On Concentration for (Regularized) Empirical Risk Minimization

Sara van de Geer
Seminar for Statistics, ETH Zürich, Zurich, Switzerland

Martin J. Wainwright
University of California, Berkeley, USA

Abstract

Rates of convergence for empirical risk minimizers have been well studied in the literature. In this paper, we aim to provide a complementary set of results, in particular by showing that after normalization, the risk of the empirical minimizer concentrates on a single point. Such results have been established by Chatterjee (The Annals of Statistics, 42(6):2340–2381 2014) for constrained estimators in the normal sequence model. We first generalize and sharpen this result to regularized least squares with convex penalties, making use of a “direct” argument based on Borell’s theorem. We then study generalizations to other loss functions, including the negative log-likelihood for exponential families combined with a strictly convex regularization penalty. The results in this general setting are based on more “indirect” arguments as well as on concentration inequalities for maxima of empirical processes.

AMS (2000) subject classification. Primary: 62E20; Secondary 60F99.

Keywords and phrases. Concentration, Density estimation, Empirical process, Empirical risk minimization, Normal sequence model, Penalized least squares

1 Introduction

Empirical risk minimization (ERM) is an important methodology in statistics and machine learning, widely used for estimating high-dimensional and/or nonparametric quantities of interest. The idea is to express the quantity of interest as the minimizer of an expected loss, known as the population or theoretical risk. Given that the distribution of data is not known or difficult to assess, one replaces the theoretical expectation by an empirical counterpart defined by samples. The technique of ERM is known under various names, including $M$-estimation and minimum contrast estimation.

By the law of large numbers, empirical averages of various types of random variables, with the i.i.d. setting being the canonical case, are close to their expectations. This elementary fact is the motivation for ERM and
the starting point for studying its theoretical properties. There is much
literature developing the theory for a broad spectrum of estimation problems.
The more recent literature takes a non-asymptotic point of view, in which
context concentration inequalities play a major role. Concentration inequal-
ities describe the amount of concentration of certain (complex) quantities
around their mean. We refer to Talagrand (1995) as a key paper in the
area, and to the books by Ledoux (2001), and more recently by Boucheron
et al. (2013). The key point is that the deviation from the mean is gen-
erally of much smaller order than the mean itself. Moreover, at least in a
certain sense, the deviation does not depend on the complexity of the orig-
inal object. In statistics, the usefulness of concentration inequalities has
been excellently outlined and studied in Massart (2000). We also refer the
reader to Koltchinskii (2011) for an in-depth treatment in the context of
high-dimensional problems.

Some statistical papers address concentration for the parameter of in-
terest itself; for instance, see Boucheron and Massart (2011) and Saumard
(2012). The present paper is along the lines of Chatterjee (2014). The latter
examines the concentration properties of constrained estimators for the
normal sequence model, or alternatively phrased in the regression setting, for
the least-squares problem with fixed design and Gaussian errors. The au-
thor shows that the statistical error of the least squares estimator satisfies a
concentration inequality where the amount of concentration still depends on
the complexity of the problem, but is in the nonparametric case of smaller
order than the statistical error itself. In Muro and van de Geer (2015),
the situation is studied where a regularization penalty based on a squared
pseudo-norm is added to the least squares loss function. In Section 2, we pro-
vide a “direct” argument for concentration of the regularized least squares in
the normal sequence setting. Our argument here is elementary, using stan-
dard facts from convex analysis (Rockafellar, 1970), and concentration for
Lipschitz functions of Gaussian vectors (Borell, 1975). Our next contribu-
tion is to extend such results to more general problems. The main obstacle
is that the direct concentration for Lipschitz functions holds only for the
Gaussian case. Accordingly, we make use of more general one-sided concen-
tration results for maxima of empirical processes, as given by Klein (2002)
and Klein and Rio (2005).

Our theory allows us to treat a number of new examples in which concen-
tration holds. However, as (asymptotically) exact values for the expectation
of maxima of the empirical process are generally not available, we typically
cannot provide explicit expressions for the point of concentration in terms
of the parameters of the model.
1.1. Set-up and notation 
Consider independent observations \( X_1, \ldots, X_n \) taking values in a space \( \mathcal{X} \), a given class \( \mathcal{F} \) of real-valued functions on \( \mathcal{X} \) and a non-negative regularization penalty \( \text{pen} : \mathcal{F} \to [0, \infty) \). The empirical measure \( P_n \) of a function \( f : \mathcal{X} \to \mathbb{R} \) is defined by the average \( P_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i) \), whereas the theoretical or population measure is given by \( P f := \mathbb{E} P_n f \).

We let \( \mathcal{F} \) denote a class of loss functions indexed by a parameter \( g \) in a parameter space \( \mathcal{G} \). As a concrete example, in the case of least-squares regression, the observations consist of covariates along with real-valued responses of the form \( \{(X_i, Y_i)\}_{i=1}^{n} \). The loss class takes the form

\[
\mathcal{F} = \left\{ f_g(x, y) = (y - g(x))^2 \mid g \in \mathcal{G} \right\},
\]

where \( \mathcal{G} \) is some underlying collection of regression functions.

