

Oracle Inequalities for Convex Loss Functions with Non-Linear Targets

MEHMET CANER*

ANDERS BREDAHL KOCK†

April 30, 2015

Abstract

This paper considers penalized empirical loss minimization of convex loss functions with unknown target functions. Using the elastic net penalty, of which the Lasso is a special case, we establish a finite sample oracle inequality which bounds the loss of our estimator from above with high probability. If the unknown target is linear this inequality also provides an upper bound of the estimation error of the estimated parameter vector. Next, we use the non-asymptotic results to show that the excess loss of our estimator is asymptotically of the same order as that of the oracle. If the target is linear we give sufficient conditions for consistency of the estimated parameter vector. We briefly discuss how a thresholded version of our estimator can be used to perform consistent variable selection. We give two examples of loss functions covered by our framework.

Keywords and phrases: Empirical loss minimization, Lasso, Elastic net, Oracle inequality, Convex loss function, Nonparametric estimation, Variable selection.

JEL classification: C13, C21, C31.

*North Carolina State University, Department of Economics, 4168 Nelson Hall, Raleigh, NC 27695. Email: mcaner@ncsu.edu.

†Aarhus University and CREATES, Department of Economics, Fuglesangs Alle 4, 8210 Aarhus V, Denmark. Financial support from the Danish National Research Foundation (DNRF78) is gratefully acknowledged. We thank the editor of the special issue Marcelo Medeiros and anonymous referee for their comments that substantially changed the paper. We thank an anonymous referee for substantially modifying and getting a better result for Theorem 1.

1 Introduction

High-dimensional data sets have become increasingly available to researchers in many fields. In economics, large datasets generated from scanner data can be found in the analysis of price formation and consumer behavior. Many macroeconomic variables are sampled rather infrequently leaving one with many variables compared to observations in models with many explanatory variables. Financial data is also of a high-dimensional nature with a large number of variables and instruments being observed in small intervals due to high-frequency trading. Alternatively, models with many variables emerge when trying to control for non-linearities in a wage regression by including basis functions of the space in which the non-linearity is supposed to be found. Including more basis functions improves the approximation at a cost of increasing number of variables in the model.

High-dimensional models are models where the number of variables are larger than the number of observations. For these reasons handling high-dimensional data sets has received a lot of attention in the econometrics and statistics literature in the recent years. Tibshirani (1996) has introduced the Lasso estimator which carries out variable selection and parameter estimation simultaneously. The theoretical properties of this estimator have been studied extensively since then in various papers and extensions such as the adaptive Lasso by Zou (2006), the bridge estimator by Huang et al. (2008), the SCAD estimator by Fan and Li (2001), the sure independence screening by Fan and Lv (2008) or the square root Lasso by Belloni et al. (2011) have been proposed. For recent reviews see, e.g., Fan et al. (2011), Bühlmann and van de Geer (2011) or Belloni and Chernozhukov (2011).

Lasso-type estimators have been useful in econometrics literature. For example Belloni et al. (2012) have established results in the context of instrumental variable estimation without imposing the hitherto much used assumption of sub-gaussianity by means of moderate deviation theorems for self-normalized random variables. Furthermore, they allow for heteroscedastic error terms which is pathbreaking and greatly widens the scope of applicability of their results.

Applications to panel data may be found in e.g. Kock (2013). The estimators have been studied in the context of GMM, factor models, and smooth penalties by, among others, Caner and Zhang (2014), Caner and Han (2014), Cheng and Liao (2013) and Fan and Li (2001). Within linear time series models oracle inequalities have been established by Kock and Callot (2013) and Negahban et al. (2012) have proposed a unified framework which is valid for regression as well as matrix estimation problems.

Most research has considered the linear regression model or other parametric models. In this paper we focus on a very general setup. Specifically we consider, penalized empirical loss minimization of convex loss functions with potentially non-linear target functions. van de Geer (2008) studies a similar setup for the Lasso which is a special case of our results for the elastic net. In chapter 6 of Bühlmann and van de Geer (2011) generalized linear models with Lasso penalty are discussed. The elastic net, which is first proposed by Zou and Hastie (2005), has the advantage over the Lasso that it deals better with highly correlated variables. Lasso has a tendency to only retain one of two relevant variables if these are highly correlated while the elastic net does not suffer from this problem. Furthermore, our study of general convex loss functions is motivated by the fact that Hebiri and van de Geer (2011) have shown that in the case of a linear model with a quadratic loss function the elastic net may behave better than the Lasso in terms of the so-called restricted eigenvalue condition. Though we can not provide a comparison like this in the general case of non-linear models and convex loss functions¹ it is useful to establish finite sample upper bounds on the estimation and prediction error of the elastic net in this general setting. From an applied perspective, Caner and Han (2014) have shown how the elastic net was useful in estimating the external habit specification model of Chen and Ludvigson (2009) by GMM. Furthermore, even though our main focus is on non-asymptotic bounds, we also present asymptotic upper bounds on the excess risk and on the estimation error for linear targets.

In this paper we

1. provide a finite sample oracle inequality for empirical risk minimization penalized by the elastic net penalty. This inequality is valid for convex loss functions and non-linear targets and contains an oracle inequality for the Lasso as a special case.
2. For the case where the target function is linear this oracle inequality can be used to establish finite sample upper bounds on the estimation error of the estimated parameter vector.
3. The finite sample inequality is used to establish asymptotic results. In particular, the excess risk of our estimator is of the same order as that of an oracle which trades off the approximation and estimation errors. When the target is linear we give sufficient conditions for consistency of the estimated parameter vector.

¹The tradeoff between the Lasso and the elastic net is illustrated by Hebiri and van de Geer (2011) relies crucially on the loss function being quadratic and does not generalize easily.

4. In the case where the target is linear we briefly explain how a thresholded version of our estimator can unveil the correct sparsity pattern.
5. We provide two examples of specific loss functions covered by our general framework. We verify in detail that the abstract conditions of the above general results are satisfied in these two common settings.

We stress here that our main objective is to establish upper bounds on the performance of the elastic net. It is not our intention to promote either the Lasso or the elastic net, merely to analyze the properties of the latter. However, we shall make some brief comments on merits of the two procedures when compared to each other. A clear ranking like the one in Hebiri and van de Geer (2011) is not available at this point for the reasons mentioned above. As mentioned, these authors only focus on quadratic loss for which a certain data augmentation trick facilitates the analysis.

We believe that the performance guarantees on the elastic net provided by this paper are useful for the applied researcher who increasingly faces high-dimensional data sets. The usefulness is enhanced by the fact that our results are valid for a wide range of loss functions.

The paper is organized as follows. Section 2 puts forward the setup and notation. Section 3 introduces the main result, the oracle inequality for empirical loss minimization of convex loss functions penalized by the elastic net. Section 4 briefly discusses consistent variable selection by a thresholded version of the elastic net. Section 5 shows that the quadratic as well as the logistic loss are covered by our framework.

2 Setup and notation

We begin by setting the stage for general convex loss minimization. The setup is similar to the Lasso one in Section 6.3 in Bühlmann and van de Geer (2011). Let (Ω, \mathcal{F}, P) be a standard probability space. Consider a sample $\{Z_i\}_{i=1}^n = \{X_i, Y_i\}_{i=1}^n$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$. Here, for the sake of exposition, \mathcal{X} can be thought of as a subset of \mathbb{R}^p for $p \geq 1$. Define $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and let \mathbf{F} be a normed real vector space with norm $\|\cdot\|$. For each $f \in \mathbf{F}$ let $\rho_f : \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function. More precisely, $f \in \mathbf{F}$ will be a function $f : \mathcal{X} \rightarrow \mathbf{R}$ and the corresponding norm will be the $L_2(P)$ -norm $(\int f^2(X)dP)^{1/2}$. Furthermore, when v is a vector in \mathbb{R}^p , $\|v\|_1 = \sum_{j=1}^p |v_j|$ denotes the ℓ_1 -norm while $\|v\|_2 = \sqrt{\sum_{j=1}^p v_j^2}$ denotes the ℓ_2 -norm. Order symbols such as o, O, o_p and O_p are used with their usual meanings. Finally, for any abstract set A , $|A|$ denotes its cardinality. Throughout the paper we shall assume:

Assumption 0: $\{Z_i\}_{i=1}^n = \{X_i, Y_i\}_{i=1}^n$ is an iid sample and the mapping $f \mapsto \rho_f(z)$ is convex for all $z \in \mathcal{Z}$.

The assumption of identically distributed data is not essential. We have made it here for convenience and our results remain valid when the data is merely independent upon a few minor modifications. The following examples provide illustrations of when the conditions in Assumption 0 are met:

Quadratic loss

Let $\mathcal{X} \subset \mathbb{R}^p$ and

$$Y_i = f^0(X_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where ϵ_i is some real error term and $f^0 \in \mathbf{F}$. Then, the standard case of quadratic loss is covered by the above setting upon choosing $\rho_f(x, y) = (y - f(x))^2$ which is clearly convex in $f(x)$. By letting \mathbf{F} only consist of linear functions $f(x) = \beta'x$ for some $\beta \in \mathbb{R}^p$ the case of linear least squares is covered. Non-linear least squares is covered by choosing $f(x) = g(\beta, x)$ for some parameter vector β .

Logistic loss

Let

$$Y_i^* = f^0(X_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where ϵ_i is independent of X_i and assumed to have a logistic distribution while $f^0 \in \mathbf{F}$. Assume that $Y_i = 1$ if $Y_i^* > 0$ and $Y_i = 0$ otherwise. Since ϵ_i has cdf $F(z) = \frac{e^z}{1+e^z}$ one gets

$$P(Y_i = 1 | X_i = x) = P(f^0(X_i) + \epsilon_i > 0 | X_i = x) = \frac{e^{f^0(x)}}{1 + e^{f^0(x)}}.$$

Note that for $\mathcal{X} \subseteq \mathbb{R}^p$ and $f^0(x) = \beta^{0'}x$ for some parameter vector $\beta \in \mathbb{R}^p$ this is the usual expression for $P(Y_i = 1 | X_i = x)$ in the logit model. The above setting is more general, however, since it allows f^0 to be non-linear.

The log-likelihood function for a given $f \in \mathbf{F}$ is then given by (for $z = (x, y)$)

$$l(f | z_1, \dots, z_n) = \sum_{i=1}^n \left[y_i \log\left(\frac{e^{f(x_i)}}{1 + e^{f(x_i)}}\right) + (1 - y_i) \log\left(1 - \frac{e^{f(x_i)}}{1 + e^{f(x_i)}}\right) \right] = \sum_{i=1}^n \left[y_i f(x_i) - \log(1 + e^{f(x_i)}) \right].$$

Hence, a sensible loss function is the negative log-likelihood

$$\rho_f(x, y) = -yf(x) + \log(1 + e^{f(x)}),$$

which is convex in $f(x)$.

Negative log-likelihood

The above two examples are both instances of the loss function being the negative of the log-likelihood². Hence, in a general setting with the negative of the log-likelihood being a convex function in $f(x)$ our results also apply. Again, a special case is $f(x) = \beta'x$.

