Annals of the Institute of Statistical Mathematics (2022) 74:167–194
https://doi.org/10.1007/s10463-021-00792-5

Asymptotic linear expansion of regularized M-estimators

Tino Werner[1]

Received: 5 September 2019 / Revised: 16 October 2020 / Accepted: 5 February 2021 /
Published online: 24 March 2021

Abstract
Parametric high-dimensional regression requires regularization terms to get interpretable models. The respective estimators correspond to regularized M-functionals which are naturally highly nonlinear. Their Gâteaux derivative, i.e., their influence curve linearizes the asymptotic bias of the estimator, but only up to a remainder term which is not guaranteed to tend (sufficiently fast) to zero uniformly on suitable tangent sets without profound arguments. We fill this gap by studying, in a unified framework, under which conditions the M-functionals corresponding to convex penalties as regularization are compactly differentiable, so that the estimators admit an asymptotically linear expansion. This key ingredient allows influence curves to reasonably enter model diagnosis and enable a fast, valid update formula, just requiring an evaluation of the corresponding influence curve at new data points. Moreover, this paves the way for optimally-robust estimators, bounding the influence curves in a suitable way.

Keywords Asymptotic linear expansion · Regularized M-estimators · Influence curves

1 Introduction

In the mid nineties, Robert Tibshirani succeeded in combining two important paradigms of fitting linear regression models, namely variable selection and shrinkage of the coefficients, in one single optimization problem, calling it the Lasso (least absolute shrinkage and selection operator, Tibshirani 1994). While already being superior to the former state-of-the-art procedures of Ridge regression and subset selection in terms of interpretability of the model and prediction accuracy (Tibshirani 1994), its popularity grew when (Efron et al. 2004) embedded the Lasso into the framework of forward stagewise regression and provided the LARS algorithm which turned out to

✉ Tino Werner
tino.werner1@uni-oldenburg.de

[1] Institute for Mathematics, Carl von Ossietzky University Oldenburg, P/O Box 2503, 26111 Oldenburg (Oldb), Germany

be more efficient than the implementations of Tibshirani and Osborne et al. (2000). A very competitive algorithm has been developed by (Friedman et al. 2007), relying on the fact that despite the Lasso for multiple regression does not have a closed form solution, a simple Lasso just concerning one single predictor has. Therefore, they apply the so-called ”shooting” algorithm to the Lasso and other suitable problems, which means that one repeatedly cycles through the variables, keeping all others fixed to their values of the previous iteration and fits the partial residual, i.e., a coordinate-wise optimization is done.

Evidently, the Lasso estimator is not robust since contamination of the data both in the responses as well as in the regressors can severely distort it. More precisely, let $y = f(x) + \epsilon$ be the underlying regression model where $(x, y) \sim F$ for a distribution F on $(\mathcal{X} \times \mathcal{Y}, \mathbb{B}(\mathcal{X} \times \mathcal{Y}))$ for the space $\mathcal{X} \subset \mathbb{R}^p$, $p \in \mathbb{N}$, of regressors and the space $\mathcal{Y} \subset \mathbb{R}$ of responses, the Borel sigma algebra $\mathbb{B}(\mathcal{X} \times \mathcal{Y})$ on $\mathcal{X} \times \mathcal{Y}$ and some error term $\epsilon \sim F_\epsilon$ with mean zero and variance $\sigma^2 \in (0, \infty)$. Regarding the distribution F as ”ideal distribution”, we assume that the true distribution of (x, y) stems from a convex contamination neighborhood that contains all distributions $F_t := (1-t)F + tH$ where H is an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and $t \in [0, 1]$ is the so-called contamination radius (see, e.g., Rieder 1994).

Robust statistics provides a diagnostic tool that quantifies the impact of contamination on the estimator. First, one of the most important contributions of robust statistics is the identification of estimators with functionals (Huber and Ronchetti 2009; Hampel et al. 2011; Rieder 1994; Maronna et al. 2006), so we write the estimation procedure as the statistical functional $T : \mathcal{F} \to \Theta$ for the space $\mathcal{F}$ of all distributions on $\mathcal{X} \times \mathcal{Y}$ and the parameter space $\Theta \subset \mathbb{R}^p$. The idea is to linearize this functional in a first-order expansion

$$T(F_t) - T(F) \approx T'(F)(F_t - F)$$

which essentially goes back to Von Mises (1947), but it also has been studied, for example, in (Reeds 1976), (Clarke 1983) and (Rieder 1994). The term T' is a functional derivative, in fact, usually a Gâteaux derivative that we specify in detail in Sect. 2. This derivative has been identified in Hampel (1974) with the so-called influence curve, i.e., we have

$$T(F_t) = T(F) + \int IC((x, y), T, F) d(F_t - F)((x, y)) + rem \tag{1}$$

for some stochastic remainder term *rem*.

Those influence curves play a major role in robust statistics since they quantify the infinitesimal influence of a single observation on the estimator. This is easily seen once we face real data where we have the empirical distribution function $\hat{F}_n$, n being the number of observations. Then, Eq. (1) becomes

$$\begin{aligned} T(\hat{F}_n) - T(F) = & \int IC((x,y),T,F)d(\hat{F}_n - F)(x,y) + rem \\ = & \frac{1}{n}\sum_i IC((x_i,y_i),T,F) + rem \end{aligned} \tag{2}$$

for observations (x_i, y_i), a remainder term of order $n^{-1/2}$ and where $\int ICdF = 0$ was used, one of the regularity conditions that are specified in Sect. 2. An estimator that allows for such an expansion in terms of influence curves is referred to as asymptotically linear estimator (ALE). The concept of influence curves is well-known, but its success is based on the much deeper theoretical fact that standard M-estimators as well as for example MD- and R-estimators are ALEs (see, e.g., Fernholz 1983).

The Gâteaux derivatives of certain penalized M-functionals have already been computed (Öllerer et al. 2015; Avella-Medina 2017). We note that (Avella-Medin 2017) already mentioned that a linearization of the estimators in the sense of an asymptotically linear expansion is necessary and pointed out that the remainder term in this linearization has to be controlled uniformly under suitable tangent sets, which in our statistical functional setting take the form of distributional neighborhoods. (Avella-Medina 2017) require this uniformity to cover bounded tangent sets which amounts to Fréchet differentiability (Averbukh and Smolyanov 1967) of the underlying statistical functional. However, (Avella-Medina 2017) restricted theirselves to contributing influence curves of regularized M-estimators. Note that an argumentation as in (Avella-Medina 2017) has not been done in the other references. (Bühlmann and Hothorn 2007) identified their functional gradient boosting algorithm with iteratively evaluating Gâteaux derivatives at the current model. There already exist results where the asymptotic linearity of penalized M-estimators has been shown (LeDell et al. 2015; Van de Geer 2016) and even remarkable results concerning penalized M-estimators in the nonconvex setting (Loh 2017), but to the best of our knowledge, in particular, a general unified theory for the asymptotic linearity of penalized M-estimators has not been provided by the aforementioned results so far.

It should be noted that even without this uniformity, the influence function has its merits, as it will provide an approximation in at least some directions, the problem being that it is not warranted that these directions cover the empirical process $\hat{F}_n - F$. Since the standard theory of ALEs does not apply for estimators based on non-differentiable target functions like the Lasso, these influence curves lack of a sound theoretical foundation. This a is major issue since simply computing the influence curves provide no guarantee that the remainder term in Eq. (2) is of order $n^{-1/2}$. But if this property does not hold, computing the asymptotic bias, i.e., $T(\hat{F}_n) - T(F)$, via influence curves can be highly misleading. The reason is that Gâteaux derivatives only allow for uniformly vanishing remainders locally on finite sets which does not suffice for the application of a functional delta method (since the chain rule is not applicable, cf. Rieder 1994, Thm. 1.2.9) which is an indispensable tool for deriving the asymptotically linear expansion. In fact, we need at least the stronger notion of differentiability in Hadamard sense, which warrants uniformity on compact sets of tangents, therefore also called compact differentiability. Once this uniformity on compact tangent sets is settled, we obtain, now on a valid base, the influence function as Gâteaux derivative. This gap is closed by our contribution, allowing all advantages like detecting regions where the input data have a high impact on the estimator by influence curves or a fast update construction when new instances appear, even for estimators like the Lasso. On top of that, the validity of the asymptotic linear expansion leads to asymptotic normality, including the corresponding simple confidence regions.

Robustness is actually often understood as the property that the estimator has a bounded influence curve. The is also called B-robustness (see, e.g., Hampel et al. 2011). In literature, robustness results by proving the boundedness of influence curves have already been established in the special case of regularized kernel-based regression problems, see (Christmann and Steinwart 2004). Since they require Fréchet-differentiability of the loss function which is not true, for example, for the $\epsilon-$ insensitive loss, (Christmann and Van Messem 2008) introduced Bouligand influences curves to prove the robustness of support vector regression estimators. The asymptotic normality of kernel-based regression methods has been shown in (Hable 2012) who also used the framework of compact differentiability of the corresponding functional. Once we face a robustness problem which frequently occurs for high-dimensional real data, a sophisticated strategy to robustify the estimator is to suitably bound the influence curve by solving constrained optimization problems (see, for example, Rieder 1994; Kohl 2005) which requires that the original estimator can be expressed in terms of influence curves. So, our contributions also pave the way for extending this principle to the case of regularized estimators.

Therefore, the main questions that we answer in this work are: Can we make assumptions under which regularized M-estimators are asymptotically linear? And if we can, is it possible to embed well-known regularized M-estimators like the Lasso into this framework?

The rest of this article is organized as follows. In Sect. 2, the general definition of $\mathcal{R}-$differentiability and relevant tools from robust statistics are revisited. In Sect. 3, results concerning asymptotic linearity are recapitulated which will be essential for the rest of this main section. Then, we transfer the results for M-estimators to the case of regularized M-estimators with convex penalties. Subsection 3.3 is the main theoretical part where we show under which conditions the asymptotic linearity is valid. For non-differentiable penalty terms, we will heavily rely on an approximation lemma of (Avella-Medina 2017). We also consider an extension of our results to ranking problems. Section 4 is devoted to concrete examples. In Sect. 5, we discuss how to deal with data-driven penalty parameters and in Sect. 6, we outline several benefits of our results in practice.

