \frac{(\mathring{v}^T\mathring{x})^2}{h} K\left(\frac{1 - \mathring{x}^T\hat{\theta}}{h}\right) f(x_s|x_{-s}) f(x_{-s}) dx_s dx_{-s}.
\end{aligned}$$

Using substitution $u = \frac{1 - \mathring{x}^T\theta^* - \mathring{x}^T\Delta}{h}$, i.e. $x_s = \frac{1 - \hat{\theta}^T_{-s}\mathring{x}_{-s} - uh}{\hat{\theta}_s}$,

$$\begin{aligned}
&\frac{1}{h}\int_{\mathbb{R}} K\left(\frac{1 - \mathring{x}^T\theta^* - \mathring{x}^T\Delta}{h}\right)(\mathring{v}^T\mathring{x})^2 f(x_s|x_{-s})dx_s \\
&= -\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\left(v_0 + v^T_{-s}x_{-s} + v_s\frac{1 - \hat{\theta}^T_{-s}\mathring{x}_{-s} - uh}{\hat{\theta}_s}\right)^2 f\left(\frac{1 - \hat{\theta}^T_{-s}\mathring{x}_{-s} - uh}{\hat{\theta}_s}\middle| x_{-s}\right)du \\
&= -\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\, v_s^2\left(\frac{1 - \hat{\theta}^T_{-s}\mathring{x}_{-s} - uh}{\hat{\theta}_s}\right)^2 f\left(\frac{1 - \hat{\theta}^T_{-s}\mathring{x}_{-s} - uh}{\hat{\theta}_s}\middle| x_{-s}\right)du \\
&- \frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\left(v_0 + v^T_{-s}x_{-s}\right)^2 f\left(\frac{1 - \hat{\theta}^T_{-s}\mathring{x}_{-s} - uh}{\hat{\theta}_s}\middle| x_{-s}\right)du \\
&- \frac{2}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)(v_0 + v^T_{-s}x_{-s})v_s\left(\frac{1 - \hat{\theta}^T_{-s}\mathring{x}_{-s} - uh}{\hat{\theta}_s}\right) f\left(\frac{1 - \hat{\theta}^T_{-s}\mathring{x}_{-s} - uh}{\hat{\theta}_s}\middle| x_{-s}\right)du \quad (83)
\end{aligned}$$

The first term in (83) can be bounded using Taylor expansion (of $x^2 f(x|x_{-s})$) around $\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*-uh}{h}$. Specifically,

$$\begin{aligned}
&v_s^2 K(u)\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-\Delta^T\mathring{x}-uh}{\hat{\theta}_s}\right)^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-\Delta^T\mathring{x}-hu}{\theta_s^*}\middle| x_{-s}\right)\\
&= v_s^2 K(u)\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\right)^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)\\
&+ v_s^2 K(u)\int_0^1\left[2\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+t\tilde{\Delta}\right) f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+t\tilde{\Delta}\middle| x_{-s}\right)\right]\tilde{\Delta}dt\\
&+ v_s^2 K(u)\int_0^1\left[\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+t\tilde{\Delta}\right)^2 f'\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+t\tilde{\Delta}\middle| x_{-s}\right)\right]\tilde{\Delta}dt\\
&\le v_s^2 K(u)\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\right)^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)+v_s^2 CK(u)\tilde{\Delta}, \qquad (84)
\end{aligned}$$

where we used Assumption 5 and the notation

$$\tilde{\Delta}\triangleq\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}-\frac{1-\hat{\theta}_{-s}^{T}\mathring{x}_{-s}-uh}{\hat{\theta}_s}.$$

Note that $\tilde{\Delta}$ is a function of u.

By Taylor expansion of the conditional density around $(1-\theta_{-s}^{*T}\mathring{X}_{-s}-uh)/\theta_s^*$ and Assumption 5,

$$\begin{aligned}
&f\left(\frac{1-\hat{\theta}_{-s}^{T}\mathring{x}_{-s}-uh}{\hat{\theta}_s}\middle| x_{-s}\right)\\
&= f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)+\int_0^1 f'\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+t\tilde{\Delta}\middle| x_{-s}\right)\tilde{\Delta}dt\\
&\le f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)+C\tilde{\Delta}.
\end{aligned}$$

Thus, for the second term in (83) we have,

$$\begin{aligned}
&-\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\left(v_0+v_{-s}^T x_{-s}\right)^2 f\left(\frac{1-\hat{\theta}_{-s}^{T}\mathring{x}_{-s}-uh}{\hat{\theta}_s}\middle| x_{-s}\right)du\\
&\le -\int_{\mathbb{R}}\frac{K(u)}{\hat{\theta}_s}(v_0+v_{-s}^T x_{-s})^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)du\\
&-\int_{\mathbb{R}}\frac{CK(u)}{\hat{\theta}_s}(v_0+v_{-s}^T x_{-s})^2\tilde{\Delta}du. \qquad (85)
\end{aligned}$$