Returning to the general setup, denote by $P_n\rho_f = \frac{1}{n} \sum_{i=1}^n \rho_f(X_i, Y_i)$ and $P\rho_f = E\rho_f(X_i, Y_i)$ the empirical and population means of the loss function for a fixed $f \in \mathbf{F}$. We shall also denote these two quantities the empirical and population risk, respectively. We define our target as the minimizer of the theoretical risk

$$f^0 := \operatorname{argmin}_{f \in \mathbf{F}} P\rho_f,$$

where it is tacitly assumed that the minimizer exists and is unique for the $\|\cdot\|$ -norm on \mathbf{F} . Then, for any $f \in \mathbf{F}$, we define the excess population risk over the target as

$$\Xi(f) := P(\rho_f - \rho_{f^0}).$$

Note that, by construction, $\Xi(f) \geq 0$ for all $f \in \mathbf{F}$. Since the joint distribution of (Y_i, X_i) is assumed to be unknown we shall consider empirical risk minimization instead of minimizing the population excess risk. Put differently, $P_n\rho_f$ is minimized. Furthermore, we will consider a linear subspace $\mathcal{F}_L = \{f_\beta(x) = \sum_{j=1}^p \beta_j \psi_j(x), \beta \in \Phi\}$ of \mathbf{F} where $\psi_j(x) : \mathcal{X} \rightarrow \mathbb{R}$ may be thought of as basis functions of \mathbf{F} . Of course, in the case where \mathcal{X} is a subset of \mathbb{R}^p , one could also think of $\psi_j(x)$ as being the j 'th coordinate projection. This is the choice we make whenever f^0 is assumed to be linear. In general, $\psi(x) = (\psi_1(x), \dots, \psi_p(x))'$ denotes a vector of (transformed) covariates. Φ is a convex subset of \mathbb{R}^p – in many cases we can even have $\Phi = \mathbb{R}^p$. In Section 5 we shall see an example of $\Phi = \mathbb{R}^p$ but also an example of Φ being a subset of \mathbb{R}^p . In case the target function f^0 is linear we will denote its parameter vector by β^0 .

²The quadratic loss example can be thought of as the negative of a Gaussian log-likelihood.

2.1 The elastic net

Define

$$\hat{f} = f_{\hat{\beta}} = \operatorname{argmin}_{f_{\beta}: \beta \in \Phi} \left(P_n \rho_{f_{\beta}} + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right) = \operatorname{argmin}_{f \in \mathcal{F}_L} \left(P_n \rho_f + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right),$$

where λ_1 and λ_2 are positive constants. Hence, we are minimizing the empirical risk *plus* an elastic net penalty. This form of penalty was originally introduced by Zou and Hastie (2005) in the case of a linear regression model. The penalty is a compromise between the ℓ_1 -penalty of the plain Lasso and the squared ℓ_2 -loss in ridge regression. Ridge regression does not perform variable selection at all – all estimated coefficients are non-zero. On the other hand, if two variables are highly correlated, the Lasso has a tendency to include only one of these. A concrete economic example where this tradeoff is relevant is provided Kallestrup-Lamb et al. (2013) who investigate the determinants of the retirement decision based on a vast Danish register data set. The authors pay particular attention the effect of a big category of health related variables. Many of these health variables may be highly correlated but it might not be the diagnosis of a single illness which causes a person to retire. As a consequence, only including one of these may not provide the full picture of the effects of health related variables.

The elastic net strikes a balance between the ridge and the Lasso and hence performs particularly well in the presence of highly correlated variables. As mentioned in the introduction this benefit has been formalized by Hebiri and van de Geer (2011) in the case of quadratic loss. In particular, they have shown that the elastic net behaves better with respect to certain restricted eigenvalue conditions than the plain Lasso. To be precise, they use a data augmentation idea which hinges crucially on the loss function being quadratic. In that way the usual Gram matrix is perturbed by λ_2 times the identity matrix ensuring that this perturbed Gram matrix has a positive restricted eigenvalue as long as $\lambda_2 > 0$. For details we refer to Section 3.1 in Hebiri and van de Geer (2011), in particular expressions (8) and (9) as well as their Theorem 1. As mentioned, the data augmentation idea hinges on the loss function being quadratic. Hence, we can not make a general comparison of the Lasso and the elastic net here. Instead we will focus on deriving oracle inequalities for the elastic net.

2.2 Assumptions

We next turn to the assumptions needed to prove oracle inequalities for the elastic net. First, define $\mathbf{F}_{\text{local}} = \{f \in \mathbf{F} : \|f - f^0\|_\infty \leq \eta\}$ for some $\eta > 0$ where $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ ³. Note that even though L_∞ ball around f^0 seems restrictive at first, there are econometric examples that satisfy that restriction in section 5. Also related to examples, L_∞ is quite standard in sieve estimation as in Chen (2007).

The margin condition requires that in $\mathbf{F}_{\text{local}} \subseteq \mathbf{F}$ the excess loss $\Xi(f)$ is bounded from below by a convex function of $\|f - f^0\|$.

Definition. We say that the **margin condition** holds with strictly convex margin function $G(\cdot)$, if for all $f \in \mathbf{F}_{\text{local}}$ we have

$$\Xi(f) \geq G(\|f - f^0\|).$$

In all examples we shall consider it can be shown that the margin condition holds for $G(u) = cu^2$ for some $c > 0$ such that for all $f \in \mathbf{F}_{\text{local}}$, $\Xi(f) \geq c\|f - f^0\|^2$. More generally, we present a sufficient condition for G to be quadratic in Section 5. The convex conjugate of $G(\cdot)$ will also play a role in the development of the oracle inequalities below. In particular, the following definition is taken from page 121 in Bühlmann and van de Geer (2011) and many more properties of convex conjugates can be found in Rockafellar (1997).

Definition. Let G be a strictly convex function on $[0, \infty)$ with $G(0) = 0$. The **convex conjugate** H of G is defined as

$$H(v) = \sup_{u \geq 0} \{uv - G(u)\}, \quad v \geq 0.$$

Lemma 3 in the appendix establishes some properties of $H(v)$. Note also that if $G(u) = cu^2$, then $H(v) = v^2/(4c)$. Furthermore, from the definition of the convex conjugate

$$uv \leq G(u) + H(v), \tag{1}$$

also known as Fenchel's inequality. For any subset S of $\{1, \dots, p\}$ and $\beta \in \mathbb{R}^p$ we define β_S such that $\beta_{S,j} = \beta_j 1_{\{j \in S\}}$ for $j = 1, \dots, p$. Letting $|S| = s$ denote the cardinality of S and assuming that the covariates are identically distributed, we define

$$\phi^2(S) := \min_{\beta \in \mathbb{R}^p} \frac{\|f_\beta\|^2}{\|\beta_S\|_2^2} = \frac{\beta' \Sigma \beta}{\|\beta_S\|_2^2},$$

³Using the stronger $\|\cdot\|_\infty$ topology on \mathbf{F} to define $\mathbf{F}_{\text{local}}$ instead of $\|\cdot\|$ turns out to be useful when verifying that the margin condition is satisfied with a quadratic margin in Section 5.

where in the second equality we use $\|f_\beta\|^2 = \beta' \Sigma \beta$ with $\Sigma = E(\psi(X_1)\psi'(X_1))$. Since

$$\phi^2(S) = \min_{\beta \in \mathbb{R}^p} \frac{\beta' \Sigma \beta}{\|\beta_S\|_2^2} \geq \min_{\beta \in \mathbb{R}^p} \frac{\beta' \Sigma \beta}{\|\beta\|_2^2},$$

for all $S \in \{1, \dots, p\}$, $\phi^2(S) > 0$ in particular when the smallest eigenvalue of the population covariance matrix is positive. Here it is very important to notice that it is the smallest eigenvalue of the *population* covariance matrix which has to be positive – not the smallest eigenvalue of the empirical covariance matrix (which may often be singular). Since it is rather standard to assume that the population covariance matrix is positive definite $\phi^2(S) > 0$ will often be satisfied.

At this point it is also worth remarking that usually a so-called restricted eigenvalue is assumed to be positive, see e.g. Bickel et al. (2009), van de Geer and Bühlmann (2009) (Section 10) or Bühlmann and van de Geer (2011) (Section 6.1.1). The restricted eigenvalue of a matrix may be larger than the smallest eigenvalue of a matrix which is what we are dealing with. However, it is usually the restricted eigenvalue of the empirical Gram matrix which has to be non-zero. On the other hand we are requiring the smallest eigenvalue of the *population* covariance matrix to be non-zero. Since the population covariance matrix is non-random no concentration inequalities have to be applied in order to verify it. This is in contrast to the classical restricted eigenvalue condition on the empirical Gram matrix which often requires considerably more work to verify. For these reasons the ranking between the traditional restricted eigenvalue condition and our population minimum eigenvalue is not clear in terms of applicability. In any case, we believe that our approach is an interesting alternative to the classical condition.

Remark: Of course one could try to combine the best of the two worlds mentioned above: namely trying to impose a restricted eigenvalue type condition on the *population* covariance matrix. However, the presence of the squared ℓ_2 -norm in the penalty of the elastic net turns out to make this a non-trivial exercise since the appropriate definition of the restricted set in the restricted eigenvalue condition would interfere adversely with the definition of the oracle below.

Before defining what we understand by the oracle estimator, define $S_\beta = \{j : \beta_j \neq 0\}$ as the subset of $\{1, \dots, p\}$ containing the indices of the non-zero coefficients, $s_\beta = |S_\beta|$ its cardinality and let Γ denote a collection of subsets of $\{1, \dots, p\}$.

Definition *The oracle estimator is defined as*

$$\beta^* = \operatorname{argmin}_{\beta: S_\beta \in \Gamma} \left\{ 3\Xi(f_\beta) + 2H \left(\frac{4\lambda_1 \sqrt{s_\beta} + 4\lambda_2 \|\beta\|_2}{\phi(S_\beta)} \right) \right\}. \quad (2)$$

Note that the definition of the oracle still leaves considerable freedom since Γ is defined by the user – a property used later for handling linear targets (see Remark 2 after Theorem 1). When, Γ is the

power set of $\{1, \dots, p\}$ the oracle estimator may equivalently be written as

$$\beta^* = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ 3\Xi(f_\beta) + 2H \left(\frac{4\lambda_1 \sqrt{s_\beta} + 4\lambda_2 \|\beta\|_2}{\phi(S_\beta)} \right) \right\}.$$

The definition of the oracle in (2) turns out to be convenient for technical reasons but it also has a useful interpretation as a tradeoff between approximation and estimation error: In the standard setting of a quadratic loss function with a linear target, i.e. $f^0(x) = x'\beta^0$, the squared ℓ_2 -estimation error of Lasso type-estimators is of the order $\frac{s \log(p)}{n}$, where p is the total number of parameters of which s are non-zero. In the case of quadratic loss and iid samples we have $\Xi(f) = E(f(X_1) - f^0(X_1))^2$ if X_1 and ϵ_1 are independent. Hence, both G and hence H are quadratic in the definition of the margin condition. Choosing $\lambda_2 = \frac{\lambda_1 \sqrt{s}}{2\|\beta\|}$ and λ_1 of the order $\sqrt{\log(p)/n}$, which are both choices we shall adhere to in the sequel, one finds that $H(\cdot)$ is of the order $\frac{s \log(p)}{n}$. This is exactly the *estimation error* under quadratic loss and motivates coining $H(\cdot)$ the *estimation error* term. The term $\Xi(f_\beta)$ is referred to as the *approximation error* and (2) shows that the oracle trades of these two terms: a lower approximation error can be obtained by increasing s_β while this also implies estimating more parameters resulting in a higher estimation error.

Finally, letting $S^* = S_{\beta^*}$, $\phi_* = \phi(S^*)$, $s^* = |S^*|$ and $f^* = f_{\beta^*}$ we denote the oracle bound (value of the objective function minimized by the oracle) by

$$2\Delta^* := 3\Xi(f^*) + 2H \left(\frac{4\lambda_1 \sqrt{s^*} + 4\lambda_2 \|\beta^*\|_2}{\phi_*} \right).$$

The inequality in Theorem 1 below is valid on a random set which we introduce next. In Theorem 2 we show that this set actually has a high probability by means of a suitable concentration inequality for suprema of empirical processes. Define the empirical process

$$\left\{ V_n(\beta) = (P_n - P)(\rho_{f_\beta}) : \beta \in \mathbb{R}^p \right\}.$$

Next, we introduce a local supremum of the empirical process in incremental form

$$Z_M = \sup_{\|\beta - \beta^*\|_1 \leq M} |V_n(\beta) - V_n(\beta^*)|. \quad (3)$$

Then we define

$$M^* = \Delta^* / \lambda_0,$$

where λ_0 is a positive sequence defined explicitly in Theorem 2, and set

$$\tau = \{Z_{M^*} \leq \lambda_0 M^*\} = \{Z_{M^*} \leq \Delta^*\}.$$

The set τ is the one we shall work on in Theorem 1 below. Note in particular, that on τ , Z_M can not be larger than Δ^* which is the minimal value of the loss function of the oracle.