2 Preliminaries

This section compiles the concepts needed for the main section. We recur to the abstract definition of $\mathcal{R}-$differentiability of maps between normed vector spaces. The second part contains the most important definitions of quantitative robust statistics like the influence curve and ALEs. For the ease of notation, we formulate them in a more general than the regression setting, so we w.l.o.g. do not consider regressor-response pairs (x, y) in this section but only observations x.

2.1 Functional derivatives

We start compiling necessary notions on functional derivatives already defined in the 1970s in the statistical context. For the exposition, we largely follow (Rieder 1994), respectively,(Averbukh and Smolyanov 1967).

Definition 1 Let X, Y be normed real vector spaces. A map $T : X \to Y$ is $\mathcal{R}$**–differentiable** in $x \in X$ if there exists $d_{\mathcal{R}}T(x) \in L(X, Y)$, i.e., the space of all continuous linear maps from X to Y, and $s_0 > 0$ such that for all directions $h \in X$ it holds that

$$T(x + sh) = T(x) + d_{\mathcal{R}}T(x)sh + \rho(sh) \ \forall |s| \leq s_0,$$

where the remainder term ρ satisfies the following conditions: **i)** $\rho(0) = 0$, **ii)** $\rho \in \mathcal{R}(X, Y)$ where $\mathcal{R}(X, Y)$ is a real vector space with $\mathcal{R}(X, Y) \cap L(X, Y) = \{0\}$, so it can be identified with

$$\left\{ \rho : X \to Y \,\middle|\, \lim_{t \to 0} \left(\frac{||\rho(th)||}{t} \right) = \rho(0) = 0 \right\}.$$

Then, the continuous linear map (w.r.t. h per definition) $d_{\mathcal{R}}T(x)$ is referred to as the $\mathcal{R}$**– derivative of** T **at** x.

Note that the definition of $\mathcal{R}(X, Y)$ above does not require any uniformity of the remainder term along a set of directions. However, this uniformity will be indispensable for the results in this paper. The following definition taken from (Rieder 1994, Sec. 1) is helpful to distinguish between different types of $\mathcal{R}$–differentiation.

Definition 2 Let X, Y, T be as in Def. 1. Let $\mathcal{S}$ be a covering of X. Define the **remainder class**

$$\mathcal{R}_{\mathcal{S}}(X, Y) := \left\{ \rho : X \to Y \,\middle|\, \lim_{t \to 0} \left(\sup_{h \in S} \left(\frac{||\rho(th)||}{t} \right) \right) = \rho(0) = 0 \ \forall S \in \mathcal{S} \right\}.$$

By Def. 2, it is clear that $\mathcal{R}_{\mathcal{S}}$–differentiability can be seen as a linear approximation of some functional T such that the remainder term converges uniformly on all sets $S \in \mathcal{S}$. In the following, we define three special concepts of $\mathcal{R}_{\mathcal{S}}$–differentiability (cf. Rieder 1994). We say that the functional T is Gâteaux or weakly differentiable resp. Hadamard or compactly differentiable resp. Fréchet or boundedly differentiable if the covering $\mathcal{S}$ of X consists of finite resp. compact resp. bounded sets. Trivially, bounded differentiability implies compact differentiability which implies Gâteaux differentiability. The derivatives coincide in this case. Moreover, continuous Gâteaux differentiability implies bounded differentiability.

For the application of an infinite-dimensional delta method (Rieder 1994, Thm. 1.3.3), it is necessary to investigate whether the chain rule holds for functional derivatives. The following theorem, cf. (Rieder 1994, Prop. 1.2.6+Thm. 1.2.9) or (Averbukh and Smolyanov 1967, Thm. 1.6), shows that for linear maps as approximations, the chain rule holds if and only if at least Hadamard differentiability holds.

Theorem 1 (**Chain rule**) *Let X, Y, Z be normed real vector spaces and let $T : X \to Y$, $U : Y \to Z$. If T and U are compactly differentiable, then the chain rule holds, i.e., $d_H(U \circ T)(x) = d_H U(T(x)) \circ d_H T(x)$ where d_H denotes the Hadamard derivative. Conversely, if the chain rule holds, the maps are already compactly differentiable.*

The chain rule does not hold for Gâteaux differentiable maps in general. Counterexamples can be found in Averbukh and Smolyanov (1967) or Fréchet (1937). Thus, regarding the abovely mentioned concepts of functional derivatives, one can state that compact differentiability is the weakest form of $\mathcal{R}$–differentiability such that the chain rule holds.

Of course, there exist examples where Hadamard-differentiability fails. One typical example is L-statistics where the underlying distribution has an unbounded support as pointed out in Van der Vaart (2000). Then, compact differentiability is impossible w.r.t. $|| \cdot ||_\infty$. Such functionals are written in the form (cf. Beutner and Zähle 2010) $T_g(F) := -\int xdg(F(x))$, so it is required that g has a compact support in (0, 1) to ensure compact differentiability. It is shown in Beutner and Zähle (2016) that if the support of g contains at least one of the boundary points of [0, 1], even the negative expectation value (where g is the identity map) is not compactly differentiable. The functional T_g covers relevant statistical functionals like the value at risk or the average value at risk as pointed out in (Beutner and Zähle 2010). In fact, (Krätschmer et al. 2012) stated that tail-dependent functionals are in general not compactly differentiable w.r.t. uniform norms.

2.2 Basic concepts of quantitative robustness

Every real data analysis requires model assumptions. However, these assumptions are in general not fulfilled, hence the real data differ from data that would have been generated by the theoretical (ideal) model. Therefore, fitting models by using the real data can be seen as if one analyzes a contaminated data set which affects the quality of the fitted model. It is not desirable to exclude potential ”outliers” from the data set (cf. Hampel et al. 2011) but to find strategies that downweight them, like iteratively reweighted least squares (IRWLS), see, e.g., (Huber and Ronchetti 2009). In the introduction, we already implicitly defined the influence function as first-order derivative of the corresponding statistical functional. We now give a precise definition of the influence function (cf. Hampel 1974).

Definition 3 Let X be a normed function space and let Θ be a normed real vector space. Let $T : X \to \Theta$ be a statistical functional. The **influence function or influence curve** of T at x for a probability measure P on X is defined as the derivative

$$IC(x,T,P) := \lim_{t\to 0}\left(\frac{T((1-t)P+t\delta_x)-T(P)}{t}\right) = \partial_t\left[T((1-t)P+t\delta_x)\right]\Big|_{t=0}$$

where δ_x denotes the Dirac measure at x.

So, the influence curve is just a special Gâteaux derivative with $h := \delta_x - P$. The influence curve can be seen as an estimate for the infinitesimal influence of a single observation on the estimator. If the influence curve is unbounded, then a single observation can have an infinite impact on the resulting estimator which is of course not desirable. For robustness properties, it is necessary that the influence curve is at least bounded. In that case, the estimator is sometimes called B-robust (see Hampel et al. 2011 or Van der Vaart 2000).

The robustification of an estimator can be done by robustifying its influence function. Minimax results for optimal-robust influence curves have been established in several works (Rieder 1994, Rieder et al. 2008; Hampel et al. 2011; Fraiman et al. 2001). However, for guaranteeing optimality of these approaches, it is crucial that the estimator is asymptotically linear (Rieder 1994, Def. 4.2.16) and that the model is smooth enough, i.e., that it is L_2–differentiable (see LeCam 1970), see Def. 5.

Definition 4 Let $(\Omega^n, \mathcal{A}^n)$ be a measurable space and let $S_n : (\Omega^n, \mathcal{A}^n) \to (\mathbb{R}^p, \mathbb{B}^p)$ be an estimator based on observations $x_1, ..., x_n$. Let $\mathcal{P} := \{P_\theta \mid \theta \in \Theta\}$ be a parametric distribution family on $(\Omega, \mathcal{A})$ for some parameter space $\Theta \subset \mathbb{R}^p$. Then, the sequence $(S_n)_n$ is **asymptotically linear at** P_{θ_0} if there exists an influence curve $\psi_{\theta_0} \in \Psi_2(\theta_0)$ such that the expansion

$$S_n = \theta_0 + \frac{1}{n}\sum_{i=1}^n \psi_{\theta_0}(x_i) + o_{P^n_{\theta_0}}(n^{-1/2})$$

holds. The family $\Psi_2(\theta_0)$ of influence curves is defined by the set of all maps η_{θ_0} that satisfy the conditions

i) $\eta_{\theta_0} \in L_2^p(P_{\theta_0})$, **ii)** $\mathbb{E}_{\theta_0}[\eta_{\theta_0}] = 0$, **iii)** $\mathbb{E}_{\theta_0}[\eta_{\theta_0}\Lambda^T_{\theta_0}] = I_p$,
where I_p denotes the identity matrix of dimension $p \times p$ and Λ_{θ_0} is the L_2–derivative at P_{θ_0}.

The notation $(S_n)_n$ emphasizes that the number n of observations grows where S_n is the estimator based on n instances. In Def. 4, condition i) is vital for integrability and for the application of the central limit theorem to conclude that S_n is asymptotically normal, i.e.,

$$\sqrt{n}(S_n - \theta_0)\circ P^n_{\theta_0} = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\theta_0}(x_i) + o_{P^n_{\theta_0}}(n^0)\right)\circ P^n_{\theta_0} \xrightarrow{w} \mathcal{N}_p(0, \mathbb{E}_{\theta_0}[\psi_{\theta_0}\psi^T_{\theta_0}]).$$

Condition ii) ensures unbiasedness of the asymptotically linear estimator. The third condition leads to uniform unbiasedness (w.r.t. θ_0), more precisely, if ψ_{θ_0} satisfies i) and ii), (Rieder 1994, Lemma 4.2.18) shows that the condition iii) is equivalent to

$$\sqrt{n}(S_n - \theta_0)(P^n_{\theta_0+t_n/\sqrt{n}}) \xrightarrow{w} \mathcal{N}_p(t, \mathbb{E}_{\theta_0}[\psi_{\theta_0}\psi^T_{\theta_0}])$$

for all $t_n \to t$ where t_n, $t \in \mathbb{R}^p$, so the asymptotic normality granted by the central limit theorem will hold locally uniformly over compacts.