With this set-up, the regularized empirical risk estimator is defined as

\[
\hat{f} = \text{arg min}_{f \in \mathcal{F}} \left\{ P_n f + \text{pen}(f) \right\},
\]

(1.1)

where \( \text{pen} : \mathcal{F} \to \mathbb{R} \cup \{\infty\} \) is some penalty function. In order to avoid digressions, we assume throughout this paper that “argmin’s” and “argmax’s” exist and are unique in a suitable sense. We define the associated target function \( f^0 := \text{arg min}_{f \in \mathcal{F}} P f \), corresponding to the population minimizer, and we let

\[
\tau^2(f) := \underbrace{P(f - f^0)}_{= \text{excess risk}} + \text{pen}(f)
\]

(1.2)
denote the penalized excess risk.

In order to simplify the exposition, we give asymptotic statements at places, using the classical scaling in which the sample size \( n \) tends to infinity. For a sequence of positive numbers \( \{z_n\}_{n=1}^{\infty} \), we write

\[
z_n = \mathcal{O}(1) \text{ if } \limsup z_n < \infty, \quad \text{and} \quad z_n = o(1) \text{ if } z_n \to 0,
\]
as well as \( z_n \ll 1 \) if both \( z_n = \mathcal{O}(1) \) and \( 1/z_n = \mathcal{O}(1) \). For two positive sequences \( \{y_n\}_{n=1}^{\infty} \) and \( \{z_n\}_{n=1}^{\infty} \), we write \( z_n = \mathcal{O}(y_n) \) if \( z_n/y_n = \mathcal{O}(1) \), along with analogous definitions for the \( o \)- and \( \ll \)-notation. We furthermore use the stochastic order symbols \( \mathcal{O}_\mathbb{P} \) and \( o_\mathbb{P} \). In all our uses of these forms of order notation, the arguments depend on \( n \), but we often omit this dependence so as to simplify notation.
The main results of this paper involve showing that under certain conditions, we have
\[ |\tau(\hat{f}) - s_0| = o_P(s_0).\]
Here \(s_0\) is a deterministic quantity defined by the problem under consideration; see Eq. 3.1 for its precise definition. Let us return to the problem of least-squares regression to illustrate.

**Example 1.1.** Consider a collection of independent real-valued response variables \(\{Y_i\}_{i=1}^n\) satisfying the regression equation
\[ Y_i = g^0(X_i) + \epsilon_i, \quad \text{for } i = 1, \ldots, n, \]
where the unknown function \(g^0\) belongs to some given collection \(G\) of regression functions, and the sequence \(\{\epsilon_i\}_{i=1}^n\) consists of i.i.d. zero-mean noise variables. The co-variables \(\{X_i\}_{i=1}^n\) are assumed to be fixed. Let \(\hat{g}\) be the least squares estimator
\[ \hat{g} := \arg\min_{g \in G} \left\{ \sum_{i=1}^n (Y_i - g(X_i))^2 \right\}, \]
so that, in the notation used above, the observations are \(\{(X_i, Y_i)\}_{i=1}^n\) and the loss function is given by \(f_g(x, y) = (y - g(x))^2\). Since there is no penalty in this case, the quantity \(\tau^2(f)\) is the excess risk \(P(f - f^0)\) where \(f^0 = f_{g^0}\) and
\[ P(f_g - f_{g^0}) = \frac{1}{n} \sum_{i=1}^n (g(X_i) - g^0(X_i))^2 =: \|g - g^0\|_n^2. \]
With this notation, our aim is to show that
\[ \|\hat{g} - g^0\|_n - s_0 = o_P(s_0) \]
for some (nonrandom) sequence \(s_0\). This example will be studied in Section 2 for the case of i.i.d. Gaussian noise (the normal sequence model), and in Section 6 for the case of sub-Gaussian noise (cf. Corollary 6.4).

In our context, the result requires the complexity of the problem to be in the nonparametric regime meaning that \(\sqrt{\frac{\log n}{n}} = o(s_0)\). When a certain convexity condition is met, the \(\log n\)-term can be removed. This convexity condition holds in the normal sequence model, as well as in all the examples given in Section 6. In Section 2, there is no \(\log n\)-term as well, but the concentration result there is for \(\sqrt{P(\hat{f} - f^0)}\) as opposed to \(\tau(\hat{f})\).
1.2. Organization The remainder of this paper is organized as follows. In Section 2, we provide a concentration result for the normal sequence model and least squares with convex penalty, based on a “direct” argument. We then consider more general models and loss functions, using the more indirect route originally taken by Chatterjee (2014). In Section 3, we discuss the deterministic counterpart of empirical risk minimization, corresponding the population-level optimization problem. Our theory requires a certain amount of curvature of the objective function around its minimum, a requirement that we term a margin condition. In Section 4, we present a concentration result (Theorem 4.1) for a general loss function. Section 5 is devoted to a more careful analysis of quadratic margin conditions. Section 6 is devoted to the detailed analysis of two examples in which the empirical process is linear in its parameter—projection estimators for densities and linearized least squares—whereas Section 7 provides results for nonparametric estimation involving exponential families. In Section 8, we present the concentration inequalities that underlie the proof of our indirect approach. In Section 9, we provide a similar result as in Section 4 but now for a shifted version of $\tau^2(\hat{f})$. Section 10 gives an upper bound for $\tau(\hat{f})$ that can be applied in the previous section as localization step. Finally, all proofs are in provided in Section 11.

2 Direct Approach to Normal Sequence Model

In this section, we analyze the concentration properties of regularized least-squares estimators in the normal sequence setting. The main contribution of this section is to provide a direct argument that generalizes and sharpens the previous result of Chatterjee (2014).