Using a Taylor expansion of $\left(\frac{1-\mathring{x}_{-s}^T\hat{\theta}_{-s}-hu}{\hat{\theta}_s}\right) f\left(\frac{1-\hat{\theta}_{-s}^T\mathring{x}_{-s}-uh}{\hat{\theta}_s}\middle| x_{-s}\right)$, the last term in (83) can be written as,

$$
\begin{aligned}
&-\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\, v_s\left(v_0+v_{-s}^T x_{-s}\right)\left(\frac{1-\mathring{x}_{-s}^T\hat{\theta}_{-s}-hu}{\hat{\theta}_s}\right) f\left(\frac{1-\hat{\theta}_{-s}^T\mathring{x}_{-s}-uh}{\hat{\theta}_s}\middle| x_{-s}\right)du\\
&=-\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\, v_s\left(v_0+v_{-s}^T x_{-s}\right)\int_0^1 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+t\tilde{\Delta}\middle| x_{-s}\right)\tilde{\Delta}dtdu\\
&-\frac{v_s v_0}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\int_0^1\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*-uh}{\theta_s^*}+t\tilde{\Delta}\right) f'\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+t\tilde{\Delta}\middle| x_{-s}\right)\tilde{\Delta}dtdu\\
&-\frac{v_s v_{-s}^T x_{-s}}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\int_0^1\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*-uh+t\tilde{\Delta}\theta_s^*}{\theta_s^*} f'\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh+t\tilde{\Delta}\theta_s^*}{\theta_s^*}\middle| x_{-s}\right)\tilde{\Delta}dtdu.
\end{aligned}
$$

Thus, using Assumption 5,

$$
\begin{aligned}
&-\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\, v_s\left(v_0+v_{-s}^T x_{-s}\right)\left(\frac{1-\mathring{x}_{-s}^T\hat{\theta}_{-s}-hu}{\hat{\theta}_s}\right) f\left(\frac{1-\hat{\theta}_{-s}^T\mathring{x}_{-s}-uh}{\hat{\theta}_s}\middle| x_{-s}\right)du\\
&\le-\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\, v_s\left(v_0+v_{-s}^T x_{-s}\right)\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*-uh}{\theta_s^*}\right) f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)du\\
&\quad-\frac{2}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\, v_s\left(v_0+v_{-s}^T x_{-s}\right)C\tilde{\Delta}du. \qquad (86)
\end{aligned}
$$

Equations (84), (85) and (86), combined with (83), imply that

$$
\begin{aligned}
&\frac{1}{h}\int_{\mathbb{R}} K\left(\frac{1-\mathring{x}^T\theta^*-\mathring{x}^T\Delta}{h}\right)(v^T x)^2 f(x_s|x_{-s})dx_s\\
&\le-\frac{v_s^2}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\right)^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)du\\
&-\frac{v_s^2}{\hat{\theta}_s}C\int_{\mathbb{R}} K(u)\tilde{\Delta}du\\
&-\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)(v_0+v_{-s}^T x_{-s})^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)du\\
&-\frac{C}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)(v_0+v_{-s}^T x_{-s})^2\tilde{\Delta}du\\
&-\frac{2}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\, v_s\left(v_0+v_{-s}^T x_{-s}\right)\left(\frac{1-\mathring{x}_{-s}^T\theta_{-s}^*-uh}{\theta_s^*}\right) f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)du\\
&-\frac{C}{\hat{\theta}_s}\int_{\mathbb{R}} K(u)\, v_s\left(v_0+v_{-s}^T x_{-s}\right)\tilde{\Delta}du,
\end{aligned}
$$

which can be further simplified to obtain,

$$\begin{aligned}
&\frac{1}{h}\int_{\mathbb{R}} K\left(\frac{1-\mathring{x}^T\theta^*-\mathring{x}^T\Delta}{h}\right)(v^Tx)^2 f(x_s|x_{-s})dx_s \\
&\leq -\frac{1}{\hat{\theta}_s}\int_{\mathbb{R}}\left(v_0+v_s\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+v_{-s}^Tx_{-s}\right)^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)K(u)du \\
&\quad -\frac{1}{\hat{\theta}_s}C\left(v_0+v_s+v_{-s}^Tx_{-s}\right)^2\int_{\mathbb{R}}K(u)\tilde{\Delta}du. \qquad (87)
\end{aligned}$$