An extension of this concept arises if one wants to estimate a transformed parameter $\tau(\theta)$ leading to so-called ”partial” influence curves (see Def. 6) in the terminology of (Rieder 1994, Def. 4.2.10), (Rieder et al. 2008). One essentially replaces the space $\Psi_2(\theta_0)$ with $\Psi_2^D(\theta_0) = \{D_{\theta_0}\psi_{\theta_0} \mid \psi_{\theta_0} \in \Psi_2(P_{\theta_0})\}$ for $D_{\theta_0} = \partial_{\theta_0}\tau$ (Rieder 1994, Rem. 4.2.11 e)), so the asymptotically linear expansion of transformed estimators in terms of partial influence curves clearly mimicks the traditional delta-method.

Asymptotic linearity has been proven, for example, for asymptotically normal *M*, *R* and *MD* estimators (Rieder 1994, Rem. 4.2.17), so especially for maximum likelihood estimators, quantiles or least squares estimators.

3 Compact differentiability of regularized M-functionals

This is the main part of this paper. We will recapitulate the results on asymptotic linearity of unpenalized M-functionals that will be transferred to the regularized case thereafter.

3.1 Asymptotic linearity of M-estimators

Throughout this subsection, let F_X be a distribution on $(\mathbb{R}^p, \mathbb{B}^p)$ and let $X_1, ..., X_n \overset{i.i.d.}{\sim} F_X$. For some $\Theta \subset \mathbb{R}^p$ (p finite), denote by $\mathcal{C}^p(\Theta)$ the space of all continuous $\mathbb{R}^p$–valued functions on Θ w.r.t. the supremum norm. A general assumption throughout this paper will be

(A0) The parametric model $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ is L_2–differentiable and if ψ_θ is an influence curve, it belongs to the set $\Psi_2(\theta)$.

Let $L : X \times \Theta \to [0, \infty)$ be a loss function and let $\varphi = \partial_\theta L$. Define the function

$$\eta : \Theta \to \mathbb{R}^p, \quad \eta(\theta) := \int \varphi(x, \theta) dF_X(x) = \mathbb{E}_{F_X}[\varphi(X, \theta)]$$

and call its empirical counterpart Z_n. The next two assumptions are

(A1) The parameter space $\Theta \subset \mathbb{R}^p$ is nonempty, compact and equals the topological closure of its interior.

(A2) The function φ satisfies $\varphi(x, \cdot) \in \mathcal{C}^p(\Theta)$ $F_X(dx)$ – a.e., $\varphi_\theta := \varphi(\cdot, \theta) \in L_2^p(F_X)$ $\forall \theta \in \Theta$.

Cor. 1 below (Rieder 1994, Cor. 1.4.5) makes use of the main result from (Jain and Marcus 1975, Thm. 1) which requires

(A3) There exists a pseudo-distance d on Θ such that $d(\theta, \theta_0) \to 0$ as θ converges to θ_0 and such that the metric integral $\int_0^1 \sqrt{H(\epsilon)} d\epsilon$ for the metric entropy H on (Θ, d) is finite

(A4) There exists $M \in L_2(F_X)$ such that $||\varphi(x,\zeta) - \varphi(x,\theta)|| \leq d(\zeta,\theta)M(x) \; \forall \zeta, \theta \in \Theta \; F_X(dx)$–a.e..
and of the following theorem (Rieder 1994, Thm. 1.4.2).

Theorem 2 (**Compact differentiability of M-estimators**) *Assume* $(A1)$, $(A2)$ *and additionally*:

(A5) *There exists a zero* $\theta_0 \in \Theta^\circ$ *of* η *and* $\eta \in \mathcal{C}^p(\Theta)$. *Moreover,* η *is locally homeomorphic at* θ_0 *with bounded and invertible derivative* $d\eta(\theta_0)$.

Then, there exists a neighborhood $V \subset \mathcal{C}^p(\Theta)$ *of* η *and a functional* $T : V \to \Theta$ *satisfying* $f(T(f)) = 0 \; \forall f \in V$. T *is compactly differentiable at* η *with derivative given by* $d_H T(\eta) = -(d\eta(\theta_0))^{-1} \circ \Pi_{\theta_0}$, *where* Π_{θ_0} *is the evaluation functional at* θ_0.

Corollary 1 *Under the assumptions* $(A0)$-$(A5)$, *the sequence* $(S_n)_n := (T \circ Z_n)_n$ *of M-estimators has the asymptotic linear expansion*

$$\sqrt{n}(S_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\theta(x_i) + o_{(F_X^n)_*}(n^0)$$

where the influence function is given by $\psi_\theta(x) := -(d\eta(\theta_0))^{-1} \varphi(x, \theta_0)$. *If the* S_n *is measurable, then asymptotic normality holds, i.e.,* $\sqrt{n}(S_n - \theta_0) \circ F_X^n \xrightarrow{w} \mathcal{N}(0, ACA^T)$ *where* $A := (d\eta(\theta_0))^{-1}$ *and* $C := \mathbb{E}_{F_X}[\varphi_{\theta_0} \varphi_{\theta_0}^T]$.

Remark 1 Note that this result can also be shown by working with bracketing numbers which leads to slightly different assumptions (Van der Vaart 2000, Thm. 5.21).

In the corollary, we denote the stochastic Landau symbol w.r.t. the inner n–fold product measure of F_X by $o_{(F_X^n)_*}$. Inner probabilities are invoked to safeguard against potential nonmeasurability issues of the Z-functions Z_n, see (Rieder 1994). The notation $o_{(F_X^n)_*}(n^0)$ actually means $o_{(F_X^n)_*}(1)$ but the argument n^0 should highlight that n is the growing quantity in the approximation.

Remark 2 (**Fréchet differentiability**) The proof uses an infinite-dimensional version of the delta method, see, e.g., (Van der Vaart and Wellner 2013), (Rieder 1994, Thm. 1.3.3), that requires the chain rule. By Thm. 1, the functionals have to be at least compactly differentiable. Since the chain rule holds for Fréchet differentiable maps, one may ask if the gap between compactly and Fréchet differentiable statistical functionals is considerable. The following two examples give an answer.

Example 1 We refer to (Rieder 1994, Thm. 1.5.1) who shows that for distribution functions F that are continuous in some neighborhood U around $a = F^{-1}(\alpha)$, the location α–quantile is compactly but not boundedly differentiable along $\mathcal{C}(U) \cap \mathbb{D}(\mathbb{R})$, provided that $f(a) > 0$, where $\mathbb{D}(\mathbb{R})$ denotes the Skorohod space, i.e., the space of all real-valued càdlàg functions.

Example 2 Another example is given by the functional $T(F, G) := \int FdG$ for distribution functions F, G. It is shown that this functional is compactly differentiable with Hadamard-derivative

$$d_H T(x, y) = \int xdG - \int ydF,$$

and the empirical version corresponding to the Wilcoxon statistic is compactly differentiable as well (cf. Gill et al. 1989; Van der Vaart and Wellner 2013). This fact has been used to prove asymptotic linearity of the area under the curve (AUC) and the cross-validated AUC as it has been done in (LeDell et al. 2015). However, (Wellner 1992) showed that T is not Fréchet differentiable if one considers the $|| \cdot ||_\infty-$norm. The given counter example relies on the fact that in the case of Fréchet differentiability, the Fréchet derivative $d_F T$ coincides with the Hadamard derivative $d_H T$, so $d_H T$ would be the only candidate for $d_F T$, but $d_H T$ does not supply the $o-$term in the first-order expansion in every case.

So, we can summarize that it is reasonable to show the asymptotic linearity by the milder requirement of compact differentiability.

3.2 The regression context

Fitting a model based on a training set by minimizing some loss function without any restriction generally leads to overfitting, especially in the case of high-dimensional data. This issue has been investigated in (Vapnik 1998) who introduced the structural risk minimization principle which performs the optimization on structures that have finite Vapnik-Chervonenkis dimension. In practice, this idea manifests itself when penalizing the loss function by a regularization term.

In the regression context, we have a regressor matrix with rows $x_i \in \mathcal{X} \subset \mathbb{R}^p$ and a response vector with components $y_i \in \mathcal{Y} \subset \mathbb{R}$. We assume a model $y_i = f(x_i) + \epsilon_i$ where $(x_i, y_i) \sim F$ i.i.d. for a distribution F on $(\mathcal{X} \times \mathcal{Y}, \mathbb{B}(\mathcal{X} \times \mathcal{Y}))$ for the Borel sigma algebra $\mathbb{B}(\mathcal{X} \times \mathcal{Y})$ on $\mathcal{X} \times \mathcal{Y}$ and some error terms $\epsilon_i \sim F_\epsilon$ i.i.d. with mean zero and variance $\sigma^2 \in (0, \infty)$. The function f may be any measurable function mapping from $\mathcal{X}$ into $\mathcal{Y}$. In this work, we assume that f is an element of the parametric function class $\mathcal{F}_\theta := \{f_\theta(x) = x\theta \mid \theta \in \Theta \subset \mathbb{R}^p\}$.

Remark 3 (**Intercept**) Note that unless specified otherwise, the first column of the regressor matrix may only consist of ones, which means that the first component of the parameter is the intercept.

We try to recover the true map f_θ by estimating θ. This is done by defining a loss function $L : (\mathcal{X} \times \mathcal{Y}) \times \Theta \to [0, \infty)$. For practical applications, we will assume that $L((x, y), f_\theta) := L(f_\theta(x), y) = 0$ if $f_\theta(x) = y$ as it was done in Christmann et al. (2009). The penalty term $J_\lambda : \Theta \to [0, \infty)$ that should enforce sparseness of the solution has to satisfy the following conditions:

(A6) The penalty term J_λ with regularization parameter $\lambda > 0$ is non-negative and convex with $J_\lambda(0_p) = 0$ (for $0_p = (0, ..., 0) \in \mathbb{R}^p$).