Let $Y_i \in R$ be a response variable and $X_i$ be a fixed co-variable in some space $\mathcal{X}$, $i = 1, \ldots, n$. The normal sequence model is given by

$$Y_i = g^0(X_i) + \epsilon_i \quad \text{for } i = 1, \ldots, n,$$

(2.1)

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. mean-zero Gaussians, and the regression vector $g^0 := (g^0(X_1), \ldots, g^0(X_n))^T$ is unknown. By a scaling argument we may assume here that the errors have unit variance. Let us write the vector of responses as $Y = (Y_1, \ldots, Y_n)^T$, and the noise vector as $\epsilon := (\epsilon_1, \ldots, \epsilon_n)^T$.

Let $\text{pen} : R^n \to R \cup \{+\infty\}$ be a complexity penalty, assumed to be convex. The regularized least squares estimator is given by

$$\hat{g} := \arg \min_{g \in R^n} \left\{ \|Y - g\|_n^2 + \text{pen}(g) \right\},$$

(2.2)
where for any vector $v \in \mathbb{R}^n$, we use the standard notation $\|v\|_n^2 := \frac{\|v\|_2^2}{n}$.

In past work, Chatterjee (2014) analyzed the concentration of the constrained variant of this estimator, given by

$$
\hat{g} := \arg\min_{g \in \mathcal{G}} \left\{ \|Y - g\|_n^2 \right\},
$$

where $\mathcal{G} \subseteq \mathbb{R}^n$ is a closed, convex set. Note that this constrained estimator (2.3) is a special case of the regularized estimator (2.2), in which the penalty function takes the form

$$
\text{pen}(g) := \begin{cases} 0 & \text{if } g \in \mathcal{G} \\ +\infty & \text{otherwise.} \end{cases}
$$

The following result guarantees that with for any convex penalty, the estimation error $\|\hat{g} - g_0\|_n$ of the regularized estimator (2.2) is sharply concentrated around its expectation $m_0 := \mathbb{E}\|\hat{g} - g_0\|_n$.

**Theorem 2.1.** For any convex penalty $\text{pen} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the error in the regularized estimator (2.2) satisfies

$$
\mathbb{P}\left( \left| \|\hat{g} - g_0\|_n - m_0 \right| \geq \sqrt{2t/n} \right) \leq \exp[-t] \quad \text{for all } t > 0.
$$

See Section 11.1 for the proof of this claim. The argument is direct, using some basic facts from convex analysis, and the concentration of Lipschitz functions of Gaussian vectors (see Borell (1975)).

**Remark 2.1.** In terms of asymptotic behavior, it follows that when $1/\sqrt{n} = o(m_0)$ — corresponding to the non-parametric regime — it holds that

$$
\left| \frac{\|\hat{g} - g_0\|_n - m_0}{m_0} \right| = o_\mathbb{P}\left( (nm_0^2)^{-1/2} \right) = o_\mathbb{P}(1).
$$

Moreover, it follows from its proof that Theorem 2.1 remains true if the population minimizer $g_0$ is replaced by any other vector $g \in \mathbb{R}^n$. Thus, for instance, we can take $g$ as the minimizer of the penalized noiseless problem

$$
g^* := \arg\min_{g \in \mathbb{R}^n} \left\{ \|g - g_0\|_n^2 + \text{pen}(g) \right\}
$$

With this choice, we also have concentration of $\|\hat{g} - g^*\|_n$ around its expectation $\mathbb{E}\|\hat{g} - g^*\|_n$. 

Remark 2.2. With the choice \((2.4)\) of penalty function, the result also applies to the constrained least squares estimate \((2.3)\) for any closed convex set \(G \subseteq \mathbb{R}^n\). In this context, Theorem 2.1 sharpens the previous result of Chatterjee (2014). To be clear, Chatterjee (2014) actually proves that \(\|\hat{g} - g^0\|_n\) concentrates around the quantity

\[s_0 := \arg \min [s^2 - E(s)],\]

where for all \(s \geq 0\), we define

\[E(s) := E(\hat{E}_n(s)), \quad \text{where} \quad \hat{E}_n(s) := \max_{g \in G: \|g - g^0\|_n \leq s} \frac{2\epsilon^T(g - g^0)}{n}.\]  

(2.6)

With this notation, he shows that

\[
\frac{\|\hat{g} - g^0\|_n - s_0}{s_0} = O_P \left( ns_0^2 \right)^{-1/4}. 
\]

(2.7)

Hence we necessarily have

\[
\frac{s_0 - m_0}{s_0} = O \left( (ns_0^2)^{-1/2} + (ns_0^2)^{-1/4} \right).
\]

In certain special cases—including linear regression—the quantities \(s_0\) and \(m_0\) are equal, but otherwise it is as yet not clear what the connection is between the two centerings.

3 Theoretical Version of Minimization Problem

For the remainder of the paper, we study the general empirical risk minimizer. We apply an approach similar to the one used in Chatterjee (2014), giving results of the flavour of Eq. 2.7—in particular, see Eqs. 4.4 and 4.7 in the sequel.

Let \(\tau(f)\) denote the penalized excess risk \((1.2)\) associated with the empirical minimizer \(\hat{f}\) from Eq. 1.1. Our goal is to establish conditions under which \(\tau(\hat{f})\) is concentrated around a deterministic quantity \(s_0\), which we now define.