We are now ready to state our assumptions:

Assumption 1. Assume the margin condition with strictly convex function $G(\cdot)$.

Assumption 2. Assume that $f^* \in \mathbf{F}_{\text{local}}$ and $f_\beta \in \mathbf{F}_{\text{local}}$ for all $\|\beta - \beta^*\|_1 \leq M^*$.

Assumption 3. Assume that $\phi(S^*) > 0$.

As discussed above, the margin condition, Assumption 1, regulates the behavior of the excess risk function. When \mathbf{F} is equipped with the $L_2(P)$ we will see in Section 5 that the margin condition is actually often satisfied with $G(\cdot)$ being quadratic. Put differently, the margin condition is satisfied in many examples with a quadratic margin.

Assumption 2 is essentially technical and enables us to use the margin condition for f^* . The first part requires that the oracle is a good approximation to the target f^0 in the sup-norm. The validity of this condition depends on how well f^0 is approximated by linear combinations of elements in $\{\psi_j\}_{j \in S}$, $S \in \Gamma$. The second part of Assumption 2 states that $f_\beta \in \mathbf{F}_{\text{local}}$ if β and β^* are sufficiently close. This condition essentially imposes a relationship between an L_1 ball around β^* and the approximation properties of f_β inside this set. From the triangle inequality, this condition is satisfied if $\max_j \|\psi_j\|_\infty \Delta^* / \lambda_0 + \|f^* - f\|_\infty \leq \eta$. In Section 5 we discuss examples satisfying Assumption 2. In particular, if \mathbf{F} consists of sufficiently smooth functions⁴, we shall exhibit choices of bases $\{\psi_j\}_{j=1}^\infty$ and collections of sets Γ such that f^* approximates f^0 to the desired degree. Assumption 3 was discussed before and is valid if the population covariance $E(\psi(X_1)\psi'(X_1))$ has full rank.

3 An Oracle Inequality

In this section we extend Theorem 6.4 of Bühlmann and van de Geer (2011) from ℓ_1 penalty (Lasso) to $\ell_1 + \ell_2^2$ penalty (Elastic Net). This is not a trivial extension since the basic inequality used to establish the result has to be altered considerably. More precisely, the inequality that ties the estimator to the oracle has to be modified. The second difference is that we use $\phi(S^*) > 0$ which is different from the compatibility condition used in the ℓ_1 -case. Compared to the linear target with quadratic loss in Hebiri and van de Geer (2011) estimated by the elastic net our proof cannot benefit from the augmented regressors idea since this idea relies crucially on the loss function being

⁴To be concrete, we shall be considering a Hölder class of function to be defined precisely in Section 5.

quadratic. In the case of general convex loss function the proof technique is entirely different and we use the margin condition, Fenchel's inequality, and a careful definition of the oracle instead. We would like to stress that Theorem 1 below is purely deterministic in the sense that there are no probabilities attached to it. It is valid on the set τ to which we shall later attach a lower bound on its probability. It also provides a finite sample result – i.e. the result is valid for any sample size and not just asymptotically.

Theorem 1. *Suppose λ_1 satisfies $\lambda_1 \geq 8\lambda_0$. Then on the set τ , under Assumptions 1-3, we have*

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^*\|_1 + \lambda_2 \|\hat{\beta} - \beta^*\|_2^2 \leq 4\Delta^* = 6\Xi(f^*) + 4H\left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi_*}\right),$$

where $H(\cdot)$ is the convex conjugate of the function $G(\cdot)$ in the margin condition.

Note that Theorem 1 provides an upper bound, $4\Delta^*$, on the excess loss of \hat{f} in terms of the excess loss of the oracle f^* as well as an extra term $H(\cdot)$, the estimation error. We shall comment much more on this extra term in the sequel. Theorem 1 can also be used to give an upper bound on the ℓ_1 -estimation error. Due to its importance the theorem warrants some detailed remarks.

1. The result of Theorem 1 reduces to the result for the Lasso in Theorem 6.4 of Bühlmann and van de Geer (2011) when we set $\lambda_2 = 0$ except for the fact that we assume ϕ_* rather than their compatibility constraint. In that sense we generalize the oracle inequality of Bühlmann and van de Geer (2011). Their oracle inequality is, with ϕ_{**} being a compatibility constant (see p.157, Bühlmann and van de Geer (2011))

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^*\|_1 \leq 6\Xi(f^*) + 4H\left(\frac{4\lambda_1\sqrt{s^*}}{\phi_{**}}\right).$$

As mentioned, the only difference between their Theorem 6.4 and the result that can be deduced from our Theorem 1 is that $\phi_{**} \geq \phi_*$.

2. Letting β^{BL} denote $\operatorname{argmin}_{\beta} \rho_{f_\beta}$, i.e. the best linear approximation, and setting $\Gamma = S_{BL} = \{j : \beta_j^{BL} \neq 0\}$ and choosing $\lambda_2 = \frac{\lambda_1\sqrt{|\beta^{BL}|}}{2\|\beta^{BL}\|_2}$ it follows that

$$\beta^* = \operatorname{argmin}_{\beta \in S_{BL}} \left\{ 3\Xi(f_\beta) + 2H\left(\frac{6\lambda_1\sqrt{|S_{BL}|}}{\phi(S_{BL})}\right) \right\}.$$

Note how we have used our discretion in making a choice of Γ which will turn out to be useful below. Since the second term in the definition of β^* does not depend on β in this case, it follows that β^* is the minimizer of $\Xi(f_\beta)$ which itself also is the minimizer of ρ_{f_β} . Hence, $\beta^* = \beta^{BL}$ in this

case. It follows that under the conditions of Theorem 1

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^{BL}\|_1 + \lambda_2 \|\hat{\beta}^{BL} - \beta^{BL}\|_2^2 \leq 6\Xi(f_{\beta^{BL}}) + 4H \left(\frac{6\lambda_1 \sqrt{|S_{BL}|}}{\phi(S_{BL})} \right). \quad (4)$$

So, Theorem 1 can be used to provide upper bounds on the ℓ_1 -distance of $\hat{\beta}$ to β^{BL} due to our freedom in defining the oracle β^* . If the target f^0 is also linear then the best linear approximation equals the target implying that $\beta^{BL} = \beta^0$, $f_{\beta^{BL}} = f^0$ and $\Xi(f_{\beta^{BL}}) = \Xi(f^0) = 0$. Using $S_{BL} = S^0 = \{j : \beta_j^0 \neq 0\}$, inequality (4) yields

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^0\|_1 + \lambda_2 \|\hat{\beta} - \beta^0\|_2^2 \leq 4H \left(\frac{6\lambda_1 \sqrt{|S^0|}}{\phi(S^0)} \right). \quad (5)$$

Hence, in case the target is linear, (5) yields an upper bound on the ℓ_1 and ℓ_2 -estimation errors which do not depend on the excess loss of the oracle. We shall make use of this fact in Section 4 on variable selection. In practice, S^0 is unknown and (5) only yield theoretical bounds.

3. In many econometric examples the margin condition (Assumption 1) is satisfied with a quadratic margin resulting in $H(v) = v^2/4c$ for a positive constant c . Under this margin and setting $\lambda_2 = \frac{\lambda_1 \sqrt{s^*}}{2\|\beta^*\|_2}$ in Theorem 1 results in

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^*\|_1 + \lambda_1 \frac{\sqrt{s^*}}{2\|\beta^*\|_2} \|\hat{\beta} - \beta^*\|_2^2 \leq 6\Xi(f^*) + \frac{36\lambda_1^2 s^*}{c\phi_*^2}.$$

Note that for $\lambda_2 = 0$, corresponding to a pure ℓ_1 -loss, Theorem 1 reduces to

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^*\|_1 \leq 6\Xi(f^*) + \frac{16\lambda_1^2 s^*}{c\phi_*^2}.$$

Theorem 1 holds under the event $\tau = \{Z_{M^*} \leq \lambda_0 M^*\} = \{Z_{M^*} \leq \Delta^*\}$ that, ideally, holds with large probability. However, there is a tradeoff between the tightness of the bound and the probability in which the bound holds. Increasing λ_0 yield a large probability, but it also increases $\lambda_1 \geq 8\lambda_0$. We derive a probability bound for τ that takes into account this tradeoff.

Assume that $\rho_f(x, y) = \gamma(y, f(x)) + c(f)$ where $c(f)$ is a constant possibly depending on f . Also, we assume that there exists a $D > 0$ such that

$$|\gamma(y, f_\beta(x)) - \gamma(y, f_{\tilde{\beta}}(x))| \leq D|f_\beta(x) - f_{\tilde{\beta}}(x)|, \quad (6)$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\beta, \tilde{\beta} \in \Phi$. In other words $\gamma(\cdot, \cdot)$ is assumed to be Lipschitz continuous in its second argument over \mathcal{F}_L with Lipschitz constant D . The reason we only need Lipschitz

continuity over \mathcal{F}_L is that it is used for a contraction inequality in connection with bounding the local supremum of the empirical process Z_M in (3). Assume furthermore that

$$\max_{1 \leq j \leq p} E\psi_j^2(X_i) \leq 1, \quad (7)$$

and for a positive constant K

$$\max_{1 \leq j \leq p} \|\psi_j\|_\infty \leq K. \quad (8)$$

Note that (8) implies (7) when $K = 1$. Assuming $\max_{1 \leq j \leq p} \|\psi_j\|_\infty \leq K$ is rather innocent since many commonly used basis functions are bounded. As we shall see in Section 5 the most critical assumption in concrete examples is the Lipschitz continuity of the loss function. With this notation in place we state the following result which builds on Theorem 14.5 Bühlmann and van de Geer (2011) (see also Corollary A.1 in van de Geer (2008)).

Theorem 2. *Assume (6)-(8), $p \geq 2$, and that $\log(p) \leq n$. Then, choosing $\lambda_0 = dD\sqrt{\frac{\log(p)}{n}}$, for some $d > 0$ yields*

$$P(\tau) \geq 1 - \left(\frac{1}{p}\right).$$

The assumption $p \geq 2$ is made for purely technical reasons and does not exclude any interesting problems. Similarly, $\log(p) \leq n$ still allows p to increase at an exponential rate in the sample size.⁵ Note that the bound is a finite sample bound. It holds for *any* choice of n, p, K and D . In particular, we may have different p, K and D for different n but we choose not to index these three quantities by n . When deriving asymptotic results we use that p (and sometimes D) may change with n .

From an asymptotic point of view Theorem 2 reveals that the measure of the set τ , on which the inequality in Theorem 1 is valid, tends to 1 as $p \rightarrow \infty$. The developments in this paper are for high dimensional problems, i.e. p increases with n . Combining Theorems 1 and 2 yields the following result

Corollary 1. *Under the assumptions of Theorems 1 and 2 with the choice of $\lambda_0 = dD\sqrt{\frac{\log(p)}{n}}$ for some positive constant d it holds with probability at least $1 - \left(\frac{1}{p}\right)$*

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^*\|_1 + \lambda_2 \|\hat{\beta} - \beta^*\|_2^2 \leq 4\Delta^* = 6\Xi(f^*) + 4H \left(\frac{4\lambda_1 \sqrt{s^*} + 4\lambda_2 \|\beta^*\|_2}{\phi_*} \right),$$

⁵This is from an asymptotic point of view, though we wish to emphasize that the inequality in Theorem 2 holds for any given sample size (satisfying the conditions of the theorem).

where $H(\cdot)$ is the convex conjugate of $G(\cdot)$.

Theorem 1 consists of the inequality in Theorem 1 with a lower bound attached to the measure of the set τ on which Theorem 1 is valid. It reveals that the excess loss of, \hat{f} , $\Xi(\hat{f})$, is close to the optimal one at the oracle. The second term on the right hand side reflects the estimation error. We have the following result for the asymptotic excess loss of \hat{f} .