The assumption that a regularization term must be non-negative is natural. On the other hand, since it penalizes the model complexity, the assumption that $J_\lambda(0_p) = 0$ is reasonable since the parameter $\theta = 0_p$ leads to an empty model which would not make sense to penalize. The convexity assumption is needed for practical applications to prevent the solution from overfitting and, of course, to guarantee the existence of a unique solution in combination with a convex loss function.

Then, we try to solve

$$R(\theta) := \mathbb{E}_F[L((x,y),f_\theta)] + J_\lambda(\theta) = \int_{\mathcal{X}\times\mathcal{Y}} L((x,y),f_\theta)dF(x,y) + J_\lambda(\theta) = \min_{\theta\in\Theta}! \quad (3)$$

by solving the empirical counterpart of the corresponding Z-equation

$$\eta^\lambda(\theta) := \int_{\mathcal{X}\times\mathcal{Y}} \varphi((x,y),\theta)dF(x,y) + J'_\lambda(\theta) \overset{!}{=} 0,$$

provided that $\partial_\theta L = \varphi$ exists and that integration and differentiation can be interchanged. If the penalty term is not of a particular interest, we suppress the superscript and just write Z_n or η.

As for L_2–differentiability of parametric regression models, we refer to (Rieder 1994, Thm. 2.4.7) for random design and to (Rieder 1994, Thm. 2.4.2) for fixed design of the regressor matrix.

3.3 Asymptotic linearity of regularized M-estimators

For clarity, we write down the following counterpart of Cor. 1 to illustrate that we regard the loss function and the penalty term separately, where the latter will be the one that is more likely to cause problems.

Corollary 2 *Assume* (*A*0), (*A*1) *and* (*A*3). *Let the assumptions* (*A*2) *and* (*A*4) *be true for the score function* $\varphi((x,y),\theta) + J'_\lambda(\theta) : \mathbb{R}^p \times \mathbb{R} \times \Theta \to \mathbb{R}^p$ *and let* (*A*5) *be true for* $\eta^\lambda(\theta)$ *provided that the derivative exists. Then, the asymptotic linear expansion in Cor.* 1 *holds with*

$$\psi(x,y) = -\left(d_H\left(\int \varphi((x,y),\theta)dF(x,y) + J'_\lambda(\theta)\right)\Bigg|_{\theta=\theta_0}\right)^{-1}\left[\varphi((x,y),\theta) + J'_\lambda(\theta)\Bigg|_{\theta=\theta_0}\right].$$

Clearly, for example, for the l_1–penalty term, the derivative J'_λ would not be reasonable. Before showing how to remedy this issue, we justify the compactness assumption of the parameter space by coercivity arguments.

3.3.1 Compactness assumption of the parameter space

Lemma 1 *Let $\mathcal{X}$, $\mathcal{Y}$, Θ be real vector spaces. Let $L : \mathcal{X} \times \mathcal{Y} \times \Theta \to [0, \infty)$ be a continuous loss function and let $J_\lambda : \Theta \to [0, \infty)$ be a convex penalty function where $J_\lambda \not\equiv 0$. Let F be a distribution on $(\mathcal{X} \times \mathcal{Y}, \mathbb{B}(\mathcal{X} \times \mathcal{Y}))$. Then, the risk function $\mathbb{E}_F[L((X, Y), \theta)] + J_\lambda(\theta)$ is coercive w.r.t. θ, so the parameter space can be restricted to a compact.*

Proof By convexity, the penalty terms always must satisfy $\lim_{||\theta|| \to \infty}(J_\lambda(\theta)) = \infty$; otherwise, it would have to be constantly zero which we excluded by assumption. The coercivity is inherited from the penalty term since the loss function is convex and by linearity of the integral, its expectation is as well, so the risk is coercive w.r.t. θ. In fact, we get $R(\theta) \to \infty$ for $||\theta|| \to \infty$, so we are allowed to restrict the parameter space to a compact due to Lemma 4. □

This reasoning is of course not new and has been already done to show the existence of solutions for the Huberized lasso (Lambert-Lacroix and Zwald 2011), for regularized kernel methods in Vito et al. (2004) or in (De los Reyes et al. 2016) for regularized functionals in the context of image restoration.

It is easy to see that the usual penalty terms like the l_1-, l_2- or elastic net penalty are coercive (Aravkin et al. 2013, Cor. 8). On the other hand, non-convex penalties do not have to be coercive, for example, the SCAD penalty (cf. Fan and Li 2001) is constant outside a neighborhood of zero whose width depends on the penalty parameter.

In fact, since we are now allowed to assume compactness of the parameter space, we face another potential issue. The compactness assumption leads to the problem that the M-estimator $\hat{\theta}_n$ may be located at the boundary of Θ. We invoke the idea of one-step estimators from Van der Vaart (2000) to make the connection with machine learning algorithms.

Having a $\sqrt{n}$–consistent preliminary solution $\tilde{\theta}_n$ of the estimating equation $Z_n(\theta) = 0$, then an application of the Newton-Raphson algorithm leads to an improved one-step solution

$$\hat{\theta}_n := \tilde{\theta}_n - (Z'_{n,0}(\tilde{\theta}_n))^{-1} Z_n(\tilde{\theta}_n).$$

$Z'_{n,0}$ is a regular matrix and converges in probability to a regular matrix Z'_0. The following theorem can be found in (Van der Vaart 2000, Thm. 5.45).

Theorem 3 *Let the notation be as above. Let the condition that for every constant M it holds that*

$$\sup_{\sqrt{n}||\theta - \theta_0|| < M} \left(||\sqrt{n}(Z_n(\theta) - Z_n(\theta_0)) - Z'_0 \sqrt{n}(\theta - \theta_0)|| \right) \xrightarrow{P} 0 \tag{4}$$

be satisfied for a regular matrix Z_0'. If it holds additionally that $\sqrt{n}(Z_n(\theta_0))$ converges to some limit, then the one-step estimator $\hat{\theta}_n$ is already $\sqrt{n}$–consistent.

Lemma 2 *Let all the notation be as above. Under* (*A*2), (*A*5) *and the additional assumptions*

(A7) *The learning procedure is $\sqrt{n}$–consistent,*

(A8a) *The function Z_n is twice differentiable w.r.t. θ,*

the one-step estimator is not located at the boundary of the parameter space.

Proof Since condition (4) is weaker than differentiability of Z_n at θ, this part is already satisfied by (A8a). The only condition of Thm. 3 that remains to be proven is the convergence of $\sqrt{n}(Z_n(\theta_0))$ to some limit Z. But since we already know by (A7) that the learning algorithm is $\sqrt{n}$–consistent, hence $\sqrt{n}(\theta_0 - \hat{\theta}_n)$ converges in probability. An application of the delta method which is possible under (A8a) provides that $\sqrt{n}(Z_n(\theta_0) - Z_n(\hat{\theta}_n))$ has a limiting distribution (which is the Dirac measure at zero) and we note that by definition, it holds that $Z_n(\hat{\theta}_n) = 0$, so the convergence of $\sqrt{n}Z_n(\theta_0)$ has been established and Thm. 3 applies. □

We admit that it is not common to assume learning rates as we did in assumption (A7), but it is more convenient just to assume consistency. Since functions that are too complex may not be able to be approximated with a predetermined rate, this assumption results in the class of approximable functions getting strictly smaller.

3.3.2 Twice differentiable Z-function

We state the following intermediate result:

Theorem 4 *Under the conditions* (*A*0), (*A*1), (*A*6), (*A*7), (*A*8*a*) *and*

(A2') $\varphi^\lambda(\cdot, \theta) \in L_2^p(F)\ \forall \theta \in \Theta$,

(A5') $\eta^\lambda(\theta_0) = 0$ *for a* $\theta_0 \in \Theta^\circ$ *and* η^λ *are locally homeomorphic at* θ_0 *with bounded and invertible derivative* $d\eta^\lambda(\theta_0)$,

the sequence $(S_n^\lambda)_n := (T \circ Z_n^\lambda)_n$ of regularized M-estimators is asymptotically linear.

Proof As a byproduct of (A1), we immediately get (A3). This is true since $\mathbb{R}^p$ is a normed space, hence the pseudo-distance is just the standard euclidean norm on $\mathbb{R}^p$ and by boundedness of Θ and since p is finite, we can conclude that the metric integral is finite. Even more general, the metric integral is finite provided that $d(\theta, \theta_0) = ||\theta - \theta_0||_2^\delta$ for some $\delta \in (0, \infty)$ (Rieder 1994, Rem. 1.4.6.b)).

From twice differentiability of η^λ, the first part of (A2) is trivially satisfied and the derivative φ of L is continuous w.r.t. θ. By (A6), (A7), (A8a) and Lemma 2, the assumption that we can restrict the parameter space Θ to a compact set is justifiable. Using this compactness of Θ, we deduce that there exists an $M(x) \in L_2(F_X)$ such that $||\varphi(x,\xi) - \varphi(x,\theta)|| \leq M(x)d(\xi,\theta)$, so we conclude that (A4) holds.

Thus, Cor. 1 is applicable and we get the desired result. □

3.3.3 Twice continuously differentiable loss function, non-differentiable penalty term

If the penalty term is non-differentiable, like the Lasso loss, then we invoke an approximation result of (Avella-Medina 2017) which uses a maximum theorem of (Berge 1963). This result is exactly what we need in the presence of non-differentiable regularization terms since despite that we cannot assume differentiability, we can at least assume continuity. The following lemma is a combination of (Avella-Medina 2017, Lemma 2) and (Avella-Medina 2017, Prop. 1).