First, we define the minimum possible excess risk

\[\tau^2_{\min} := \min_{f \in \mathcal{F}} \tau^2(f) = \min_{f \in \mathcal{F}} \{P(f - f^0) + \text{pen}(f)\}.
\]

For each \(s \geq \tau_{\min}\), we define the sub-level set \(\mathcal{F}_s := \{f \in \mathcal{F} \mid \tau(f) \leq s\}\), and the functions

\[\hat{E}_n(s) = \max_{f \in \mathcal{F}_s} (P_n - P)(f^0 - f), \quad \text{and} \quad E(s) := E(\hat{E}_n(s)).\]
From hereon, we refer to minimizing the function \( s \mapsto s^2 - E(s) \) as the theoretical problem, and to minimizing \( s \mapsto s^2 - \hat{E}_n(s) \) as the empirical problem.

The two central objects in our analysis are the minimizers

\[
\hat{s} := \arg \min_{s \geq \tau_{\min}} \{s^2 - \hat{E}_n(s)\}, \quad \text{and} \quad s_0 := \arg \min_{s \geq \tau_{\min}} \{s^2 - E(s)\}
\]  

(3.1)
defined by the empirical and theoretical problems respectively. The reader should recall that throughout this paper, we tacitly assume that the minimizers are unique. Note that the theoretical minimizer \( s_0 \) is deterministic, and our results guarantee concentration of \( \tau(\hat{f}) \) around this number. An exact expression for the scalar \( s_0 \) is not evident, but we provide some bounds in Lemma 10.3.

The random quantity \( \hat{s} \) turns out to be equivalent to the risk \( \tau(\hat{f}) \) associated with original ERM estimate (1.1), as guaranteed by the following lemma.

**Lemma 3.1 (Key equivalence).** We have \( \tau(\hat{f}) = \hat{s} \).

See Section 11.2 for the proof of this claim.

With this set-up, the idea for proving concentration of \( \hat{s} \) around \( s_0 \) is as follows. First of all, we can use standard concentration inequalities (Klein and Rio, 2005) to ensure that \( s^2 - \hat{E}_n(s) \) concentrates around \( s^2 - E(s) \) uniformly in \( s \). Based on this uniform convergence, we want to ensure that the empirical minimizer \( \hat{s} = \tau(\hat{f}) \) concentrates around the theoretical minimizer \( s_0 \). This last step requires that the the theoretical function \( s \mapsto s^2 - E(s) \) is suitably curved around its minimum, and we refer to such a curvature condition as a *margin condition*. In this section, we introduce an abstract margin condition, with some specific examples given in Section 5.

**Definition 3.1.** Given a strictly convex increasing function \( G \) on \( \mathbb{R}^+ = [0, \infty) \) with \( G(0) = 0 \) and a subset \( I \subset \mathbb{R}^+ \), we say that the \( G \)-margin condition holds over \( I \) if

\[
[s^2 - E(s)] - [s_0^2 - E(s_0)] \geq G(|s - s_0|) \quad \text{for all} \ s \in I.
\]  

(3.2)

Some particular choices of the set \( I \) play an important role in our analysis. In particular, let \( \delta \geq 0 \) and \( \tau_{\max} \geq s_0 + \delta \) be constants, and suppose that \( s_0 > \tau_{\min} \). The following choices of \( I \) are used in our analysis:

- the *two-sided \( G \)-margin condition* holds when \( I = [\tau_{\min}, s_0) \cup (s_0 + \delta, \tau_{\max}] \).
on Concentration for (Regularized) Empirical Risk Minimization

- the right-sided $G$-margin condition holds when $I = (s_0 + \delta, \tau_{\text{max}}]$. 
- If no reference is made to any range, it means that the condition holds over $I = [\tau_{\text{min}}, s_0) \cup (s_0, \infty)$, corresponding to $\delta = 0$ and $\tau_{\text{max}} = \infty$.

Note that our definition of the two-sided margin condition does not allow for a gap for values of $s$ to the left of $s_0$. This choice is merely for simplicity, and is valid in our examples.

In general, these margin conditions are non-trivial to verify, because it is not clear how the function $s \mapsto E(s)$ behaves: generally only upper bounds are available. Nevertheless, in certain situations, as detailed in Section 5, it is possible to derive an important special case, namely quadratic margin behavior.

This special case of a quadratic margin is of particular interest:

**Definition 3.2.** We say that a quadratic margin condition is met with margin constant $c > 0$ when the conditions in Definition 3.1 hold with the function $G(u) = \frac{u^2}{c^2}$.

When the two-sided condition in Definition 3.2 holds with $\delta = 0$, then it corresponds to a form of strong convexity of the function $s \mapsto s^2 - E(s)$ at $s_0$.