Corollary 2. *Assume that $p \in O(\exp(n^a))$ and $|S^*| \in O(n^b)$ for some $a > 0$ and $b \geq 0$, as well as $\lambda_1 \leq L\lambda_0$ for some $L \geq 8$. Then, under the assumptions of Theorems 1 and 2, with $\lambda_2 = \lambda_1\sqrt{s^*}/2\|\beta^*\|_2$ and the choice of $\lambda_0 = dD\sqrt{\frac{\log(p)}{n}}$ for some positive constant d one has*

$$\limsup_{n \rightarrow \infty} \Xi(\hat{f}) \leq 6 \limsup_{n \rightarrow \infty} \Xi(f^*),$$

with probability approaching one if $a + b < 1$ and ϕ_* is bounded away from zero.

Corollary 2 shows that asymptotically the excess loss of \hat{f} will be of the same order as that of the oracle. In the case where f^0 is linear we know from Remark 2 above that we can choose Γ such that $f^* = f^0$ and hence $\Xi(f^*) = 0$. In this case Corollary 2 actually reveals that the excess loss of \hat{f} tends to zero. We next investigate the case of linear f^0 in more detail.

3.1 Linear target

In the case where the target function f^0 is linear Theorem 1 can be used to deduce the following result.

Corollary 3. *Assume that f^0 is linear.*

a) *Then, under the assumptions of Theorems 1 and 2*

$$\Xi(\hat{f}) \leq 4H\left(\frac{4\lambda_1\sqrt{|S^0|} + 4\lambda_2\|\beta^0\|_2}{\phi(S^0)}\right), \quad (9)$$

$$\|\hat{\beta} - \beta^0\|_1 \leq \frac{4}{\lambda_1}H\left(\frac{4\lambda_1\sqrt{|S^0|} + 4\lambda_2\|\beta^0\|_2}{\phi(S^0)}\right), \quad (10)$$

with probability at least $1 - \left(\frac{1}{p}\right)$.

b) *If, furthermore, the margin is quadratic such that $H(v) = v^2/(4c)$ for some $c > 0$, and we choose $\lambda_2 = \frac{\lambda_1\sqrt{|S^0|}}{2\|\beta^0\|_2}$ as well as $\lambda_1 \leq L\lambda_0$ for some $L \geq 8$, then, by (9) and (10)*

$$\Xi(\hat{f}) \leq \frac{36}{c} \frac{\lambda_1^2 |S^0|}{\phi^2(S^0)} \leq \frac{36}{c} \frac{(LdD)^2 \log(p) |S^0|}{n\phi^2(S^0)}, \quad (11)$$

$$\|\hat{\beta} - \beta^0\|_1 \leq \frac{36L}{c} \frac{\lambda_1 |S^0|}{\phi^2(S^0)} \leq \frac{36LdD}{c} \sqrt{\frac{\log(p)}{n}} \frac{|S^0|}{\phi^2(S^0)}, \quad (12)$$

with probability at least $1 - \left(\frac{1}{p}\right)$.

The bounds (9) and (10) bound the ℓ_1 -estimation error of the elastic net estimator for any type of loss function satisfying the conditions of Theorem 1. Note that there is no excess loss from the oracle entering in the upper bound. This is due to the fact that this is zero when the target is linear. The last two bounds in Corollary 3 specialize to the case where the quadratic margin condition is satisfied. We stress again, as we shall see later (see Section 5), that the margin condition is indeed quadratic in many econometric examples. One sees from the above Corollary that the rate of convergence of the elastic net estimator in the ℓ_1 -norm is $\frac{\sqrt{n}}{\sqrt{\log(p)}} \frac{1}{|S^0|}$ provided that $\phi^2(S^0)$ is bounded away from zero. Also (12) implies consistency of $\hat{\beta}$ for β^0 .

Corollary 4. *Assume that $p = O(\exp(n^a))$ and $|S^0| = O(n^b)$ for some $a > 0$ and $b \geq 0$, and f^0 is linear and set $\lambda_2 = \frac{\lambda_1 \sqrt{|S^0|}}{2\|\beta^*\|_2}$ as well as $\lambda_1 \leq L\lambda_0$ for some $L \geq 8$. Let the quadratic margin condition be satisfied and assume that $\phi^2(S^0)$ is bounded away from 0. Then, under the assumptions of Theorems 1 and 2,*

$$\|\hat{\beta} - \beta^0\|_1 \xrightarrow{p} 0,$$

if $a + 2b < 1$.

Corollary 4 shows that the elastic net can be consistent even when the dimension p increases at a subexponential rate in the sample size. Note, however, that the number of relevant variables, $|S^0|$ can not increase faster than the square root of the sample size (a can be put arbitrarily close to 0 to see this). Even though the total number of variables can be very large, the number of relevant variables must still be quite low. This is in line with previous findings for the linear model in the literature. We also remark that this requirement is slightly stricter than the one needed when considering the excess loss in Corollary 2 (in that corollary we only needed $a + b < 1$). Also note that the conditions in Corollary 4 are merely sufficient. For example, one can let $\phi^2(S^0)$ tend to zero at the price of reducing the growth rate of p and $|S^0|$.

So far we have not discussed how to choose the tuning parameters λ_1 and λ_2 . It is common practice to use 5 or 10-fold cross validation for this purpose. We too suggest using one of these cross validation procedures to select λ_1 and λ_2 .

4 Variable Selection

In this section we briefly comment on how the results in Section 3 can be used to perform consistent variable selection in the case where f^0 is a linear function⁶. First note, that the results in Corollaries 3 and 4 can be used to provide rates of convergence of $\hat{\beta}$ for β^0 in the ℓ_1 -norm in the case of a linear target. If one also assumes that $\min \{|\beta_j^0| : \beta_j^0 \neq 0\}$ is bounded away from zero by at least the rate of convergence of $\hat{\beta}$, it follows that no non-zero β^0 will be classified as such (see Lounici (2008) or Kock and Callot (2013)). Put differently, the elastic net possesses the *screening property*.

In order to remove all non-zero variables one may threshold the elastic net estimator by removing all estimates below a certain cutoff value. Standard arguments show that choosing the threshold of the order of the rate of convergence yields consistent model selection asymptotically. Since thresholding is a generic technique which is not specific to our setup we shall not elaborate further on this at this stage.

5 Econometric Examples

In this section we present sufficient conditions for validity of Assumptions 1-3 in econometric models. Recall that these Assumptions are sufficient for Theorem 1 to be valid. We also comment on sufficient conditions for the assumptions underlying Theorem 2. We focus on showing that the Lipschitz assumption (6) is satisfied. The conditions (7) and (8) on the basis functions are true for many families of basis functions and are assumed satisfied. First, we present a general sufficient condition for loss functions to satisfy a quadratic margin condition. This condition is then illustrated by two examples.

Assume that the loss function is of the form $\rho_f(x, y) = \rho(f(x), y)$ such that it only depends on x through $f(x)$. By Doob's representation, see Lemma 1.13 in Kallenberg (2002), we define

$$l(f(X), X) := E[\rho(f(X), Y) | (X, f(X))].$$

Furthermore, by iterated expectations, it suffices to show that $l(f(x), x)$ satisfies the margin condition in order to verify that this is the case for $\rho_f(x, y) = \rho(f(x), y)$. The target function will be

$$f^0(x) = \operatorname{argmin}_{f \in \mathbf{F}} l(f(x), x),$$

⁶If the target function is not linear we do not find it sensible to talk about consistent variable selection in a linear approximation of the target. Hence, this section restricts attention to the case where the target is linear.

which is the minimizer of the loss function and hence a natural choice. Next, fix an $x \in \mathcal{X}$ and assume that $l(a, x)$ is twice continuously differentiable in its first argument in an $\|\cdot\|_\infty$ -neighborhood of radius $\eta > 0$ around $f^0(x)$ with second derivative bounded from below by $2c > 0$, for $c > 0$. Then it follows by Lagrange's form of the remainder term in a Taylor series that for some \tilde{a} on the line segment joining $f(x)$ and $f^0(x)$

$$\begin{aligned} l(f(x), x) &= l(f^0(x), x) + l'_1(f^0(x), x)(f(x) - f^0(x)) + \frac{l''_{11}(\tilde{a}, x)}{2}(f(x) - f^0(x))^2 \\ &\geq l(f^0(x), x) + c(f(x) - f^0(x))^2, \end{aligned}$$

for all $f \in \mathbf{F}$ such that $|f(x) - f^0(x)| < \eta$, and $l'_1(\cdot, \cdot)$, $l''_{11}(\cdot, \cdot)$ represent the first and second order partial derivatives of l with respect to its first argument. Assuming this is valid for all $x \in \mathcal{X}$ implies that

$$l(f(x), x) - l(f^0(x), x) \geq c(f(x) - f^0(x))^2, \quad (13)$$

for all $f \in \mathbf{F}_{\text{local}} = \{f \in \mathbf{F} : \|f - f^0\|_\infty \leq \eta\}$. This yields, using that the $(X_i, Y_i)_{i=1}^n$ are identically distributed,

$$\Xi(f) = P\rho_f - P\rho_{f^0} = [E\rho_f(X_1, Y_1) - E\rho_{f^0}(X_1, Y_1)] = E[l(f(X_1), X_1) - l(f^0(X_1), X_1)] \geq c\|f - f^0\|^2,$$

for all $f \in \mathbf{F}_{\text{local}}$ such that the margin condition is satisfied with $G(x) = cx^2$. Put differently, the above equality shows that it suffices to establish a lower bound on the second derivative of the conditional expectation of the loss function in order to show that the margin condition holds with quadratic margin.

We next use the above result to verify that some typical loss functions encountered in econometrics satisfy Assumptions 1-3.

5.1 Quadratic loss

Assume that the data is generated by the i.i.d. sequence

$$Y_i = f^0(X_i) + \epsilon_i \quad (14)$$

for $X_i \in \mathcal{X} \subseteq \mathbb{R}$.⁷ We show that the quadratic loss function

$$\rho(f(x), y) = (y - f(x))^2, \quad (15)$$

⁷It is not difficult to generalize this to \mathcal{X} being some subset of a normed space by modifying the definition of sub-gaussianity in footnote 8 below slightly.

for $f \in \mathbf{F}$ can be encompassed by our general theory. The quadratic loss function is probably the most widely used loss in regression analysis. The main obstacle in fitting this type of loss into our general theory is that Theorem 2 and Corollary 1 rely on the loss function being Lipschitz continuous in order to lower bound the probability of the event τ . However, the quadratic loss is only locally Lipschitz continuous. Nevertheless, this can still deliver an oracle inequality which holds with high probability if both the error term and covariates have enough moments and the target function f^0 is bounded on compact subsets of \mathbb{R} . Before stating the first result note that for any $f \in \mathbf{F}$, if ϵ_1 is independent of X_1

$$\Xi(f) = E(Y_1 - f(X_1))^2 - E(Y_1 - f^0(X_1))^2 = E(f(X_1) - f^0(X_1))^2, \quad (16)$$

such that the excess loss reduces to the mean square error (MSE) in case of a quadratic loss function. (16) also reveals that the margin condition is satisfied with a quadratic margin, even without using the technique from above, which implies that Assumption 1 is satisfied.