Using $-\frac{1}{\hat{\theta}_s}=\frac{\hat{\theta}_s-\theta_s^*}{\hat{\theta}_s\theta_s^*}-\frac{1}{\theta_s^*}$ we can further rewrite equation (87) to obtain

$$\begin{aligned}
&\frac{1}{h}\int_{\mathbb{R}} K\left(\frac{1-\mathring{x}^T\theta^*-\mathring{x}^T\Delta}{h}\right)(\mathring{v}^T\mathring{x})^2 f(x_s|x_{-s})dx_s \\
&\leq -\frac{1}{\theta_s^*}\int_{\mathbb{R}}\left(v_0+v_s\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+v_{-s}^Tx_{-s}\right)^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)K(u)du \\
&+\frac{\hat{\theta}_s-\theta_s^*}{\hat{\theta}_s\theta_s^*}\int_{\mathbb{R}}\left(v_0+v_s\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}+v_{-s}^Tx_{-s}\right)^2 f\left(\frac{1-\theta_{-s}^{*T}\mathring{x}_{-s}-uh}{\theta_s^*}\middle| x_{-s}\right)K(u)du \\
&-\frac{1}{\hat{\theta}_s}C\left(v_0+v_s+v_{-s}^Tx_{-s}\right)^2\int_{\mathbb{R}}K(u)\tilde{\Delta}du.
\end{aligned}$$

Observe that,

$$\begin{aligned}
\tilde{\Delta} &= \frac{1}{\theta_s^*\hat{\theta}_s}\left[\left(1-x_{-s}^Tw_{-s}^*-b^*\right)(\hat{\theta}_s-\theta_s^*)+\theta_s^*(\hat{w}_{-s}-w_{-s}^*)^Tx_{-s}\right] \\
&\quad+\frac{1}{\theta_s^*\hat{\theta}_s}\left[\theta_s^*(\hat{\theta}_0-\theta_0^*)+hu(\theta_s^*-\hat{\theta}_s)\right] \\
&= \frac{1}{\theta_s^*\hat{\theta}_s}\left[\theta_s^*\begin{pmatrix}1\\ \frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}\\ x_{-s}\end{pmatrix}^T\begin{pmatrix}\Delta_0\\ \Delta_s\\ \Delta_{-s}\end{pmatrix}+hu(\theta_s^*-\hat{\theta}_s)\right].
\end{aligned}$$

By Assumption 4,

$$\int_{\mathbb{R}}K(u)\tilde{\Delta}du\leq\frac{1}{\theta_s^*\hat{\theta}_s}\left[\theta_s^*\begin{pmatrix}1\\ \frac{1-\mathring{x}_{-s}^T\theta_{-s}^*}{\theta_s^*}\\ x_{-s}\end{pmatrix}^T\begin{pmatrix}\Delta_0\\ \Delta_s\\ \Delta_{-s}\end{pmatrix}\right].$$

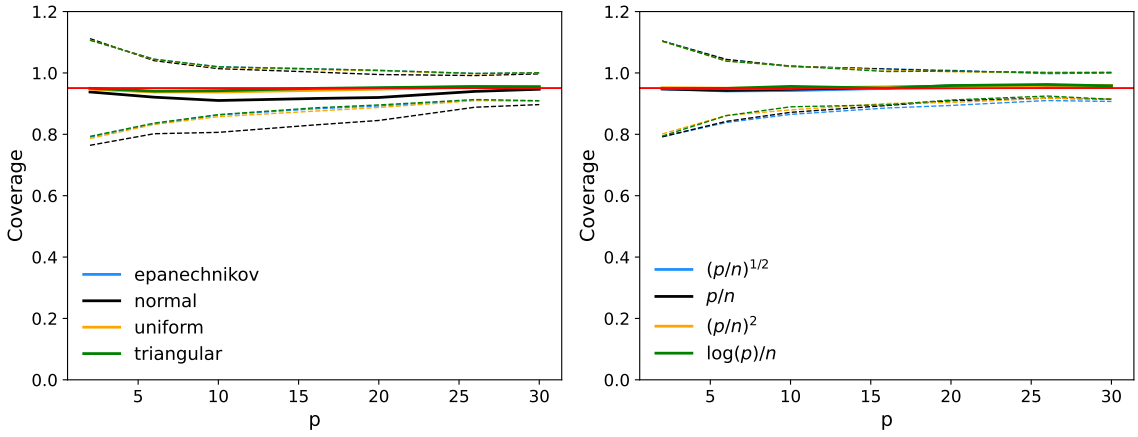


Figure 7: Mean and standard deviation bounds for coverage ratios for different choices of kernel (plot a) and bandwidth (plot b). Based on 500 simulations with $n = 500$. The red line depicts the target coverage.