**Remark 3.1.** In Lemma 5.1, we show that if the function $s \mapsto E(s)$ is concave, then the quadratic margin condition holds with margin constant $c = 1$. This concavity condition holds in the normal sequence setting, as exploited by Chatterjee (2014). In the latter paper, the map $s \mapsto \hat{E}_n(s)$ is concave, which implies concavity of the function $s \mapsto E(s)$ as well. Going back to the general case, we note that concavity of $s \mapsto E(s)$ has another nice feature: it implies that $s \mapsto s^2 - E(s)$ is strictly convex. In that situation, the peeling device used to prove Theorem 4.1 below can be avoided, and the log $n$-factor resulting from the peeling device can be removed. This is a special case of Theorem 4.2. As we will see in Theorem 10.1, concavity of $E(\cdot)$ is also very useful towards showing that certain upper and lower bounds for $\tau(\hat{f})$ are of the same order as $s_0$ (see the discussion following Lemma 10.3). Summarizing: assuming concavity of $E(\cdot)$ would simplify the exposition drastically. However, apart from the situation included in Lemma 5.3, it is an assumption difficult to check. This is the reason why our general result relies on an abstract margin condition.

4 Concentration of ERM

We now turn to the statement of our main result on concentration of ERM in the general setting. By concentration of ERM we actually mean
concentration of the (penalized) excess risk $\tau^2(\hat{f})$ of the (penalized) empirical risk minimizer. We begin by specifying some conditions that underlie the result. First, we require a uniform boundedness condition. The reason is that the concentration inequality of Klein and Rio (2005), cited here as Theorem 8.1, requires uniformly bounded functions.

**Condition 4.1 (K-uniform boundedness).** The function class $F$ is uniformly bounded, meaning that

$$K := \max_{f \in F} \|f - f^0\|_\infty < \infty.$$  

We note that this condition can be removed if one first shows that, for a suitable constant $K$, the minimizer $\hat{f}$ satisfies the bound $\|\hat{f} - f^0\|_\infty \leq K$ with high probability. Section 10 provides a mean to do so in the case where the (penalized) loss is convex in the parameter.

When Condition 4.1 holds, one may take

$$\tau_{\text{max}}^2 := 2K + \text{pen}(f^0).$$

However, in order to obtain a sharper result, one may first want to prove that $\tau^2(\hat{f})$ is much smaller than $2K + \text{pen}(f^0)$ with high probability. In fact, there is a substantial literature on techniques for showing a high probability bound for $\tau(\hat{f})$ for various problems. We present such a result in Section 10. As we discuss in Section 9, similar results exist for the shifted version.

In addition, we require that the population risk $f \mapsto Pf$ be suitably curved around its minimum, which we refer to as *curvature condition*. For each function $f \in F$, define the variance

$$\sigma^2(f) := \frac{1}{n} \sum_{i=1}^n \left[ E f^2(X_i) - (E f(X_i))^2 \right].$$

With this notation, we have:

**Definition 4.1.** A quadratic curvature condition with constant $C > 0$ is said to hold if

$$P(f - f^0) \geq \frac{\sigma^2(f - f^0)}{C^2} \quad \text{for all } f \in F. \quad (4.1)$$

We take a quadratic curvature condition as basis for our results. An extension to more general curvature is omitted here to avoid digressions.

Our main result is stated in terms of a suitable bound $J(s)$ on the expected supremum of the empirical process $E(s) = E[\max_{f \in F_s} (P_n - P)(f^0 - f)]$, which we formalize as follows:
Condition 4.2 (Empirical process bound). There is an increasing function \( J : [\tau_{\text{min}}, \infty) \to [0, \infty) \) with \( J(\tau_{\text{min}}) = 0 \) such that the function \( s \mapsto \frac{J(s)}{(s^2 - \tau_{\text{min}}^2)} \) is decreasing, and such that

\[
E(s) \leq J(s) \quad \text{for all } s \geq \tau_{\text{min}}. \tag{4.2}
\]

A bounding function \( J \) can often be derived from empirical process theory results, for example via a chaining argument (e.g., van der Vaart and Wellner (1996)). Condition 4.2 is invoked in Theorems 4.1 and 4.2; we also use it to prove a high probability bound \( r_0 \) for \( \tau(\hat{f}) \), as stated in Theorem 10.1.

The following theorem is our main result. Its proof is based on comparing the minima of the two functions \( Q(s) = s^2 - E(s) \) and \( \hat{Q}(s) = s^2 - \hat{E}_n(s) \). We assume that the \( G \)-margin condition (3.2) holds with some strictly increasing and convex function \( G \), and we let \( G^*(v) := \sup_u \left\{ uv - G(u) \right\} \) denote its convex conjugate. The quadratic margin condition is a special case, which is presented after the theorem and in asymptotic terms. One also sees that the theorem involves a scalar \( r_0 \). It serves as a preliminary upper bound for \( \tau(f) \), see Remark 4.1.

**Theorem 4.1.** Suppose that:

- Conditions 4.1 and 4.2, as well as the curvature condition (4.1) with constant \( C \) hold.
- The right-sided \( G \)-margin condition (3.2) holds over the interval \((s_0 + \hat{\delta}, \tau_{\text{max}}]\).