Lemma 1. *Assume that $\max_{1 \leq j \leq p} \|\psi_j\|_\infty \leq K$, set $\lambda_2 = \frac{\lambda_1 \sqrt{s^*}}{2\|\beta^*\|_2}$ and suppose that, for some $L \geq 8$, $8\lambda_0 \leq \lambda_1 \leq L\lambda_0$ and $\frac{36L^2\lambda_0 s^* K}{\eta\phi_*^2} \leq 1$. Also, assume that $\|f^* - f^0\|_\infty \leq \eta/2$, with $E(f^* - f^0)^2 \leq \frac{6\lambda_1^2 s^*}{\phi_*^2}$. Finally, assume that the population covariance matrix of the covariates $\Sigma = E(\psi(X_1)\psi'(X_1))$ has full rank. Then,*

a) *Assumptions 1-3 are valid and on the set τ*

$$E(\hat{f} - f^0)^2 + \lambda_1 \|\hat{\beta} - \beta^*\|_1 + \lambda_2 \|\hat{\beta} - \beta^*\|_2^2 \leq 6E(f^* - f^0)^2 + \frac{36L^2\lambda_0^2 s^*}{\phi_*^2}. \quad (17)$$

b) *if, $\max_{1 \leq j \leq p} E\psi_j^2(X_1) \leq 1$, $\sup_{|x| \leq C_n} |f^0(x)| \leq F_{C_n} < \infty$ for all $C_n > 0$, $\Phi = \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq G < \infty\}$ and X_1, ϵ_1 are sub-gaussian⁸ one has for all $C_n > 0$ that (17) is valid with probability at least $1 - \left(\frac{1}{p}\right) - 2\alpha n \exp(-\delta C_n^2)$ for $\lambda_0 = dD_n \sqrt{\frac{\log(p)}{n}}$ with $D_n = 2(C_n + 2F_{C_n} + GK)$ and $d > 0$.*

Lemma 1 a) provides sufficient conditions for Assumptions 1-3, and a bound similar to the one in Theorem 1 which is valid on the set τ . Lemma 1 b) also provides a lower bound on the probability of the set τ in the case of a non-Lipschitz continuous loss function. Some of the assumptions of Lemma 1 may still seem rather high level but they are not very restrictive. We next give a concrete example of when they are satisfied. We shall assume that x has support in $[-1, 1]$, which can be achieved by a sigmoidal transformation to the covariates. f^0 will be assumed to belong to a Hölder

⁸A real random variable V is said to be sub-gaussian if $P(|V| \geq x) \leq \alpha \exp(-\delta x^2)$ for some positive constants α and δ .

class of order $1/2 < r < \infty$ (or in less standard terminology, it is r -smooth as in Chen (2007)⁹). In this case one may choose \mathcal{F}_L to consist of p th degree polynomials, i.e. $\psi_j(x) = x^j$, $j = 1, \dots, p$, and $\Gamma = \{1, \dots, s^*\}$. We use polynomials as an example, but other basis function can also be used with no loss of generality (see Chen (2007)).

The assumption $\max_{1 \leq j \leq p} \|\psi_j\|_\infty \leq K$ requires the basis functions ψ_j to be bounded. Since the covariates take values on $\mathcal{X} = [-1, 1]$ it follows that $\|\psi_j\|_\infty = \|x^j\|_\infty = 1$ for all $j = 1, \dots, p$ and hence the assumption is satisfied with $K = 1$. Moreover, it also implies that $\max_{1 \leq j \leq p} E\psi_j^2(X_1) \leq 1$ which is needed in part b) of Lemma 1.

The approximation requirements $\|f^* - f^0\|_\infty \leq \eta/2$ and $E(f^* - f^0)^2 \leq \frac{6\lambda_1^2 s^*}{\phi_*^2}$ state that the target should be approximated well by the oracle and imply that $f^* \in \mathbf{F}_{\text{local}}$. As explained previously, one may choose $f^* = f^0$ when f^0 is linear. Hence, the two approximation requirements are trivially satisfied in the linear case. In general, these assumptions depend on the target class of functions \mathbf{F} and the choice of basis functions.

The condition (1) is not very restrictive either. Assuming that ϕ_* is bounded away from 0 it suffices to show that $\lambda_0 s^* \rightarrow 0$. This is a restriction on the sparsity. Assuming Σ to have full rank is reasonably innocent as discussed in Section 2.

The additional assumptions of part b) of Lemma 1 are used to establish a lower bound on the probability of τ on which (17) is valid in the absence of Lipschitz continuity of the loss function. First, $\max_{1 \leq j \leq p} E\psi_j^2(X_1) \leq 1$, imposes a further boundedness assumption on the basis functions, which, as discussed above, is trivially satisfied since $K = 1$. The assumption $\sup_{|x| \leq C_n} |f^0(x)| \leq F_{C_n} < \infty$ requires f^0 to be locally bounded and is satisfied if f^0 is continuous. Continuity of f^0 is ensured if the target is linear or r -smooth for $r > 0$ (the latter being covered by our working example). Hence, our theory covers the case of a linear target with quadratic loss. If the variables are bounded, F_{C_n} is also bounded independently of the sample size. The assumption $\|\beta\|_1 \leq G$ is rather innocent since G can be chosen arbitrarily large¹⁰. The sub-gaussianity assumption on the error term and covariates enforces light tails and is recurrent in the high dimensional regression literature. Bounded random variables are always sub-gaussian.

In conclusion, all conditions of Lemma 1 are valid when f^0 belongs to a Hölder class of order at least $1/2$ and the covariates have support $[-1, 1]$. In particular, it is sufficient to take \mathcal{F}_L as the

⁹Following Chen (2007), we say that a function f is r -smooth if it has derivatives up to order $\lfloor r \rfloor$ and the r -th derivative is Hölder continuous with an exponent of $r - \lfloor r \rfloor$. Here $\lfloor r \rfloor$ denotes the greatest integer strictly less than r . This is also often called a Hölder class of order r , see eg Tsybakov (2009).

¹⁰It is also straightforward to generalize to the setting where G is a sequence depending on the sample size.

family of all p -dimensional polynomials, and $\Gamma = \{1, \dots, s^*\}$.

Remark: Even though the sub-gaussianity assumptions on the covariates and the error terms is standard, we would like to stress that our theory is also applicable for much heavier tails. A version of Lemma 1 b) can be developed even when we only have $E|X_1|^r, E|\epsilon_1|^r \leq \kappa < \infty$ for some $r \geq 2$. As long as the covariates and error terms posses enough moments Lemma 1 b) still applies. More precisely, a slight change of the last part of the proof of part b) yields that $1 - \left(\frac{1}{p}\right) - \frac{\kappa n}{C_n^r}$. So the price to pay for the increased generality is that the last term in the lower bound on the probability no longer tends to 0 exponentially fast in C_n such that C_n now has to be at least of the order $O(n^{1/r})$.

5.1.1 Asymptotic results for quadratic loss

We now return to the general setting of Lemma 1. To illustrate the usefulness of Lemma 1 we remark that it can be used to establish the following asymptotic result.

Corollary 5. *Let the assumptions of Lemma 1 be satisfied. Also, assume that $p \in O(\exp(n^a))$ and $s^* \in O(n^b)$ for some $a > 0$ and $b \geq 0$, and let $F_{C_n} \in O(n^{\tilde{d}})$ for some $\tilde{d} \geq 0$. Choose $C_n = \sqrt{\frac{2}{\delta} \log(n)}$ ¹¹. Then, if ϕ_*^2 is bounded away from 0, one has with probability tending to one,*

a)

$$\limsup_{n \rightarrow \infty} E(\hat{f} - f^0)^2 \leq 6 \limsup_{n \rightarrow \infty} E(f^* - f^0)^2,$$

if $a + b + 2\tilde{d} < 1$.

b)

$$|\hat{\beta} - \beta^0| = \|\hat{\beta} - \beta^0\|_1 \rightarrow 0,$$

if the target f^0 is linear and $a + 2\tilde{d} < 1$.

The assumption $F_{C_n} \in O(n^{\tilde{d}})$ is not overly restrictive since even when $f^0(x) = \exp(\mu x^2)$ for some $\mu > 0$ one has that $F_{C_n} = \sup_{|x| \leq C_n} |f^0(x)| = \exp(2\frac{\mu}{\delta} \log(n)) = n^{2\frac{\mu}{\delta}}$. So the assumption of a polynomial growth of F_{C_n} can be satisfied even by functions increasing exponentially fast by choosing $\tilde{d} = 2\frac{\mu}{\delta}$. In the case where we only assume that the covariates and the error terms have bounded r th moments a similar argument shows that f^0 can increase at the rate of an r th degree polynomial. We have assumed in Corollary 5 that the assumptions of Lemma 1 are valid. Then we

¹¹Where $\delta > 0$ is the subgaussianity constant of footnote 8.

may choose \mathcal{F}_L to consist of polynomials of degree p and $\Gamma = \{1, \dots, s\}$ to meet the assumptions of Lemma 1.

5.2 Logistic regression

Next we verify that the logistic loss satisfies Assumptions 1-3. This is done by verifying that the second derivative with respect to the first argument of the conditional loss function is bounded away from zero as discussed in the beginning of Section 5. Let $(Y_i, X_i)_{i=1}^n$ be i.i.d. As seen in Section 2 the loss in case of logistic regression is

$$\rho(f(x), y) = -yf(x) + \log(1 + \exp(f(x))). \quad (18)$$

Here y, x represent the values of the random variable Y_i and X_i . Then we have the following result.

Lemma 2. *Assume that $\epsilon_0 \leq \pi(x) \leq 1 - \epsilon_0$, for all $x \in \mathcal{X}$ for some $0 < \epsilon_0 \leq 1/2$, and $\max_{1 \leq j \leq p} \|\psi_j\|_\infty \leq K$. Set $\lambda_2 = \frac{\lambda_1 \sqrt{s^*}}{2\|\beta^*\|_2}$ and take for some constant $L \geq 8$, $8\lambda_0 \leq \lambda_1 \leq L\lambda_0$ and suppose that*

$$\frac{36L^2\lambda_0 s^* K}{\eta\phi_*^2 c} \leq 1,$$

for $c = \frac{\epsilon_0 e^{-\eta}}{2(1+e^\eta/\epsilon_0)^2}$. Assume that $\|f^ - f^0\|_\infty \leq \eta/2$, with $\frac{3}{2}\Xi(f^*) \leq \frac{9\lambda_1^2 s^*}{\phi_*^2 c}$ and finally that the population covariance matrix of the covariates $\Sigma = E(\psi(X)\psi'(X))$ has full rank. Then,*

a) Assumptions 1-3 are valid and on the set τ ,

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^*\|_1 + \lambda_2 \|\hat{\beta} - \beta^*\|_2^2 \leq 6\Xi(f^*) + \frac{36\lambda_1^2 s^*}{\phi_*^2 c}. \quad (19)$$

b) if, furthermore, $\frac{1}{n} \sum_{i=1}^n \max_{1 \leq j \leq p} E\psi_j^2(X_i) \leq 1$ one has that (19) is valid with probability at least $1 - \left(\frac{1}{p}\right)$ upon choosing $\lambda_0 = 2d\sqrt{\frac{\log(p)}{n}}$ (where d is the positive constant from Theorem 2).

Note that except for $\epsilon_0 \leq \pi(x) \leq 1 - \epsilon_0$ for all $x \in \mathcal{X}$ the assumptions of Lemma 2 are a subset of the ones in Lemma 1. Hence, we know that they are satisfied when the covariates have support $[-1, 1]$ and f^0 belongs to a Hölder class of order $r > 1/2$ since then we may choose $\psi_j(x) = x^j$, $j = 1, \dots, p$ and $\Gamma = \{1, \dots, s\}$ as discussed after Lemma 1.

(19) in Lemma 2 gives upper bounds on the excess risk as well as the estimation error in the logit model. Corollaries 2-4 can now be used to establish asymptotic results for these quantities.

By the definition of λ_0 it is not difficult to see that the second term on the right hand side tends to zero if $s^* \in o(n/\log(p))$ and ϕ_* is bounded away from zero. This in turn reveals that $\Xi(\hat{f})$

is of the same order of magnitude as the excess risk of the oracle $\Xi(f^*)$ which parallels part a) in Corollary 5. Mimicking the arguments in part b) of that Corollary one sees that

$$\|\hat{\beta} - \beta^0\|_1 \rightarrow 0,$$

with probability tending to one if ϕ_*^2 is bounded away from 0 and the target f^0 is linear as long as $s^* \log(p)/n \rightarrow 0$.