Lemma 3 (**Approximating influence curves**) *Assume that the parameter space $\Theta \subset \mathbb{R}^p$ is compact and that the loss function is twice continuously differentiable w.r.t. θ. If there exists a sequence $(J_\lambda^m)_m$ with $J_\lambda^m \in \mathcal{C}^\infty(\Theta)$ that converges to J_λ in the Sobolev space $W^{2,2}(\Theta)$, i.e.,*

$$||J_\lambda^m - J_\lambda||_{W^{2,2}} = \left(\sum_{|\alpha|\leq 2} \int_\Theta |\partial^\alpha (J_\lambda^m(\theta) - J_\lambda(\theta))|^2 d\theta \right)^{1/2} \longrightarrow 0,$$

then $\lim_m(T_m) = T$ where T_m denotes the M-functional that intends to find the zero of the Z-equation corresponding to R_m where R_m denotes the risk function where J_λ is replaced by J_λ^m, cf. Eq. 3. If ψ_θ^m denotes the influence curve corresponding to the functional T_m, the limiting influence curve equals the one for the functional T. Both the limiting behaviour of T_m and ψ_θ^m do not depend on the particular choice of the approximating sequence J_λ^m.

Theorem 5 *Assume that there exists a sequence $(J_\lambda^m)_m$ with $J_\lambda^m \in \mathcal{C}^\infty(\Theta)$ of regularization functions that converge to J_λ in the Sobolev space $W^{2,2}(\Theta)$. Under the conditions* (A0), (A1), (A2'), (A5'), (A6), (A7) *and*

(A8b) *the loss function L is convex and twice continuously differentiable w.r.t. θ,*

the sequence $(S_n^\lambda)_n$ of regularized M-estimators is asymptotically linear.

Proof From Thm. 4, we can conclude that the estimator has an asymptotic linear expansion and that it is asymptotically normal if the respective assumptions collected there are satisfied. But since this is just an asymptotic property up to a remainder term of order $n^{-1/2}$, it suffices to approximate J_λ by J_λ^m such that the difference in the respective influence functions is negligible, i.e., the difference is already of order $n^{-1/2}$. Note that by continuity of the Gâteaux derivative w.r.t. the direction and by Lemma 3, it holds that $\lim_m(IC(x, T_m, P)) = IC(x, T, P)$.

Finally, we can conclude that we can work with infinitely differentiable penalty terms satisfying the conditions of the previous subsection but that this results in the same asymptotic linear expansion as if we used the true non-differentiable penalty term. Thus, we only need the existence of an approximating sequence of penalty terms. □

Remark 4 (Öllerer et al. 2015, Lemma 5.4) showed for the Lasso and a concrete hyperbolic tangent approximation of the penalty term that the influence function of the approximating estimator derived by (Öllerer et al. 2015, Prop. 4.1) converges to the influence function of the Lasso. So, (Avella-Medina 2017) generalized their result with Lemma 3 for any losses and penalties satisfying the given conditions.

Note again that the main difficulty for non-differentiable regularization terms is the translation of M- to Z-equations. The approximation result elegantly avoids a tedious case-by-case study under which conditions an M-estimator w.r.t. a certain regularized loss function can be written as Z-estimator and provides a universal result.

3.4 Extension to ranking

So far, we concentrated on regression problems where a data set $(x_1, y_1), \ldots, (x_n, y_n)$ is given with the goal to compute a model $f_{\hat{\theta}}$ such that the y_i are fitted by $f_{\hat{\theta}}(x_i)$. Ranking problems are different since they only intend to recover the true ordering of the responses which clearly does not require an exact recovery of the response values themselves. In the ranking setting, we assume the same underlying model as in the first part of this section, with the only difference that we have to invoke a joint distribution $F_r : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$ on the measurable space $((\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}), \mathbb{B}((\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})))$ where the notation F_r is introduced to distinguish it from the joint distribution in the previous part of this section.

In contrast to prediction problems, it is not the goal to recover the true values of the y_i but just to predict their true order which implies that loss functions that are based on the residuals $y_i - f_{\hat{\theta}}(x_i)$ are not appropriate. Therefore, the ranking model can be fitted by minimizing a ranking loss function $L^r : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \times \Theta \to [0, \infty)$ which quantifies a ranking loss, that is a loss suffered from a misranking of a pair of instances, see Clémençon et al. (2008). Defining a penalty term and the ranking risk analogously to the risk R in Eq. 3, the corresponding Z-equation resulting from the problem to minimize the regularized risk is

$$Z_n^{r,\lambda}(\theta) := \frac{1}{n(n-1)} \sum\sum_{i \neq j} \varphi^r(((x_i, y_i), (x_j, y_j)), \theta) + J_\lambda'(\theta) \overset{!}{=} 0$$

where $\varphi^r = \partial_\theta L^r$ is the score function, hence the first term of $Z_n^{r,\lambda}$ is the empirical counterpart of the expected score.

We can easily adapt our results to the ranking setting and conclude compact differentiability of regularized ranking functionals and asymptotic linearity of the sequence $(S_n^{r,\lambda})_n := (T \circ Z_n^{r,\lambda})_n$ of regularized ranking M-estimators.

Theorem 6 *Define*

$$\eta^{r,\lambda} : \Theta \to \mathbb{R}^p, \quad \eta^{r,\lambda}(\theta) := \int \varphi^r(((x,y),(x',y')),\theta)dF^r(((x,y),(x',y'))) + J'_\lambda(\theta).$$

Then, under the conditions $(A0)$, $(A1)$, $(A5)$, $(A6)$, $(A7)$, $(A8a)$ *and*

((A2r)') $\varphi^r(\cdot,\theta) + J'_\lambda(\theta) \in L_2^p(F^r)\ \forall \theta \in \Theta$,

the sequence $(S_n^{r,\lambda})_n := (T \circ Z_n^{r,\lambda})_n$ *of regularized ranking M-estimators is asymptotically linear.*

Proof This directly follows from Thm. 4 since the optimization is done w.r.t. θ, whereas the dimension of the space of the other arguments of φ is not explicitly used in the proof. □

Similarly, an analogue to Thm. 5 holds for the ranking setting.

Remark 5 It is important to emphasize that ranking loss functions like the hard ranking loss (Clémençon et al. 2008, Sec. 2) and related losses are not continuous and not convex, so they fail the assumptions of these theorems (however, the hard ranking loss is bounded, so combining it with a suitable regularity term again leads to a coercive target function). Examples for which the regularity conditions hold are smooth convex surrogates of those ranking losses, see (Clémençon et al. 2013) for an overview.

4 Examples for asymptotically linear estimators in machine learning

The conditions for asymptotic linearity of the regularized M-estimators in the previous section are quite general. The goal of this section is to provide examples for machine learning algorithms to which the derived results can be applied and to specify the required conditions for each procedure.

4.1 Lasso

Lasso regression (cf. (Bühlmann and Van De Geer 2011)) is an l_1–penalized least squares regression, i.e.,

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta \in \Theta}\left(\frac{1}{n}||Y - X\beta||_2^2 + \lambda||\beta||_1\right)$$

where we denote the regressor matrix by $X \in \mathbb{R}^{n \times p}$ and the response vector by $Y \in \mathbb{R}^n$. The lasso regression results in a shrinkage of the coefficients and in sparsity of the fitted model. The score function for the non-regularized loss is given by

$$\varphi(\cdot, \beta) = \frac{2}{n} X^T (Y - X\beta). \tag{5}$$

We invoke the approximation of the non-differentiable penalty term. There exists an example of such a smooth penalty term converging to the absolute value in Avella-Medina (2017).

Theorem 7 *Assume* $(A0)$, $(A1)$ *and*

((A2^{Lasso}**)')** *the ideal distribution F has finite fourth moments,*

(A5') *the true solution* β^0 *lies in the interior of* Θ *and the derivative* $d\eta^\lambda(\beta^0)$ *is invertible,*

(A7^{Lasso}**)** $||\beta^0||_1 = o(\sqrt{n/\ln(p)})$ *and that the regularization parameter in dependence of n is chosen in the range of* $\lambda_n = o(\sqrt{\ln(p)/n})$.

Then, the sequence $(S_n^{Lasso})_n := (T \circ Z_n^{Lasso})_n$ *of Lasso estimators is asymptotically linear.*

Proof We need to verify the conditions of Thm. 5. Consider a smooth approximation J_λ^m of the absolute value in the sense of the Sobolev space $W^{2,2}$, as given in Öllerer et al. (2015) or Avella-Medina (2017), respectively. Then, we set

$$\tilde{J}_\lambda^m(\beta) := \sum_i J_\lambda^m(\beta_i) \longrightarrow \sum_i |\beta_i| = ||\beta||_1,$$

and thus

$$\nabla_\beta \tilde{J}_\lambda^m(\beta) = (\partial_{\beta_1} \tilde{J}_\lambda^m(\beta), ..., \partial_{\beta_p} \tilde{J}_\lambda^m(\beta)) \longrightarrow (\mathrm{sign}(\beta_1), ..., \mathrm{sign}(\beta_p)) = \nabla_\beta ||\beta||_1$$

and the (component-wise) convergence of the Hessian holds as well due to the properties of $W^{2,2}$. For this idea, we refer to (Öllerer et al. 2015, Lemma 5.4). The loss function and the approximating penalty term are smooth, hence (A8b) is satisfied and Lemma 3 is applicable.

The target function is coercive w.r.t. β (see Lemma 4). This holds because as $||\beta|| \to \infty$, the penalty will tend to infinity and so does the target function. Note that this does not hold for the loss function itself since $||\beta|| \to \infty$ can result in a small loss. One may argue that even in the penalized case, it can happen that $||(x, y, \beta)|| \to \infty$ without resulting in the target function growing as well. If, for example, $y = 0$ and $\|x\|$ is large, then $y = x\beta$ for $\beta = 0_p$. But in this case, we do not lose anything if we restrict the parameter space. Furthermore, we can write the optimization problem in the form

$$\min(||Y - X\beta||_2^2/n) \quad s.t. \quad ||\beta||_1 \leq c_\lambda$$

for some constant c_λ depending on λ. So, we have a convex optimization problem with a continuous, strictly convex and coercive target function, so by Werner (2006), there definitely exists a solution β^0 of η^λ and the local homeomorphicity around the solution follows.

Combining ((A2^{Lasso})') with Eq. (5), we derive that the score function is square-integrable w.r.t. the distribution F. Then, (A2') is satisfied and by boundedness of the integral by the previous assumption and by compactness of the parameter space, this derivative is bounded.