Consider any scalar \( r_0 > 0 \) satisfying the critical inequality

\[
r_0^2 \geq \tau_{\text{min}}^2 + \frac{4KJ(r_0)}{C^2}.
\]

Then we have

\[
P\left( \tau_{\text{max}} \geq \tau(\hat{f}) > s_0 + \max\{\hat{\delta}, \delta(t)\} \right) \leq \exp[-t] \quad \text{for all } t > 0, \tag{4.3}
\]

where the quantity \( \delta(t) \) is bounded as

\[
G(\delta(t)) \leq G^* \left( c_0 \sqrt{\left[ t + \log(1 + \sqrt{n\tau_{\text{max}}^2})/n \right]} \right) + c_0 \left( (s_0 + r_0) \sqrt{\left[ t + \log(1 + \sqrt{n\tau_{\text{max}}^2})/n \right]} + \left[ t + \log(1 + \sqrt{n\tau_{\text{max}}^2})/n \right] \right),
\]
for some constant $c_0 = c_0(C, K)$. Moreover, if the two-sided version of the $G$-margin condition (3.2) holds over $[\tau_{\text{min}}, s_0] \cup (s_0 + \delta, \tau_{\text{max}})$, then we have

$$P\left( |\tau(\hat{f}) - s_0| > \max\{\hat{\delta}, \delta(t)\}, \tau(\hat{f}) \leq \tau_{\text{max}} \right) \leq 2\exp[-t] \quad \text{for all } t > 0.$$

**Asymptotics** If $G$ is the quadratic function $G(u) = u^2/c^2$, then its convex conjugate $G^\ast(v) = c^2v^2/4$ is also quadratic. Thus, under the two-sided quadratic margin condition and with the scalings $c = \mathcal{O}(1)$, $C = \mathcal{O}(1)$, $K = \mathcal{O}(1)$ and $r_0 \asymp s_0$, we find that

$$\delta(t) = \mathcal{O}\left( (\log n/n)^{1/2} + (s_0^2 \log n/n)^{1/4} \right)$$

for each fixed $t$. This means that

$$\left| \frac{\tau(\hat{f}) - s_0}{s_0} \right| = \mathcal{O}_P\left( \left( \frac{\log n}{ns_0^2} \right)^{1/2} + \left( \frac{\log n}{ns_0^2} \right)^{1/4} \right). \quad (4.4)$$

So whenever $\sqrt{\log n/n} = o_P(s_0)$ we are guaranteed that $|\tau(\hat{f}) - s_0| = o_P(s_0)$.

**Remark 4.1.** We later state Theorem 10.1 in Section 10, which guarantees that under the conditions of Theorem 4.1, any scalar $r_0$ satisfying the critical inequality is a high-probability bound on $\tau(\hat{f})$. Moreover, Lemma 10.3 provides sufficient conditions for the equivalence $s_0 \asymp r_0$.

**Remark 4.2.** The concentration result is for $\tau(\hat{f})^2 = P(\hat{f} - f^0) + \text{pen}(\hat{f})$. In general for the penalized case, it is not clear to us whether concentration holds for the excess risk $P(\hat{f} - f^0)$ itself, with the normal sequence model of Section 2 being one exception.

When the function $s \mapsto s^2 - \hat{E}_n(s)$ is strictly convex, then it is possible to remove the $\log(n)$-factor from Theorem 4.1. We detail such a refinement in the next theorem, where we for simplicity restrict ourselves to (two-sided) quadratic margin behavior. The result is then comparable to Chatterjee (2014).

**Theorem 4.2.** Suppose that the function $s \mapsto s^2 - \hat{E}_n(s)$ is strictly convex, and moreover that:

- Conditions 4.1 and 4.2, as well as the quadratic curvature condition (4.1) with constant $C$ hold.
The two-sided quadratic margin condition holds in the range $[\tau_{\text{min}}, \tau_{\text{max}}]$ with constant $c$.

Consider any scalar $r_0 > 0$ satisfying the critical inequality

$$r_0^2 \geq \tau_{\text{min}}^2 + \frac{KJ(r_0)}{C^2}.$$  

Then for all $t > 0$, we have

$$P\left(|\tau(\hat{f}) - s_0| > \delta(t)\right) \leq 3\exp[-t],$$

where

$$\delta(t) := c^2C\sqrt{\frac{t}{n}} + c\sqrt{4C(s_0 + r_0)}\sqrt{\frac{t}{n}} + \frac{(c^2C^2 + 5K)t}{n}.$$  

Asymptotics  

Theorem 4.2 leads to the same asymptotics as Theorem 4.1 but now without the log-factor: when the constants $c$, $C$ and $K$ remain bounded and $s_0 \asymp r_0$, then

$$\left|\frac{\tau(\hat{f}) - s_0}{s_0}\right| = O_P\left(\left(\frac{1}{ns_0^2}\right)^{1/2} + \left(\frac{1}{ns_0^2}\right)^{1/4}\right).$$

5 Conditions for Quadratic Margin Behavior

In this section, we investigate conditions under which the (right-sided) quadratic margin condition holds over an appropriate range. In particular, we extend the setting of Chatterjee (2014) to the case where one has a strictly convex penalty in Lemma 5.1, and to approximate forms of concavity in Lemmas 5.2 and 5.3. We note that it is possible to formulate different results with other combinations of conditions, but we omit such extensions here.

**Lemma 5.1.** Let $\mathcal{F} := \{f_g \mid g \in \mathcal{G}\}$ be a class of loss functions indexed by a parameter $g \in \mathcal{G}$. Assume that $\mathcal{G}$ is a convex subset of a linear vector space, the mapping $g \mapsto f_g - Pf_g$ is linear, and that for some $q \in (1, 2]$, the mapping $g \mapsto \tau_{2/q}(f_g)$ is convex. Letting $\tau_{\text{max}} := (M + 1)s_0$ for some constant $M > 0$, the right-sided quadratic margin condition holds in the range $(s_0, \tau_{\text{max}}]$ with constant

$$c = \sqrt{2q^{-1}(q - 1)(M + 1)\frac{2(2-q)}{q}}.$$
Moreover, when \( q = 2 \) and \( s_0 > \tau_{\min} \), then the (two-sided) quadratic margin condition holds with \( c = 1 \).