6 Conclusion

This paper has established finite sample, oracle bounds for the penalized empirical loss minimization procedure, for convex loss and elastic net penalty. We have also seen that the results for a Lasso penalty are a special case of ours. For the case where the target is linear the oracle inequality can be used to deduce finite sample upper bounds on the estimation error of $\hat{\beta}$. The oracle inequality can also be used to show that the excess loss of our estimator is asymptotically of the same order as that of the oracle. Also, when the target is linear we give sufficient conditions for $\hat{\beta}$ to be consistent for β^0 . To illustrate our framework we give two examples of settings which fit into our theory – the quadratic and the logistic loss.

Future avenues of research include, but are not limited to, proposing theoretically justified data driven methods for the choice of tuning parameters in the case where the truth is not linear. For these, good references are Koltchinskii (2011) and Massart (2007). Also, bounds which are valid for dependent data are interesting for time series analysts. It would be of potential interest to derive results which illustrate the tradeoff between the Lasso and the elastic net as was done by Hebiri and van de Geer (2011) in the case of a linear model with quadratic loss. However, as it is mentioned in the introduction this does not seem to be a trivial task.

Appendix

We start by proving Theorem 1. As a by-product of the proof we provide a version of the basic inequality (6.28) of Bühlmann and van de Geer (2011) which covers the elastic net using a different proof technique. In particular, using $\phi^2(S^*) > 0$, instead of a compatibility condition allows us to prove Theorem 1 directly without considering several subcases separately. The presence of the squared ℓ_2 -norm in the penalty of the elastic net complicates the proof compared to the one for the plain Lasso since the penalty is no longer a norm.

Throughout the appendix we shall let S_c denote the complement of the set S for any set S ¹².

Proof of Theorem 1. This step uses the convexity of the loss function ρ_f . First, define for $0 \leq t \leq 1$

$$\tilde{\beta} = t\hat{\beta} + (1-t)\beta^*,$$

where

$$t = \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1}.$$

To simplify notation define $\tilde{\Xi} = \Xi(f_{\tilde{\beta}})$ and $\Xi^* = \Xi(f_{\beta^*})$ and set $\tilde{f} = f_{\tilde{\beta}}$, $\hat{f} = f_{\hat{\beta}}$, $f^* = f_{\beta^*}$. Furthermore, define

$$\text{pen}_\lambda(f_\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2.$$

Note that the objective function can be written as $R_n(f_\beta) = P_n(\rho_{f_\beta}) + \text{pen}_\lambda(f_\beta)$ and that $R_n(f_\beta)$ is convex. By the minimizing property of $\hat{\beta}$ one has that

$$P_n \rho_{\hat{f}} + \lambda_1 \|\hat{\beta}\|_1 + \lambda_2 \|\hat{\beta}\|_2^2 \leq P_n \rho_{f^*} + \lambda_1 \|\beta^*\|_1 + \lambda_2 \|\beta^*\|_2^2. \quad (20)$$

By the convexity of $\beta \mapsto \rho_{f_\beta}$ and the linearity of the P_n -integral it follows that

$$\begin{aligned} P_n \rho_{\tilde{f}} + \lambda_1 \|\tilde{\beta}\|_1 + \lambda_2 \|\tilde{\beta}\|_2^2 &\leq t P_n \rho_{\hat{f}} + (1-t) P_n \rho_{f^*} + t \lambda_1 \|\hat{\beta}\|_1 + (1-t) \lambda_1 \|\beta^*\|_1 + t \lambda_2 \|\hat{\beta}\|_2^2 + (1-t) \lambda_2 \|\beta^*\|_2^2 \\ &\leq P_n \rho_{f^*} + \lambda_1 \|\beta^*\|_1 + \lambda_2 \|\beta^*\|_2^2, \end{aligned} \quad (21)$$

where the second inequality follows from (20). Rearranging (21), define $A(f^*)$ and $B(f^*)$ as

$$A(f^*) := [P_n(\rho_{\tilde{f}} - \rho_{f^*})] \leq \text{pen}_\lambda(f^*) - \text{pen}_\lambda(\tilde{f}) =: B(f^*). \quad (22)$$

The next steps involve finding a lower bound on $A(f^*)$, an upper bound on $B(f^*)$ and use these bounds to get the oracle inequality. First, we start by finding a lower bound on $A(f^*)$. To that end, note that for any integrable random variables X, Y, Z

$$P_n(X - Y) = (P_n - P)(X) - (P_n - P)(Y) + P(X - Z) - P(Y - Z). \quad (23)$$

Now recall that for any $f \in \mathbf{F}$, the excess risk is defined as

$$\Xi(f) = P(\rho_f - \rho_{f^0}), \quad (24)$$

and

$$V_n(f) = (P_n - P)(\rho_f). \quad (25)$$

¹²It will be clear from the context of which set S is a subset.

Use $X = \rho_{\tilde{f}}$, $Y = \rho_{f^*}$, $Z = \rho_{f^0}$ in (23) and (24)-(25) to get

$$A(f^*) = V_n(\tilde{f}) - V_n(f^*) + \Xi(\tilde{f}) - \Xi(f^*). \quad (26)$$

By assumption

$$\sup_{\beta \in \{\|\beta - \beta^*\|_1 \leq M^*\}} |V_n(f_\beta) - V_n(f^*)| \leq \lambda_0 M^*. \quad (27)$$

Linearity of f_β implies a bijection between f_β and β . Note that

$$\|\tilde{\beta} - \beta^*\|_1 = \|t(\hat{\beta} - \beta^*)\|_1 = \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1} \|\hat{\beta} - \beta^*\|_1 \leq M^*.$$

This means that any M^* is suitable as far as Assumption 2 holds and $\tilde{\beta}$ is among the β the supremum is taken over in (27). So via (26) (27)

$$A(f^*) \geq -\lambda_0 M^* + \Xi(\tilde{f}) - \Xi(f^*). \quad (28)$$

Inequality (28) will serve as the lower bound on $A(f^*)$. Next, we consider the upper bound on $B(f^*)$.

We start by analyzing the terms in our penalty functions. From the reverse triangle inequality and $\beta_{S_c^*}^* = 0$,

$$\|\beta^*\|_1 - \|\tilde{\beta}\|_1 \leq \|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_1 - \|\tilde{\beta}_{S_c^*}\|_1. \quad (29)$$

Furthermore, by continuity of the norm

$$0 \leq \|\beta_{S^*}^*\|_2 - \|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_2 \leq \|\tilde{\beta}_{S^*}\|_2$$

Squaring both sides, using that $\|\beta_{S^*}^*\|_2 = \|\beta^*\|_2$ and rearranging yields that

$$\begin{aligned} \|\beta^*\|_2^2 - \|\tilde{\beta}\|_2^2 &\leq 2\|\beta_{S^*}^*\|_2 \|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_2 \\ &\quad - \|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_2^2 - \|\tilde{\beta}_{S_c^*}\|_2^2 \\ &= 2\|\beta_{S^*}^*\|_2 \|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_2 - \|\tilde{\beta} - \beta^*\|_2^2, \end{aligned} \quad (30)$$

where for the last equality we use $\|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_2^2 + \|\tilde{\beta}_{S_c^*}\|_2^2 = \|\tilde{\beta} - \beta^*\|_2^2$. Next, using Section 2.2, $\|f_\beta\|^2 = \beta' \Sigma \beta \geq \phi(S)^2 \|\beta\|_2^2$. Then, by Assumption 3

$$\|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_2 \leq \|\tilde{f} - f^*\| / \phi(S^*). \quad (31)$$

Also, we have

$$\|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_1 \leq \sqrt{s^*} \|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_2. \quad (32)$$

Recall that

$$\begin{aligned}
B(f^*) &= \text{pen}_\lambda(f^*) - \text{pen}_\lambda(\tilde{f}) \\
&= \lambda_1(\|\beta^*\|_1 - \|\tilde{\beta}\|_1) \\
&\quad + \lambda_2(\|\beta^*\|_2^2 - \|\tilde{\beta}\|_2^2).
\end{aligned} \tag{33}$$

Next, using (29) and (30) on the right side of (33)

$$\begin{aligned}
B(f^*) &\leq \lambda_1\|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_1 - \lambda_1\|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_1 \\
&\quad + 2\lambda_2\|\beta_{S^*}^*\|_2\|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_2 - \lambda_2\|\tilde{\beta} - \beta^*\|_2^2.
\end{aligned} \tag{34}$$

Now add and subtract the term $\lambda_1\|\tilde{\beta}_{S^*} - \beta_{S^*}^*\|_1$ from the right side of (34) and use (31)-(32) to get

$$B(f^*) \leq [2\lambda_1\sqrt{s^*} + 2\lambda_2\|\beta_{S^*}^*\|_2] \frac{\|\tilde{f} - f^*\|}{\phi(S^*)} - \text{pen}_\lambda(f_{\tilde{\beta} - \beta^*}), \tag{35}$$

where we used that $\text{pen}_\lambda(f_{\tilde{\beta} - \beta^*}) = \lambda_1\|\tilde{\beta} - \beta^*\|_1 + \lambda_2\|\tilde{\beta} - \beta^*\|_2^2$.

Since we assume $f^* \in \mathbf{F}_{\text{local}}$, and since $f_{\tilde{\beta}} \in \mathbf{F}_{\text{local}}$ (because $\|\tilde{\beta} - \beta^*\|_1 \leq M^*$ as shown above) by Assumption 3, we get using (1) and the margin condition that

$$\begin{aligned}
(2\lambda_1\sqrt{s^*} + 2\lambda_2\|\beta^*\|_2) \frac{\|\tilde{f} - f^*\|}{\phi(S^*)} &= \frac{\|\tilde{f} - f^*\|}{2} \left[\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)} \right] \\
&\leq G\left(\frac{\|\tilde{f} - f^*\|}{2}\right) + H\left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)}\right) \\
&\leq \Xi(\tilde{f})/2 + \Xi(f^*)/2 + H\left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)}\right),
\end{aligned} \tag{36}$$

where we use (1) for the first inequality, with $u = \frac{\|\tilde{f} - f^*\|}{2}$ and $v = \frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)}$, and the second estimate uses the triangle inequality, convexity of $G(\cdot)$ and the margin condition. Substitute (36) into (35) to have

$$B(f^*) \leq \Xi(\tilde{f})/2 + \Xi(f^*)/2 + H\left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)}\right) - \text{pen}_\lambda(f_{\tilde{\beta} - \beta^*}). \tag{37}$$

Since we have both the upper bound on $B(f^*)$, equation (37), and the lower bound on $A(f^*)$, equation (28), we can combine them via (22) to get

$$-\lambda_0 M^* + \Xi(\tilde{f}) - \Xi(f^*) \leq \Xi(\tilde{f})/2 + \Xi(f^*)/2 + H\left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)}\right) - \text{pen}_\lambda(f_{\tilde{\beta} - \beta^*}). \tag{38}$$

We can rewrite (38) as

$$\Xi(\tilde{f})/2 + \text{pen}_\lambda(f_{\tilde{\beta}-\beta^*}) \leq \lambda_0 M^* + \frac{3}{2}\Xi(f^*) + H \left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)} \right). \quad (39)$$

Next, we use the definition $\Delta^* = \lambda_0 M^* = \frac{3}{2}\Xi(f^*) + H \left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)} \right)$ to have

$$\Xi(\tilde{f})/2 + \lambda_1\|\tilde{\beta} - \beta^*\|_1 + \lambda_2\|\tilde{\beta} - \beta^*\|_2^2 \leq 2\Delta^*, \quad (40)$$

by $\text{pen}_\lambda(f_{\tilde{\beta}-\beta^*}) = \lambda_1\|\tilde{\beta} - \beta^*\|_1 + \lambda_2\|\tilde{\beta} - \beta^*\|_2^2$.