The Lasso is generally inconsistent, but under (A7^{Lasso}), it follows from Bühlmann and Van De Geer (2011) that the Lasso is $\sqrt{n}$–consistent in this case. Note that despite we solve a convex optimization problem assuming that the true solution is already located in the interior of Θ, that does not suffice to guarantee that the computed solution does not lie on the boundary of the parameter space. Finally, Thm. 5 applies and the assertion is proven. □

4.2 Elastic net

The elastic net (cf. Zou and Hastie 2005) can be regarded as a compromise between Lasso and Ridge regression. Given two penalty parameters λ_1, λ_2, the elastic net solution is given by

$$\hat{\beta}^{EN} = \text{argmin}_\beta \left(\frac{1}{n} ||Y - X\beta||_2^2 + \lambda_1 ||\beta_1||_1 + \lambda_2 ||\beta||_2^2 \right)$$

and by defining $\alpha := \frac{\lambda_2}{\lambda_1 + \lambda_2}$, this can be rewritten as a convex combination of l_1- and l_2- penalties, i.e.,

$$\hat{\beta}^{EN} = \text{argmin}_\beta \left(\frac{1}{n} ||Y - X\beta||_2^2 + (1 - \alpha) ||\beta_1||_1 + \alpha ||\beta||_2^2 \right)$$

where $(1 - \alpha)||\beta||_1 + \alpha||\beta||_2^2$ is referred to as the elastic net penalty.

Corollary 3 *Under the assumptions of Thm. 7, the sequence* $(S_n^{orthEN})_n := (T \circ Z_n^{orthEN})_n$ *of elastic net estimators with orthonormal design is asymptotically linear.*

Proof Note that for orthonormal design, the EN solution is just a rescaled Lasso solution with factor $\frac{1}{1+\lambda_2}$. In this case, we can simply rescale the influence function derived in Öllerer et al. (2015), proving the result. □

Corollary 4 *Under the assumptions of Thm. 7, the sequence* $(S_n^{EN})_n := (T \circ Z_n^{EN})_n$ *of elastic net estimators is asymptotically linear.*

Proof It is shown in Zou and Hastie (2005) that the elastic net can be rewritten as a special Lasso with the augmented data

$$X^* := \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I_p \end{pmatrix}, \quad y^* := \begin{pmatrix} y \\ 0_p \end{pmatrix}$$

and the penalty $\gamma := \frac{\lambda_1}{\sqrt{1+\lambda_2}}$. If $\hat{\beta}^{Lasso}$ is the respective Lasso solution, the elastic net solution is a rescaling with factor $\frac{1}{1+\lambda_2}$ as before. □

Using these results and the idea of Öllerer et al. (2015), the respective influence curve of the elastic net can be computed by just adapting the influence curve for the Lasso computed in Öllerer et al. (2015) and Thm. 7.

4.3 Adaptive Lasso

The adaptive Lasso (cf. Zou 2006) is a two-stage estimator that first computes the standard Lasso estimator (or any $\sqrt{n}$–consistent estimator), denoted by $\hat{\beta}^{init}$, and then in a second step, one minimizes

$$\frac{1}{n}||Y - X\beta||_2^2 + \lambda \sum_j \frac{|\beta_j|}{|\hat{\beta}_j^{init}|}.$$

Borrowing the consistency requirements for the adaptive Lasso from Zou (2006), we have the following result.

Theorem 8 *Assume* $(A0)$, $(A1)$ *and*

((**A2**^{Lasso})') *the ideal distribution F has finite fourth moments*,

(A5') *the true solution β^0 lies in the interior of Θ and the derivative $d\eta^\lambda(\beta^0)$ is invertible*,

(**A7**^{ALasso}) *The regularization parameter in dependence of n satisfies $\lambda_n = o(\sqrt{n})$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$ for $\gamma > 0$.*

Then, the sequence $(S_n^{adapt})_n := (T \circ Z_n^{adapt})_n$ of adaptive Lasso estimators is asymptotically linear and the influence curve of the j–th component of $\hat{\beta}^{adapt}$ is given by

$$IC((x_0, y_0), \hat{\beta}_j^{adapt}(\lambda), P_{\beta^0}) = \begin{cases} 0, & \hat{\beta}_j^{init}(\lambda) = 0 \\ 0, & \hat{\beta}_j^{adapt}(\lambda) = 0 \\ IC((x_0, y_0), \hat{\beta}_j^{Lasso}(\lambda/|\hat{\beta}_j^{init}(\lambda)|)), & \text{otherwise} \end{cases}.$$

where we denote by $\hat{\beta}^{Lasso}(\lambda)$ the Lasso estimator using the penalty parameter λ.

Proof Obviously, if $\hat{\beta}_j^{init} = 0$, we immediately know that $\hat{\beta}_j^{adapt} = 0$. Hence, if we have the initial solution, we can rewrite the adaptive Lasso optimization problem as

$$\hat{\beta}^{adapt}_{\hat{S}^{init}_{\lambda}} = \text{argmin}_{\beta_{\hat{S}^{init}(\lambda)}} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \hat{S}^{init}(\lambda)} (y_i - x_{ij}^T \beta_j)^2 + \lambda \sum_{j \in \hat{S}^{init}(\lambda)} \frac{|\beta_j|}{|\hat{\beta}_j^{init}|} \right)$$

where $\hat{S}^{init}(\lambda) := \{j \mid \hat{\beta}_j^{init} \neq 0\}$. Then, this optimization problem is just a Lasso optimization problem with a weighted penalty term which can be approximated coordinate-wisely in the spirit of Avella-Medina (2017).

The corresponding influence function has implicitly been derived in (Avella-Medina 2017). Note that by our method, we would only derive $|\hat{S}^{init}(\lambda)|$ components of the influence function. However, it was proven in (Öllerer et al. 2015) that the components of the influence function corresponding to the coefficients that are excluded from the model are zero, i.e., if the Lasso in the first step already sets some coefficients to zero, the final coefficients will be zero, so we can just plug in zeroes into the respective components of the influence function, providing the usual asymptotic linear expansion.

Since the first step does not compute the final non-zero coefficients but just regularizing weights, its influence implicitly arises in this expansion as a factor, leading to the stated result. □

Remark 6 (**Partial influence curves**) Note that the influence curves derived in Thm. 8 correspond to the concept of "partial" influence functions (see Def. 6). This is true since in the proof of Thm. 8, we are implicitly using the smooth transformation $\beta \mapsto \beta_{\hat{S}^{init}_{\lambda}}$ to derive the components of the influence curve corresponding to the coefficients that not already have been excluded from the model in the initialization step. In other words (after suitable renumeration of the columns), we get the matrix $D_{\beta_{\hat{S}^{init}(\lambda)}} := (\text{diag}(1, \hat{s}^{init}), 0_{p-\hat{s}^{init}}) \in \mathbb{R}^{\hat{s}^{init} x p}$ where $\hat{s}^{init} := |\hat{S}^{init}(\lambda)|$.

4.4 Relations to other work

Van de Geer 2016 derived a result on asymptotic linearity of the de-sparsified graphical Lasso estimator. Note that according to Banerjee et al. (2008) (see also Friedman et al. 2008), the loss function that is optimized in the case of the standard graphical Lasso is convex and with the convex penalty, we are still in the case of convex optimization problems. However, the main difference between this work and our theory is that the estimator is matrix-valued and not vector-valued since it is a covariance and no regression estimator.

Van de Geer (2016) defines asymptotic linearity in the case of matrix-valued estimators $\hat{M}_n$ in the sense that the difference $\hat{M}_n - M_0$ is linear, up to some remainder term where M_0 is the truth. Our framework required that the remainder term is of order $n^{-1/2}$ which is generalized in Van de Geer (2016) to the condition that the infinity norm of the remainder term is of order $n^{-1/2}$. Since influence curves for such estimators are very complicated, Van de Geer (2016) use regularity arguments to derive their results in order to avoid this issue.

Note that in an extended understanding of sparsity, one indeed could use estimators like the graphical Lasso to derive regression estimators. Since the linear regression estimator is given by $Cov(X,X)^{-1}Cov(X,Y)$, one may use the graphical Lasso or related methods to derive estimators which are ”sparse” in the sense that many cells of the covariance matrices are set to zero. However, this would not lead to sparse regression estimators as in our sense.

Van de Geer (2014) show that the oracle requirements of the Lasso that concern the penalty term actually only use the facts that the l_1–penalty is weakly decomposable, i.e., $||\theta||_1 = ||\theta_S||_1 + ||\theta_{S^c}||_1$ for $S \subset \{1, ..., p\}$, and that it satisfies the triangle inequality and the dual norm equality. They therefore suggest other penalty terms that satisfy these conditions. As long as the penalty term is sufficiently regular or a sufficiently regular approximation exists, we can transfer our results on asymptotic linearity of the corresponding estimator to the respective case.

Remark 7 So far, our results cover the case of convex penalty terms. However, we do not explicitly need this requirement except for justifying the assumption that the solution lies in a suitable compact so that the parameter space can be reduced to this compact. We decided to formulate the theoretical results for this special case since the most popular regularized M-estimators like the Lasso (which additionally has a convex loss function) can be embedded into this framework, as we have shown in Thm's 7 and 8. For the robustness problems arising when optimizing convex target functions, see Sect. 6.

However, sophisticated results from Loh and Wainwright (2015), Lee (2015) or Negahban et al. (2012) on local optima and consistency results for regularized M-estimators in the nonconvex setting can be combined with our work since our theorems require the assumption of the existence of a suitable solution and consistency. It would be beyond the scope of this work to gather all existing results that provide the correctness of several of our assumptions in the theorems, so we restrict ourselves to the presented results.

Remark 8 (**Asymptotic normality**) Note that the additional assumption of measurability of the sequences of estimators provides asymptotic normality of the estimating sequence due to Cor. 1. Of course, we are not the first ones with results on asymptotic normality. See, for example, (Loh 2017, Cor. 1) showing under which conditions regularized M-estimators of a very general form, including the Lasso, are asymptotically normal.