We note that the latter two-sided quadratic margin condition corresponds to the favourable setting of the normal sequence model, as studied by Chatterjee (2014).

Asymptotics When \( q < 2 \) in the above lemma, the idea is that one first proves by separate means that \( \tau(\hat{f}) = \mathcal{O}_P(s_0) \). There is a large literature on upper bounds on this type (e.g., see Koltchinskii (2011) and references therein); in Section 10, we provide a result of this type. Having established such an upper bound, one can then take \( M = \mathcal{O}(1) \).

We sometimes write \( \hat{E}_n(\cdot) = \hat{E}_n^\tau(\cdot) \) and \( E(\cdot) = E^\tau(\cdot) \) so as to highlight their dependence on \( \tau \). For \( f \in \mathcal{F} \), define the functionals

\[
\varsigma^2(f) := c^2\sigma^2(f - f^0) + \text{pen}(f), \quad \text{and} \quad \varsigma_{\min} := \min_{f \in \mathcal{F}} \varsigma(f),
\]

where \( c > 0 \) is some constant. Moreover, for \( s \geq \varsigma_{\min} \), let us define

\[
\hat{E}^\varsigma_n(s) := \max_{f \in \mathcal{F} : \varsigma(f) \leq s} (P_n - P)(f^0 - f), \quad \text{and} \quad E^\varsigma(s) := \mathbb{E}\hat{E}^\varsigma_n(s).
\]

Lemma 5.2. Suppose that the function \( s \mapsto E^\varsigma(s) \) is concave, and that

\[
\frac{\varsigma^2(f)}{A^2} \leq \tau^2(f) \leq A\varsigma^2(f), \quad \text{for all} \ f \in \mathcal{F}, \ \text{and} \ \tau_f \leq \tau_{\max},
\]

where \( A^2 = 1 + \epsilon \) for some \( \epsilon \geq 0 \) satisfying \( \sqrt{\epsilon}(1 + \epsilon) < 1/2 \). For some \( M > 0 \), define

\[
\tau_{\max} := (M + 1)s_0, \quad \text{and} \quad \delta := 2\left[\sqrt{\epsilon}(2\sqrt{\epsilon}M + 1)\right]^{1/2}s_0.
\]

Then when \( s_0 > \tau_{\min} \), the quadratic margin condition holds with constant \( c = 4 \) over the set \( [\tau_{\min}, s_0) \cup (s_0 + \delta, \tau_{\max}] \).

Asymptotics As in Lemma 5.1, one may first prove by separate means that \( \tau(\hat{f}) = \mathcal{O}_P(s_0) \) and then take \( M = \mathcal{O}(1) \).

Lemma 5.2 requires the function \( E^\varsigma \) to be concave. We now present conditions under which this is indeed the case.

Lemma 5.3. Let \( \mathcal{F} := \{f_g \mid g \in \mathcal{G}\} \) be a class of loss functions indexed by the parameter \( g \) in a parameter space \( \mathcal{G} \). Assume \( \mathcal{G} \) is a convex subset of a linear vector space, and that \( g \mapsto \sqrt{\text{pen}(f_g)} \) is convex and \( g \mapsto f_g - Pf_g \) is linear. Then the function \( s \mapsto \hat{E}^\varsigma_n(s) \) is concave.
In fact, we show concavity of the empirical version $\hat{E}_n^\varsigma$, which then implies concavity of $E^\varsigma$. The reasoning is along the lines of Chatterjee (2014), and along the lines of the corresponding part of the proof of Lemma 5.1.

6 Some Pure Cases

We present in this section, as well as in Section 7, some special cases where our main results can be applied. We write the conclusions in the form of corollaries. Yet, for deriving these one does need a few steps, which are given in Section 11 in the form of small proofs.

Let us make a few technical comments before proceeding. In order to avoid cumbersome expressions, we assume from now on that $\log n/(ns_0^2) \to 0$ so that the leading term in the asymptotics (as in Eq. 4.4 or Eq. 4.7) is the one with the power $1/4$. Moreover, we require a certain technical assumption\(^1\)—namely a form of stability of the rates if only a change is made in the constants. To be precise, we require in this and the next section that Condition 4.2 holds and define $u \mapsto \Phi(u) := (J^{-1}(u))^2 - \tau_{\min}^2$. We assume it is increasing and strictly convex, and let its convex conjugate be $v \mapsto \Phi^*(v) = \sup_u \{uv - \Phi(u)\}$. “Stability” says that $\Phi^*$ remains of the same order at two distinct evaluations at bounded arguments: $\Phi^*(\alpha) \simeq \Phi^*(\alpha')$ when $\alpha$ and $\alpha'$ stay away from zero and infinity. If $u \mapsto (E^{-1}(u))^2 - \tau_{\min}^2$ itself is increasing and strictly convex, we assume in this and the next section that $E(\cdot)$ is “stable”. From the discussion following Lemma 10.3, one sees that in that case in Theorems 4.1 and 4.2 one may take $r_0 \approx s_0$.

We now turn to the “pure” cases of this section. They are “pure” in the sense that the empirical process enters in a linear manner. The simplest example of such a pure case is the normal sequence model studied in Section 2, and we examine some other examples here. In our examples the quadratic margin condition holds (possibly only the one-sided version).