The inequality (40) yields the desired oracle inequality but for $\tilde{\beta}$ instead of $\hat{\beta}$. However, it also follows from (40) (using $\Delta^* = \lambda_0 M^*$) that

$$\lambda_1\|\tilde{\beta} - \beta^*\|_1 \leq 2\Delta^* = 2\lambda_0 M^*,$$

which in turn yields (using $\lambda_1 \geq 4\lambda_0$)

$$\|\tilde{\beta} - \beta^*\|_1 \leq 2\frac{\lambda_0}{\lambda_1} M^* \leq M^*/2.$$

Next, note that by the definitions of $\tilde{\beta}$ and t

$$\tilde{\beta} - \beta^* = t\hat{\beta} + (1-t)\beta^* - \beta^* = t(\hat{\beta} - \beta^*) = \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1}(\hat{\beta} - \beta^*).$$

So

$$\frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1} \|\hat{\beta} - \beta^*\|_1 = \|\tilde{\beta} - \beta^*\|_1 \leq M^*/2,$$

which upon rearranging yields $\|\hat{\beta} - \beta^*\|_1 \leq M^*$. But this means that all the above derivations are valid with $\hat{\beta}$ replacing $\tilde{\beta}$ by simply starting from (20) instead of (21). In particular, (40) yields

$$\Xi(\hat{f}) + 2\lambda_1\|\hat{\beta} - \beta^*\|_1 + 2\lambda_2\|\tilde{\beta} - \beta^*\|_2^2 \leq 4\Delta^* = 6\Xi(f^*) + 4H \left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi(S^*)} \right), \quad (41)$$

which implies the bound in Theorem 1.

□

Proof of Theorem 2. Set

$$\zeta = D \left[4\Lambda \left(\frac{K}{3}, n, p \right) + \frac{tK}{3n} + \sqrt{\frac{2t}{n}} \sqrt{1 + 8\Lambda \left(\frac{K}{3}, n, p \right)} \right] \quad (42)$$

with $\Lambda(\frac{K}{3}, n, p) = \sqrt{\frac{2\log 2p}{n}} + \frac{K\log 2p}{3n}$. Then, for all $t > 0$ Bühlmann and van de Geer (2011) show (Theorem 14.5) that

$$P(Z_M \leq M\zeta) \geq 1 - \exp(-t).$$

Next note that there exist a constant $c > 0$ (whose value may change throughout the display below) such that

$$\Lambda\left(\frac{K}{3}, n, p\right) \leq c \left(\sqrt{\frac{\log(2)}{n}} + \sqrt{\frac{\log(p)}{n}} + \frac{\log(2)}{n} + \frac{\log(p)}{n} \right) \leq c \sqrt{\frac{\log(p)}{n}}.$$

Hence, choosing $t = \log(p)$ implies that there exists a constant $\tilde{c} > 0$ (whose value may change throughout the display below) such that

$$\zeta \leq \tilde{c}D \left[\sqrt{\frac{\log(p)}{n}} + \frac{\log(p)}{n} + \sqrt{\frac{\log(p)}{n}} + \left(\frac{\log(p)}{n} \right)^{3/4} \right] \quad (43)$$

$$\leq \tilde{c}D \sqrt{\frac{\log(p)}{n}} := \lambda_0. \quad (44)$$

This implies

$$P(\tau) = P(Z_{M^*} \leq \lambda_0 M^*) \geq P(Z_{M^*} \leq \zeta M^*) \geq 1 - \exp(-\log(p)) = 1 - \left(\frac{1}{p}\right).$$

□

Proof of Theorem 1. The bound in Theorem 1 is valid on the set τ . Theorem 2 provides the stated lower bound on the probability of τ . Combining these two results gives the theorem.

□

Lemma 3. *Let $G(u)$ be a strictly convex function on $[0, \infty)$, with $G(0) = 0$. The convex conjugate $H(v) = \sup_{u \geq 0} [uv - G(u)]$, $v \geq 0$ satisfies*

1. *H is non-negative and non-decreasing.*

2. *H is convex.*

3. *H is right-continuous at 0.*

Proof. The non-negativity of H follows from $H(v) \geq [0 \cdot v - G(0)] = 0$ for all $v \geq 0$. Let $0 \leq v_1 \leq v_2$. Then, since $[uv_1 - G(u)] \leq [uv_2 - G(u)]$ for all $u \geq 0$,

$$H(v_1) = \sup_{u \geq 0} [uv_1 - G(u)] \leq \sup_{u \geq 0} [uv_2 - G(u)] = H(v_2),$$

and so H is non-decreasing.

The convexity of H may be found in Theorem 12.2 of Rockafellar (1997). Here, for the sake of completeness, we give a more direct argument. For any $0 < \lambda < 1$ and $v_1, v_2 \geq 0$

$$\begin{aligned} H(\lambda v_1 + (1 - \lambda)v_2) &= \sup_{u \geq 0} [u(\lambda v_1 + (1 - \lambda)v_2) - G(u)] = \sup_{u \geq 0} [\lambda(uv_1 - G(u)) + (1 - \lambda)(uv_2 - G(u))] \\ &\leq \lambda \sup_{u \geq 0} (uv_1 - G(u)) + (1 - \lambda) \sup_{u \geq 0} (uv_2 - G(u)) = \lambda H(v_1) + (1 - \lambda)H(v_2), \end{aligned}$$

establishing the convexity of H .

To establish that H is right continuous at 0 note first that $H(0)$ is a lower bound for $\{H(x_n)\}_{n=1}^\infty$ for any sequence $x_n \downarrow 0$ since H is non-decreasing. Hence, $\{H(x_n)\}_{n=1}^\infty$ is a bounded non-increasing sequence and so it possesses a limit which furthermore satisfies $H(0) \leq \inf_n H(x_n) = \lim_n H(x_n)$. It suffices to show that $H(0) \geq \inf_{x>0} H(x) = \inf_n H(x_n)$ to conclude $H(0) = \inf_n H(x_n) = \lim_n H(x_n)$. We assume the converse to reach a contradiction., i.e. assume that $H(0) < \inf_{x>0} H(x)$. In particular, $H(0) < \inf_{0<\lambda<1} H((1 - \lambda)x)$ for all $x > 0$ such that there exists an $\epsilon > 0$ satisfying $\inf_{0<\lambda<1} H((1 - \lambda)x) = H(0) + \epsilon$. But by the convexity of H it holds for all $0 < \lambda < 1$ that

$$\lambda H(0) + (1 - \lambda)H(x) \geq H((1 - \lambda)x) \geq H(0) + \epsilon.$$

By continuity of the left hand side in λ it follows that

$$H(0) = \lim_{\lambda \uparrow 1} [\lambda H(0) + (1 - \lambda)H(x)] \geq H(0) + \epsilon,$$

which is a contradiction and so we can't have $H(0) < \inf_{x>0} H(x)$ and we conclude that H is right-continuous at 0. \square

Proof of Corollary 2. First, note that the probability with which the inequality in Theorem 1 is valid tends to one. Also, this inequality implies

$$\Xi(\hat{f}) \leq 6\Xi(f^*) + 4H\left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi_*}\right).$$

Hence, it suffices to show that $\limsup_{n \rightarrow \infty} H\left(\frac{4\lambda_1\sqrt{s^*} + 4\lambda_2\|\beta^*\|_2}{\phi_*}\right) = \limsup_{n \rightarrow \infty} H\left(\frac{6\lambda_1\sqrt{s^*}}{\phi_*}\right) = 0$.

To this end, observe that with $\lambda_1 \in O(\lambda_0)$

$$\frac{6\lambda_1\sqrt{s^*}}{\phi_*} \in O\left(\sqrt{\frac{n^a}{n}n^b}\right) = O\left(n^{a/2+b/2-1/2}\right) \subseteq o(1),$$

which yields the desired result by the-right continuity of H established in Lemma 3. \square

Proof of Corollary 3. The first two inequalities, (9) and (10), follow from Theorem 1 upon using the same reasoning as in remark 2 preceding Theorem 1. In particular, choose $\Gamma = S^0$ in the definition of the oracle. This implies $\Xi(f^*) = \Xi(f^0) = 0$ (as seen in remark 2). (11) and (12) follows from (9) and (10) under the given assumptions by simple insertion. \square

Proof of Corollary 4. First note that the probability with which inequality (12) is valid tends to one. It remains to be shown that the right hand side of (12) tends to zero. But under the stated conditions the right hand side is of order

$$O\left(\sqrt{\frac{n^a}{n}}n^b\right) = O\left(n^{a/2+b-1/2}\right) \subseteq o(1),$$

where the last inclusion follows from the assumption $a + 2b < 1$. \square

Proof of Lemma 1. a) *The Analysis of Assumption 1.* First, note that by the second derivative test $\rho(f(x), y)$ is a convex function in $f(x)$. Since, the second derivative is constant, and equal to 2 this also shows that the margin condition is satisfied with a quadratic margin and $2c = 2$, i.e. $c = 1$. This is of course already clear from (16) prior to Lemma 1. The analysis of Assumption 2 is slightly more involved:

The Analysis of Assumption 2. We show that $f_\beta \in \mathbf{F}_{\text{local}}$ under the stated conditions. More precisely, we must show that $\|f_\beta - f^0\|_\infty \leq \eta$. Since

$$\|f_\beta - f^0\|_\infty \leq \|f_\beta - f^*\|_\infty + \|f^* - f^0\|_\infty \leq \|f_\beta - f^*\|_\infty + \eta/2,$$

it suffices to show that $\|f_\beta - f^*\|_\infty \leq \eta/2$. To this end, note that

$$|f_\beta(x) - f^*(x)| = \left| \sum_{j=1}^p (\beta_j - \beta_j^*) \psi_j(x) \right| \leq \|\beta - \beta^*\|_1 \max_{1 \leq j \leq p} |\psi_j(x)|,$$

which implies $\|f_\beta - f^*\|_\infty \leq M^*K$. Hence, it suffices to show that $M^*K \leq \eta/2$. To do so, recall

that by using $H(v) = v^2/4c$ (with $c = 1$), $\Xi(f^*) = E(f^* - f^0)^2$ and $\lambda_2 = \frac{\lambda_1 \sqrt{s^*}}{2\|\beta^*\|_2}$

$$\begin{aligned}
M^* &= \frac{\Delta^*}{\lambda_0} = \frac{1}{\lambda_0} \left((3/2)E(f^* - f^0)^2 + H \left(\frac{4\lambda_1 \sqrt{s^*} + 4\lambda_2 \|\beta^*\|_2}{\phi(S^*)} \right) \right) \\
&= \frac{1}{\lambda_0} \left((3/2)E(f^* - f^0)^2 + \left(\frac{4\lambda_1 \sqrt{s^*} + 4\lambda_2 \|\beta^*\|_2}{\phi(S^*)} \right)^2 / 4 \right) \\
&= \frac{1}{\lambda_0} \left((3/2)E(f^* - f^0)^2 + \frac{9\lambda_1^2 s^*}{\phi^2(S^*)} \right) \\
&\leq \frac{1}{\lambda_0} \left(\frac{18\lambda_1^2 s^*}{\phi^2(S^*)} \right) \\
&\leq \frac{18L^2 \lambda_0 s^*}{\phi^2(S^*)},
\end{aligned}$$

such that $M^*K \leq \eta/2$ under the stated assumptions.

The Analysis of Assumption 3. The validity of Assumption 3 follows from the fact that Σ is assumed to have full rank. This is sufficient for Assumption 3 to be valid as argued in Section 2.2.