Remark 9 We point out that the asymptotic normality that is established by the asymptotic linear expansion does not contradict the statements in Pötscher and Leeb (2009) and Pötscher and Schneider (2009) who derive that there does not exist a uniformly consistent estimator for the distribution of the Lasso resp. the adaptive Lasso estimator. However, our assumption (A7^{Lasso}) resp. (A7^{ALasso}) allows for the application of (Pötscher and Leeb 2009, Thm. 5) resp. (Pötscher and Schneider 2009, Thm. 4) that provides the asymptotic distribution of the Lasso resp. the adaptive Lasso estimator and shows that the convergence holds with rate $n^{-1/2}$, i.e., the error is already captured by our remainder term of order $n^{-1/2}$ in the asymptotic linear expansion. Since our assumptions correspond to the case $e = 0$ in Pötscher and Leeb (2009) resp. $m = 0$ in Pötscher and Schneider (2009), the results on nonuniformity from (Pötscher and Leeb 2009, Thm. 13) resp. (Pötscher and Schneider 2009, Thm. 12) are not valid.

5 Data-driven penalty parameters

It is common that asymptotic results for regularized methods allow for the case that the regularization parameter is data-driven which manifests itself in a sequence $(\lambda_n)_n$ of regulariztion parameters. The same is true for Boosting where the amount of regularization does not depend on a penalty parameter but implicitly on the number of iterations such that a diverging sequence of iterations is the analogue to a sequence $(\lambda_n)_n$ with $\lambda_n \to 0$ for $n \to \infty$.

To keep the asymptotic results valid uniformly for $n \to \infty$, results for Lasso methods as in Bühlmann and Van De Geer (2011) or Zou (2006) and for boosting methods as, for example, in Bühlmann (2006) require penalty parameters which fall into a suitable range in dependence of n or numbers of iterations that grow sufficiently slowly w.r.t. n.

As for our results, we would need a suitable degree of approximation, i.e., a suitable sequence $(m_n)_n$, leading to a sequence of regularization terms of the form $(J_{\lambda_n}^{m_n})_n$, to get similar statements.

In fact, we already used sequences $(m_n)_n$ implicitly when proving Thm. 5. Our argument was to set m_n sufficiently large to get a degree of approximation that leads to an error term which is already absorbed by the remainder term.

If we are concerned about sequences of penalty terms, we essentially need to have a sequence $(m_n)_n$ which again grows sufficiently fast to keep the error term small enough. Since we assumed that J_λ is approximable by a sequence of smooth penalty terms J_λ^m and since λ usually just enters as a factor, a diminishing sequence of regularization terms still keeps the approximability valid since for smaller penalty parameters, the regularization term gets ”less wiggly”, so we assume that for fixed n, one would generally need a smaller number m for a smaller λ than for a large λ.

A general approximation to the best of our knowledge is out of reach, however, for a given penalty term with a given approximation sequence, one could derive conditions for the sequence $(m_n)_n$ according to the sequence of regularization terms. In particular, this holds for the Lasso and the Adaptive Lasso.

6 Practical implications

An easily seen advantage of asymptotically linear estimators is the opportunity to perform fast updates once new observations are available. Clearly, once an initial estimator is available, one just has to evaluate the influence curve in the observations to construct a new estimator. In general, estimators of the kind

$$S_n^1 := \hat{\theta}_n + \frac{1}{n}\sum_{i=1}^{n} \psi_{\hat{\theta}_n}(x_i)$$

are known as one-step estimators (see, p.e. Rieder 1994 for more details). If model selection already has been done, the influence curves just need to be localized to the respective components, leading to partial influence curves as detailed out in Werner (2019). The practical use of such one-step estimators has been shown extensively in Kohl (2005).

We provide the solid theoretical foundation that allows the influence curves that already have been computed, for example, by Öllerer et al. (2015) or Avella-Medina (2017), to enter model diagnosis which was not evident before since the asymptotically linear expansion in terms of influence curves was not proven. First, the asymptotic normality provides asymptotic covariances that allow standard confidence intervals for the estimators. Moreover, the influence curves indicate which observations would have a high impact on the estimator. These model diagnostic tools are especially important when facing black box algorithms to compute the regularized M-estimators.

The motivating learning algorithms, i.e., the Lasso or the adaptive Lasso, have a disadvantage which becomes problematic on real data. Due to the convex loss function, the corresponding estimators are not robust which indicates that outliers, both in the regressor matrix and in the response vector, can significantly distort them. It has been shown in (Alfons et al. 2013, Thm. 1) that the so-called breakdown point of the Lasso is n^{-1}, so even a single outlier in the response vector suffices to let the estimator become unreliable. This obviously also holds for the adaptive Lasso or the elastic net.

There already exist robust variants of the Lasso like the Huberized Lasso, originally introduced in Rosset and Zhu (2007) and studied in Chen et al. (2010b) and Chen et al. (2010a) which replaces the quadratic loss by the Huber loss. However, Chang et al. (2018) showed that despite the Huberized Lasso is robust against outliers in the response vector, it is not robust against outliers in the regressor matrix. Chang et al. (2018) themselves suggest the Tukey Lasso which is actually a robust variant of the adaptive Lasso where in the first step, a robust MM-estimator (see Maronna et al. (2006)) is computed which is used for the adaptive penalty in the second step where the penalized biweight function (Maronna et al. 2006, Sec. 2.2.4) is minimized. This estimator is also robust against outliers in the regressor matrix.

More formally, one can consider outliers as ”contamination” of the assumed model, the so-called ”ideal model”, see Rieder (1994) for more details. This issue is especially relevant for high-dimensional data since every cell in the regressor matrix can potentially be contaminated which makes it very probable that a data matrix has contaminated cells which has been pointed out in Alqallaf et al. (2009).

These issues can be handled by using suitable loss functions, especially those whose derivatives $\partial_\theta L(x, \theta) = \varphi(x, \theta)$ tend to zero for large $|x|$, so-called ”redescenders” (see Maronna et al. 2006). However, it is even more sophisticated to robustify the influence curves themselves which requires that the original non-robust estimator can be expressed in terms of influence curves, so which directly uses our results about asymptotic linearity. The robustification of the influence curve is done by solving a specified optimization problem like minimizing the supremal absolute value of the influence curve, minimizing the covariance of the influence curve subject to a bias bound or minimizing the maximal MSE. These problems are known as MBRE, OBRE and OMSE problem, respectively, and have been studied in detail in Rieder (1994) and Kohl (2005) for different contamination models.

It is important to point out that just replacing the original loss function by a ”robust counterpart” does not provide any optimality statements which indeed hold for estimators based on optimally robust influence curves which use the original loss function and suitably bound the corresponding influence curve. As consequence, we can directly work with the influence curves derived, for example, by Öllerer et al. (2015). Although we did not provide a construction principle for the sequence $(m_n)_n$ in the previous section, thanks to the approximation principle of the penalty term we can directly use the influence function corresponding to $J_{\lambda_n}^{\infty}$ for a given λ_n.

It would be desirable to extend those results to the case of high-dimensional regularized regression or estimation. The first step has been done with the theory on asymptotic linearity of the regularized M-estimators in this work. The next step for future work is to develop an algorithm that computes robust variants of the influence curves corresponding to regularized M-estimators for regression.

7 Conclusion

We studied under which conditions a regularized M-estimator allows for an asymptotically linear expansion. We provided a general theory for the asymptotically linear expansion of such estimators and gave concrete examples of machine learning algorithms which our theory covers. Of course, from the asymptotic linear expansion, the asymptotic normality can be directly derived which allows for standard confidence intervals for the estimators.

Influence curves for a wide range of estimators have already been computed in the literature as the Gâteaux derivative of the corresponding statistical functional, but this does not guarantee that the remainder term in the asymptotically linear expansion vanishes over suitable (i.e., compact) tangent sets. We closed this gap by proving the compact differentiability of the regularized M-functionals, leading to the desired uniformity and therefore to the validity of the influence curves.

However, we concentrated on linear regression models in this work. An extension to other areas of machine learning will be a subject of future work.

Miscellaneous

The L_2–differentiability originally comes from LeCam (1970).

Definition 5 Let $\mathcal{P} := \{P_\theta \mid \theta \in \Theta\}$ be a family of probability measures on some measurable space $(\Omega, \mathcal{A})$ and let Θ be a subset of $\mathbb{R}^p$. Then, $\mathcal{P}$ is **L_2–differentiable at** θ_0 if there exists $\Lambda_{\theta_0} \in L_2^p(P_{\theta_0})$ such that

$$\left|\left|\sqrt{dP_{\theta_0+h}} - \sqrt{dP_{\theta_0}}\left(1 + \frac{1}{2}h^T \Lambda_{\theta_0}\right)\right|\right|_{L_2} = o(||h||)$$

for $||h|| \to 0$. In this case, the function Λ_{θ_0} is the L_2–derivative and $I_{\theta_0} := \mathbb{E}_{\theta_0}[\Lambda_{\theta_0}\Lambda_{\theta_0}^T]$ is the Fisher information of $\mathcal{P}$ at θ_0.

Note that the L_2–differentiability is a special case of the wider concept of L_r–differentiability (cf. Rieder and Ruckdeschel 2001). The L_2–differentiability holds for many distribution families, including normal location and scale families, Poisson families, gamma families, and even for ARMA, ARCH and GPD families (Rieder et al. 2008, Pupashenko et al. 2015). A standard example of a distribution family that is not L_2–differentiable is the model $\mathcal{P} := \{U([0, \theta]) \mid \theta \in \Theta\}$.

The following definition of partial influence curves and the corresponding asymptotically linear expansion in terms of such partial influence curves is borrowed from (Rieder 1994, Def. 4.2.10) and Rieder et al. (2008).