Taking expectation over x_{-s}, using Assumptions 3 and 4, and a change of variables, we obtain,

$$\begin{aligned}
&\int_{\mathcal{X}_{-s}} \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{1-\mathring{x}^T\theta^* - \mathring{x}^T\Delta}{h}\right)(\mathring{v}^T\mathring{x})^2 f(x_s|x_{-s})f(x_{-s})dx_s dx_{-s} \\
&\leq \mathbb{E}_{X|Y=1}\left[\frac{(\mathring{v}^T\mathring{X})^2}{h}K\left(\frac{1-Y\mathring{X}^T\theta^*}{h}\right) + \frac{|\hat{\theta}_s - \theta_s^*|}{\hat{\theta}_s}\frac{(\mathring{v}^T\mathring{X})^2}{h}K\left(\frac{1-Y\mathring{X}^T\theta^*}{h}\right)\right] + \frac{Cr}{\theta_s^*\hat{\theta}_s^2} \\
&\leq \mathbb{E}_{X|Y=1}\left[\frac{1}{h}K\left(\frac{1-Y\mathring{X}^T\theta^*}{h}\right)(\mathring{v}^T\mathring{X})^2\right] + \frac{|\hat{\theta}_s - \theta_s^*|}{\hat{\theta}_s\theta_s^*}C\left[1 + 2\mu_1^f + \mu_2^f\right] + \frac{Cr}{\theta_s^*\hat{\theta}_s^2}.
\end{aligned}$$

Finally,

$$\mathbb{E}\left[\frac{1}{h}K\left(\frac{1-Y\mathring{X}^T\theta^* - \mathring{X}^T\Delta}{h}\right)\langle\mathring{v},\mathring{X}\rangle^2\right] - \mathbb{E}\left[\frac{1}{h}K\left(\frac{1-Y\mathring{X}^T\theta^*}{h}\right)\langle\mathring{v},\mathring{X}\rangle^2\right] \leq Cr.$$

■

C.4 Simulations

Throughout the paper, simulations were carried out using a Gaussian kernel and bandwidth $h = (p/n)^{1/4}$. To assess the robustness of the reported results to different choices of kernel and h, we calculate the median and standard deviation of coverage ratios for different values of p, as in Section 5.3. The results are reported in Figure 7.

References

Heather Battey, David R. Cox, and Michelle V. Jackson. On the linear in probability model for binary data. *Royal Society Open Science*, 6(5):190067, 2019. doi: 10.1098/rsos.190067.

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

David R. Cox and Nanny Wermuth. Response models for mixed binary and quantitative variables. *Biometrika*, 79(3):441–461, 1992. ISSN 00063444.

Victor H de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-Normalized Processes: Limit Theory and Statistical Applications*, volume 204. Springer, 2009.

Bradley Efron. Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64(328):1278–1302, 1969.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9: 1871–1874, 2008.

Marcelo Fernandes, Emmanuel Guerre, and Eduardo Horta. Smoothing quantile regressions. *Journal of Business & Economic Statistics*, 39(1):338–357, 2021.

Edgar C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185, 1954.

Xuming He, Xiaoou Pan, Kean Ming Tan, and Wen-Xin Zhou. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 232(2):367–388, 2023.

Joel L Horowitz. Bootstrap methods for median regression models. *Econometrica*, 66(6): 1327–1351, 1998.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and Sellamanickam Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *International Conference on Machine Learning (ICML)*, pages 408–415, 2008.

Thorsten Joachims. Making large-scale SVM learning practical. Technical report, 1998.

Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9(44): 1343–1368, 2008.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 2013.

Yuh-Jye Lee and Olvi L Mangasarian. SSVM: A smooth support vector machine for classification. *Computational optimization and Applications*, 20:5–22, 2001.

Ambrose Lo. Demystifying the integrated tail probability expectation formula. *The American Statistician*, 2018.

Olvi L. Mangasarian. Generalized support vector machines. *Advances in Large Margin Classifiers*, 2000.

Vladimir Spokoiny. Bernstein-von Mises theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*, 2013.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Johan A. K. Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

Kean Ming Tan, Heather Battey, and Wen-Xin Zhou. Communication-constrained distributed quantile regression with optimal statistical guarantees. *Journal of Machine Learning Research*, 23(272):1–61, 2022.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Boxiang Wang, Le Zhou, Yuwen Gu, and Hui Zou. Density-convoluted support vector machines for high-dimensional classification. *IEEE Transactions on Information Theory*, 69(4):2523–2536, 2023.

Xiaozhou Wang, Zhuoyi Yang, Xi Chen, and Weidong Liu. Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20, 2019.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning (ICML)*, page 116, 2004.

Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(1):53, 2016.