Inequality (17) follows upon using $H(v) = v^2/4c$ with $c = 1$ as well as $\Xi(f) = E(f - f^0)^2$ for all $f \in \mathbf{F}$ in Theorem 1.

b) Next, we turn to part b) of the lemma. This result is derived based on Theorem 2. Hence, we verify the assumptions of that theorem. The two boundedness conditions are valid by assumption. Next, we establish the local Lipschitz continuity. Note that on $\mathcal{A} = \{\max_{1 \leq i \leq n} |X_i| \vee |\epsilon_i| \leq C_n\}$

$$\begin{aligned}
\left| \frac{\partial \rho(f_\beta(X_i), Y_i)}{\partial f_\beta(X_i)} \right| &= 2|Y_i - f_\beta(X_i)| = 2(|\epsilon_i + f^0(X_i) - f_\beta(X_i)|) \leq 2 \left(|\epsilon_i| + F_{C_n} + \left| \sum_{j=1}^p \beta_j \psi_j(x) \right| \right) \\
&\leq 2(|\epsilon_i| + F_{C_n} + \|\beta\|_1 \max_{1 \leq j \leq p} \|\psi_j\|_\infty) \leq 2(C_n + F_{C_n} + GK),
\end{aligned}$$

for all $i = 1, \dots, n$. So, on the set \mathcal{A} , the first derivative of the loss function is bounded and hence the loss function is Lipschitz continuous on this set. This implies that

$$P(\tau^c) \leq P(\tau^c \cap \mathcal{A}) + P(\mathcal{A}^c).$$

By the above arguments $\rho(f(x), y)$ is Lipschitz continuous on \mathcal{A} with Lipschitz constant $2(C_n + F_{C_n} + GK)$. Hence, by Theorem 2 $P(\tau^c \cap \mathcal{A}) \leq \left(\frac{1}{p}\right)$ and the Lipschitz constant D_n in the definition of $\lambda_0 = dD_n \sqrt{\frac{\log(p)}{n}}$ in Theorem 2 may be taken to be $2(C_n + F_{C_n} + GK)$. Next, through subgaussianity of X_1, ϵ_1 , by a union bound it follows that $P(\mathcal{A}^c) \leq 2\alpha n \exp(-\delta C_n^2)$ for positive constants α and δ . This yields the stated lower bound on the probability of τ (on which inequality (17) is valid). \square

Proof of Corollary 5. First note that the choice of C_n and $p \rightarrow \infty$ ensure that the probability with which inequality (17) is valid tends to one. To prove part a) it suffices to show that $\lambda_0^2 s^* \rightarrow 0$ (this follows from (17)) which in turn is implied by

$$(C_n^2 + F_{C_n}^2) \frac{\log(p)}{n} s^* \in O\left((\log(n) + n^{2\tilde{d}}) \frac{n^a}{n} n^b\right) \subseteq o(1),$$

where the first inclusion is by assumption and the second follows from $a + b + 2\tilde{d} < 1$.

Regarding part b), choosing $\Gamma = S^0$ in the definition of the oracle implies $\Xi(f^*) = \Xi(f^0) = 0$ (as seen in remark 2 after Theorem 1) since the best linear predictor of a linear target is just the target itself (in this case we of course also have $s^* = s^0 = 1$ and $\beta^* = \beta^0$). Hence, we deduce from (17) that

$$|\hat{\beta} - \beta^0| = \|\hat{\beta} - \beta^0\|_1 \leq \frac{9L^2\lambda_0}{2\phi_*^2},$$

with probability tending to one for $\lambda_0 = dD_n \sqrt{\frac{\log(p)}{n}}$ with $D_n = 2(C_n + F_{C_n} + GK)$ and $d > 0$. So, it suffices to see that $\lambda_0 \rightarrow 0$ which is implied by

$$(C_n^2 + F_{C_n}^2) \frac{\log(p)}{n} \in O\left((\log(n) + n^{2\tilde{d}}) \frac{n^a}{n}\right) \subseteq o(1),$$

if $a + 2\tilde{d} < 1$. □

Proof of Lemma 2. The proof is similar to the one of Lemma 6.8 in Bühlmann and van de Geer (2011). First, note that $\rho(f(x), y)$ is a convex function in $f(x)$ since it can be written as the sum of convex functions: the first right hand side term in (18) of $\rho(f(x), y)$ is linear in $f(x)$, and the second term has positive second derivative. We start by showing that Assumptions 1-3 are satisfied.

Step 1. *The Analysis of Assumption 1.* We show that one may choose $G(x) = cx^2$ for some positive constant c to be defined precisely below. To do so we follow the general route laid out in the beginning of Section 5. Define

$$l(f(x), x) = E[\rho(f(X), Y) | (X, f(X)) = (x, f(x))] = -\pi(x)f(x) + \log(1 + \exp(f(x))), \quad (45)$$

where $\pi(x) = E(Y | (X, f(X)) = (x, f(x)))$. (45) is minimized with respect to $f \in \mathbf{F}$ at $f^0(x) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$. Hence, $f^0(x) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$. Note that the second order partial derivative of $l(f(x), x)$ with respect to $f(x)$ is

$$\frac{\partial^2 l(a, x)}{\partial a^2} \Big|_{a=f(x)} = \frac{\exp(f(x))}{1 + \exp(f(x))} \left(1 - \frac{\exp(f(x))}{1 + \exp(f(x))}\right) = \frac{\exp(f(x))}{(1 + \exp(f(x)))^2}. \quad (46)$$

So we must show that $\frac{\exp(f(x))}{(1+\exp(f(x)))}$ is bounded from below by a constant for $f \in \mathbf{F}_{\text{local}}$. To do so it suffices to bound $f(x)$ from above and below. To this end, note that for all $x \in \mathcal{X}$

$$f^0(x) - \|f - f^0\|_\infty \leq f(x) \leq f^0(x) + \|f - f^0\|_\infty,$$

which implies that for all $f \in \mathbf{F}_{\text{local}}$

$$f^0(x) - \eta \leq f(x) \leq f^0(x) + \eta. \quad (47)$$

Furthermore, since $f^0(x) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$ and $\epsilon_0 \leq \pi(x) \leq 1 - \epsilon_0$, we get

$$\log\left(\frac{\epsilon_0}{1-\epsilon_0}\right) \leq f^0(x) \leq \log\left(\frac{1-\epsilon_0}{\epsilon_0}\right).$$

Together with (46) and (47) this implies that

$$\frac{\partial^2 l(a, x)}{\partial a^2} \Big|_{a=f(x)} \geq \frac{\frac{\epsilon_0}{1-\epsilon_0} e^{-\eta}}{\left(1 + \frac{1-\epsilon_0}{\epsilon_0} e^\eta\right)^2} \geq \frac{\epsilon_0 e^{-\eta}}{(1 + e^\eta/\epsilon_0)^2} > 0,$$

for all $x \in \mathcal{X}$. Hence, one may use $2c = \frac{\epsilon_0 e^{-\eta}}{(1+e^\eta/\epsilon_0)^2}$.

Step 2. *The Analysis of Assumption 2.* We show that $f_\beta \in \mathbf{F}_{\text{local}}$ under the stated conditions. More precisely, we must show that $\|f_\beta - f^0\|_\infty \leq \eta$. Since

$$\|f_\beta - f^0\|_\infty \leq \|f_\beta - f^*\|_\infty + \|f^* - f^0\|_\infty \leq \|f_\beta - f^*\|_\infty + \eta/2,$$

it suffices to show that $\|f_\beta - f^*\|_\infty \leq \eta/2$. To this end, note that

$$|f_\beta(x) - f^*(x)| = \left| \sum_{j=1}^p (\beta_j - \beta_j^*) \psi_j(x) \right| \leq \|\beta - \beta^*\|_1 \max_{1 \leq j \leq p} |\psi_j(x)|,$$

which implies $\|f_\beta - f^*\|_\infty \leq M^* K$. Hence, it suffices to show that $M^* K \leq \eta/2$. To do so, recall that, by using $H(v) = v^2/4c$, and $\lambda_2 = \frac{\lambda_1 \sqrt{s^*}}{2\|\beta^*\|_2}$

$$\begin{aligned} M^* &= \frac{\Delta^*}{\lambda_0} = \frac{1}{\lambda_0} \left((3/2)\Xi(f^*) + H\left(\frac{4\lambda_1 \sqrt{s^*} + 4\lambda_2 \|\beta^*\|_2}{\phi(S^*)}\right) \right) \\ &= \frac{1}{\lambda_0} \left((3/2)\Xi(f^*) + \left(\frac{4\lambda_1 \sqrt{s^*} + 4\lambda_2 \|\beta^*\|_2}{\phi(S^*)}\right)^2 / 4c \right) \\ &= \frac{1}{\lambda_0} \left((3/2)\Xi(f^*) + \frac{9\lambda_1^2 s^*}{\phi^2(S^*)c} \right) \\ &\leq \frac{1}{\lambda_0} \left(\frac{18\lambda_1^2 s^*}{\phi^2(S^*)c} \right) \\ &\leq \frac{18L^2 \lambda_0 s^*}{\phi^2(S^*)c}, \end{aligned}$$

such that $M^*K \leq \eta/2$ under the stated assumptions.

Step 3. *The Analysis of Assumption 3.* The validity of Assumption 3 follows from the fact that Σ is assumed to have full rank. This is sufficient for Assumption 3 to be valid as argued in Section 2.2. Inequality (19) follows from Theorem 1 upon using $H(v) = v^2/4c$ with $c = \frac{\epsilon_0 e^{-\eta}}{2(1+e^\eta/\epsilon_0)^2}$.

Next, we turn to part b) of Lemma 2. It suffices to verify the assumptions of Theorem 2 since (19) is valid on the set τ . First, note that $\rho(f(x), y)$ is Lipschitz continuous in $f(x)$ for all $y \in \mathcal{Y}$ since

$$\left| \frac{\partial \rho(a, y)}{\partial a} \right|_{a=f(x)} = \left| -y + \frac{e^{f(x)}}{1+f(x)} \right| \leq 2,$$

and so D in 6 may be chosen to be 2. The two boundedness assumptions on the basis functions are valid by assumption. So, using $H(v) = v^2/4c$ in Theorem 1 yields

$$\Xi(\hat{f}) + \lambda_1 \|\hat{\beta} - \beta^*\|_1 + \lambda_2 \|\hat{\beta} - \beta^*\|_2^2 \leq 6\Xi(f^*) + 4H \left(\frac{4\lambda_1 \sqrt{s^*} + 4\lambda_2 \|\beta^*\|_2}{\phi_*} \right) \leq 6\Xi(f^*) + 36 \frac{\lambda_1^2 s^*}{\phi_*^2 c}.$$

□

References

- Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6), 2369–2429.
- Belloni, A. and V. Chernozhukov (2011). High dimensional sparse econometric models: An introduction. *Inverse Problems and High-Dimensional Estimation*, 121–156.
- Belloni, A., V. Chernozhukov, and L. Wang (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* 98(4), 791–806.
- Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 37(4), 1705–1732.
- Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag, New York.
- Caner, M. and X. Han (2014). Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators. *Journal of Business and Economic Statistics* (forthcoming).

- Caner, M. and H. H. Zhang (2014). Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics* (forthcoming).
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* 6, 5549–5632.
- Chen, X. and S. C. Ludvigson (2009). Land of addicts? an empirical investigation of habit-based asset pricing models. *Journal of Applied Econometrics* 24(7), 1057–1093.
- Cheng, X. and Z. Liao (2013). Select the valid and relevant moments: An information-based lasso for gmm with many moments, second version. Technical report, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.
- Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5), 849–911.
- Fan, J., J. Lv, and L. Qi (2011). Sparse high dimensional models in economics. *Annual review of economics* 3, 291–317.
- Hebiri, M. and S. van de Geer (2011). The smooth-lasso and other $l_1 + l_2$ -penalized methods. *Electronic Journal of Statistics* 5, 1184–1226.
- Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics* 36(2), 587–613.
- Kallenberg, O. (2002). *Foundations of modern probability*. Springer.
- Kallestrup-Lamb, M., A. B. Kock, and J. T. Kristensen (2013). Lassoing the determinants of retirement. Technical report, School of Economics and Management, University of Aarhus.
- Kock, A. and L. Callot (2013). Oracle inequalities for high dimensional vector autoregressions. *CREATES Research Paper*, available on *arXiv*.
- Kock, A. B. (2013). Oracle efficient variable selection in random and fixed effects panel data models. *Econometric Theory* 29, 115–152.

- Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Springer-Verlag, New York.
- Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics* 2, 90–102.
- Massart, P. (2007). *Concentration inequalities and model selection*. Springer-Verlag, New York.
- Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Statistical Science* 27(4), 538–557.
- Rockafellar, R. T. (1997). *Convex Analysis*, Volume 28. Princeton University Press.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.
- van de Geer, S. and P. Bühlmann (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* 3, 1360–1392.
- van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* 36(2), 614–645.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476), 1418–1429.
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), 301–320.