Definition 6 Let $(\Omega^n, \mathcal{A}^n)$ be a measurable space and let $S_n : (\Omega^n, \mathcal{A}^n) \to (\mathbb{R}^q, \mathbb{B}^q)$ be an estimator for the transformed quantity of interest $\tau(\theta)$. Assume that $\tau : \Theta \to \mathbb{R}^q$ is differentiable at $\theta_0 \in \Theta$ where $\Theta \subset \mathbb{R}^p$ and $q \leq p$. Denote the Jacobian by $\partial_{\theta_0}\tau =: D_{\theta_0} \in \mathbb{R}^{q \times p}$. Then, the set of **partial influence curves** is defined by

$$\Psi_2^D(\theta_0) := \{\eta_{\theta_0} \in L_2^q(P_{\theta_0}) \mid \mathbb{E}_{\theta_0}[\eta_{\theta_0}] = 0,\ \mathbb{E}_{\theta_0}[\eta_{\theta_0}\Lambda_{\theta_0}^T] = D_{\theta_0}\}.$$

The sequence $(S_n)_n$ is asymptotically linear at P_{θ_0} if there exists a partial influence curve $\eta_{\theta_0} \in \Psi_2^D(\theta_0)$ such that the expansion

$$S_n = \tau(\theta_0) + \frac{1}{n}\sum_{i=1}^n \eta_{\theta_0}(x_i) + o_{P_{\theta_0}^n}(n^{-1/2})$$

is valid.

For the following lemma, we refer to Evgrafov and Patriksson (2003) and Levitin and Tichatschke (1998).

Lemma 4 *Let $f : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}$ be continuous, where $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} \subset \mathbb{R}^m$, $\Theta \subset \mathbb{R}^k$. Define $\Xi(x, y) := \operatorname{argmin}_\theta(f(x, y, \theta))$. If f is* **coercive w.r.t.** *θ, i.e., the sets $\{\theta \in \Theta \mid f(x, y, \theta) \leq c\}$ are bounded for all $c \in \mathbb{R}$ for every $x \in \mathcal{X}$, $y \in \mathcal{Y}$, then $\min_\theta(f(x, y, \theta)) > -\infty$ and $\Xi(x, y)$ is nonempty and compact for any x, y.*

Acknowledgements The results presented in this paper are part of the author's PhD thesis (Werner (2019)) supervised by P. Ruckdeschel at Carl von Ossietzky University Oldenburg. I thank an anonymous referee for valuable comments that really helped to improve the quality of the paper. I also thank Prof. T. Dickhaus for calling my attention to the work of Pötscher and Leeb.

References

Alfons, A., Croux, C., Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics, 7*(1), 226–248.

Alqallaf, F., Van Aelst, S., Yohai, V. J., Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics, 37*(1), 311–331.

Aravkin, A. Y., Burke, J. V., Pillonetto, G. (2013). Sparse/robust estimation and Kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *The Journal of Machine Learning Research, 14*(1), 2689–2728.

Avella-Medina, M. (2017). Influence functions for penalized M-estimators. *Bernoulli, 23*(4B), 3178–3196.

Averbukh, V., Smolyanov, O. (1967). The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys, 22*(6), 201–258.

Banerjee, O., Ghaoui, L. E., d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning Research, 9,* 485–516.

Berge, C. (1963). *Topological Spaces: Including a treatment of multi-valued functions, vector spaces, and convexity*. Courier Corporation.

Beutner, E., Zähle, H. (2010). A modified functional delta method and its application to the estimation of risk functionals. *Journal of Multivariate Analysis, 101*(10), 2452–2463.

Beutner, E., Zähle, H. (2016). Functional delta-method for the bootstrap of quasi-hadamard differentiable functionals. *Electronic Journal of Statistics, 10*(1), 1181–1222.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics, 34*(2), 559–583.

Bühlmann, P., Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22*(4), 477–505.

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

Chang, L., Roberts, S., Welsh, A. (2018). Robust lasso regression using tukey's biweight criterion. *Technometrics, 60*(1), 36–47.

Chen, X., Wang, Z. J., McKeown, M. J. (2010a). Asymptotic analysis of robust lassos in the presence of noise with large variance. *IEEE Transactions on Information Theory, 56*(10), 5131–5149.

Chen, X., Wang, Z. J., McKeown, M. J. (2010b). Asymptotic analysis of the Huberized lasso estimator. In *2010 IEEE International conference on acoustics speech and signal processing (ICASSP),*, pages 1898–1901. IEEE.

Christmann, A., Steinwart, I. (2004). On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research, 5,* 1007–1034.

Christmann, A., Van Messem, A. (2008). Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research, 9*(May), 915–936.

Christmann, A., Van Messem, A., Steinwart, I. (2009). On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Statistics and Its Interface, 2*(3), 311–327.

Clarke, B. R. (1983). Uniqueness and fréchet differentiability of functional solutions to maximum likelihood type equations. *The Annals of Statistics, 11*(4), 1196–1205.

Clémençon, S., Lugosi, G., Vayatis, N. (2008). Ranking and empirical minimization of U-statistics. *The Annals of Statistics, 36*(2), 844–874.

Clémençon, S., Depecker, M., Vayatis, N. (2013). An empirical comparison of learning algorithms for nonparametric scoring: The TreeRank algorithm and other methods. *Pattern Analysis and Applications, 16*(4), 475–496.

De los Reyes, J. C., Schönlieb, C.-B., Valkonen, T. . (2016). The structure of optimal parameters for image restoration problems. *Journal of Mathematical Analysis and Applications, 434*(1), 464–500.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*(2), 407–499.

Evgrafov, A., Patriksson, M. (2003). Stochastic structural topology optimization: discretization and penalty function approach. *Structural and Multidisciplinary Optimization, 25*(3), 174–188.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Fernholz, L. (1983). Lecture notes in statistics. In *Von Mises Calculus for Statistical Functionals*, volume 19. Springer.

Fraiman, R., Yohai, V. J., Zamar, R. H. (2001). Optimal robust m-estimates of location. *The Annals of Statistics, 29*(1), 194–223.

Fréchet, M. (1937). Sur la notion de différentielle dans l'analyse générale.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics, 1*(2), 302–332.

Friedman, J., Hastie, T., Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432–441.

Gill, R. D., Wellner, J. A., Præstgaard, J. (1989). Non- and semi-parametric maximum likelihood estimators and the von mises method (part 1)[with discussion and reply]. *Scandinavian Journal of Statistics, 16*(2), 97–128.

Hable, R. (2012). Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis, 106,* 92–117.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*(346), 383–393.

Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W. (2011). *Robust statistics*: *The approach based on influence functions*, (Vol. 114). John Wiley & Sons.

Huber, P. J., Ronchetti, E. (2009). *Robust Statistics*. Wiley.

Jain, N., Marcus, M. (1975). Central limit theorems for C(S)-valued random variables. *Journal of Functional Analysis, 19*(3), 216–231.

Kohl, M. (2005). *Numerical contributions to the asymptotic theory of robustness*. PhD thesis, University of Bayreuth.

Krätschmer, V., Schied, A., Zähle, H. (2012). Qualitative and infinitesimal robustness of tail-dependent statistical functionals. *Journal of Multivariate Analysis, 103*(1), 35–47.

Lambert-Lacroix, S., Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics, 5,* 1015–1053.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics, 41*(3), 802–828.

LeDell, E., Petersen, M., van der Laan, M. (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic Journal of Statistics, 9*(1), 1583.

Lee, S. (2015). An additive sparse penalty for variable selection in high-dimensional linear regression model. *Communications for Statistical Applications and Methods, 22*(2), 147–157.

Levitin, E., Tichatschke, R. (1998). On smoothing of parametric minimax-functions and generalized max-functions via regularization. *Journal of Convex Analysis, 5,* 199–220.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics, 45*(2), 866–896.

Loh, P.-L., Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research, 16,* 559–616.

Maronna, R., Martin, R., Yohai, V. (2006). Robust statistics: Theory and methods. *Annals of Statistics, 30,* 17–23.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B. (2012). A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers. *Statistical Science, 27*(4), 538–557.

Öllerer, V., Croux, C., Alfons, A. (2015). The influence function of penalized regression estimators. *Statistics, 49*(4), 741–765.

Osborne, M. R., Presnell, B., Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis, 20*(3), 389–403.

Pötscher, B. M., Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis, 100*(9), 2065–2082.

Pötscher, B. M., Schneider, U. (2009). On the distribution of the adaptive lasso estimator. *Journal of Statistical Planning and Inference, 139*(8), 2775–2790.

Pupashenko, D., Ruckdeschel, P., Kohl, M. (2015). L2 differentiability of generalized linear models. *Statistics & Probability Letters, 97*(C), 155–164.

Reeds, J. (1976). *On the definition of von Mises functionals*. PhD thesis, Harvard University.

Rieder, H. (1994). *Robust asymptotic statistics* (Vol. 1). Springer Science & Business Media.

Rieder, H., Kohl, M., Ruckdeschel, P. (2008). The cost of not knowing the radius. *Statistical Methods & Applications, 17*(1), 13–40.

Rieder, H., Ruckdeschel, P. (2001). Short proofs on L_r-differentiability. *Statistics & Risk Modeling, 19*(4), 419–426.

Rosset, S., Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics, 35*(3), 1012–1030.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58,* 267–288.

Van de Geer, S. (2014). Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics, 41*(1), 72–86.

Van de Geer, S. (2016). *Estimation and testing under sparsity*. Springer.

Van der Vaart, A. (2000). *Asymptotic statistics,* (Vol. 3). Cambridge University Press.

Van der Vaart, A., Wellner, J. (2013). *Weak convergence and empirical processes*: *With applications to statistics*. Springer Science & Business Media.

Vapnik, V. (1998). *Statistical learning theory* (Vol. 1). New York: Wiley.

Vito, E. D., Rosasco, L., Caponnetto, A., Piana, M., Verri, A. (2004). Some properties of regularized kernel methods. *Journal of Machine Learning Research, 5,* 1363–1390.

Von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics, 18*(3), 309–348.

Wellner, J. A. (1992). Empirical processes in action: A review. *International Statistical Review/Revue Internationale de Statistique,* 247–269.

Werner, D. (2006). *Funktionalanalysis*. Springer.

Werner, T. (2019). *Gradient-Free Gradient Boosting*. PhD thesis, Carl von Ossietzky Universität Oldenburg.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101,* 1418–1429.

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.