More precisely, consider a class of the form $\mathcal{F} := \{f_g \mid g \in \mathcal{G}\}$, where $\mathcal{G}$ is a convex subset of a normed linear vector space $(\mathcal{G}, \|\cdot\|)$. The pure case corresponds to problems in which the mapping $g \mapsto f_g - Pf_g$ is linear, and moreover, we have $P(f_g - f^0) = \|g - g^0\|^2$, where $g_0 = \arg\min_{g \in \mathcal{G}} Pf_g$, which ensures the equivalence $f^0 = f_{g^0}$.

6.1. Density Estimation Using Projection Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution $P$ taking values in a space $\mathcal{X}$. For a sigma-finite measure $\nu$ on $\mathcal{X}$, let $\|\cdot\|$ denote the $L^2(\nu)$-norm. Let $\mathcal{G}$ be a convex subset of a linear vector space $\mathcal{G} \subset L^2(\nu)$, and suppose that density

\(^1\)To be clear, it may be non-trivial to check but is “believable” in many settings.
\( g^0 : = dP/d\nu \) is a member of the model class \( \mathcal{G} \). With this set-up, we consider the estimator
\[
\hat{g} := \arg\min_{g \in \mathcal{G}} \left\{ -P_ng + \|g\|^2/2 + \lambda^2 I^q(g) \right\},
\]
where \( I \) denotes some pseudo-norm on \( \bar{\mathcal{G}} \), the exponent \( q \in (1, 2] \), and \( \lambda \geq 0 \) is a regularization parameter.

In order to analyze the concentration properties of this estimator using our general theory, we begin by casting it within our framework. For each \( g \in \mathcal{G} \), define \( f_g := -g + \frac{1}{2}\|g\|^2 \), as well as the associated function class \( \mathcal{F} := \{ f_g \mid g \in \mathcal{G} \} \). With these choices, for all \( g \in \mathcal{G} \), we have \( (P_n - P)f_g = -(P_n - P)g \), and moreover
\[
P(f_g - f_{g^0}) = -P(g - g^0) + \|g\|^2/2 - \|g^0\|^2/2 = \|g - g^0\|^2.
\]

We split our analysis into several cases, depending on the nature of the penalty \( I \).

6.1.1. Case 1: No Penalty. In this case, the estimator is simply
\[
\hat{g} := \arg\min_{g \in \mathcal{G}} \left\{ -P_ng + \|g\|^2/2 \right\}.
\]

**Corollary 6.1.** Suppose that
- Condition 4.1 holds with \( K = \mathcal{O}(1) \), and
- the function \( s \mapsto E(s) \) is strictly increasing.

Then
\[
\left| \frac{\|\hat{g} - g^0\| - s_0}{s_0} \right| = \mathcal{O}_P(n s_0^2)^{-1/4}.
\]

If in addition, the expected supremum of the empirical process satisfies a bound of the form \( E(s) \leq \mathcal{J}(s) = A s^{1-\alpha}/\sqrt{n} \) for constants \( A > 0 \) and \( \alpha \in (0, 1) \) independent of \( n \), then we have
\[
s_0 = \mathcal{O}(n^{-\frac{1}{2(1+\alpha)}}).
\]

As a concrete example, if \( \mathcal{G} \) is some subset of \( k \)-times differentiable densities, then we expect a bound of this form with \( \alpha = \frac{1}{2k} \), which leads to the familiar minimax scaling \( s_0 = \mathcal{O}(n^{-\frac{k}{2k+1}}) \), and the bound
\[
\left| \frac{\|\hat{g} - g^0\| - s_0}{s_0} \right| = \mathcal{O}_P(n^{-\frac{1}{4(2k+1)}}).
\]
6.1.2. Case 2: Quadratic Penalty. Now let us consider the same estimator with a quadratic penalty:

\[ \hat{g} := \arg \min_{g \in G} \left\{ -P_n g + \|g\|^2/2 + \lambda^2 I^2(g) \right\}. \]

Note that for this estimator, we have \( \tau_{\text{min}}^2 = \min_{g \in G} \left\{ \|g - g^0\|^2/2 + \lambda^2 I^2(g) \right\} \).

**Corollary 6.2.** Suppose that:
- Condition 4.1 holds with \( K = \mathcal{O}(1) \),
- the function \( s \mapsto E(s) \) is strictly increasing.

Then

\[ \frac{\left| \frac{\tau(\hat{f}) - s_0}{s_0} \right|}{O_P \left( (n s_0^2)^{-1/4} \right)} = \mathcal{O}(1). \]

If in addition, there are constants \( \alpha \in (0, 1) \), \( A \), \( A_0 \), all independent of \( n \), such that

\[ E(s) \leq J(s) = \frac{As}{\sqrt{n} \lambda^\alpha} \quad \text{for all } s \geq \tau_{\text{min}}, \quad \text{and} \quad \tau_{\text{min}} \leq \frac{A_0}{\sqrt{n} \lambda^\alpha}, \]

then \( s_0 = \mathcal{O}(1/(\sqrt{n} \lambda^\alpha)) \).

When \( \tau_{\text{min}}^2 \approx \text{pen}(f^0) = \lambda^2 I^2(f^0) \), the condition \( \tau_{\text{min}} \leq A_0/(\sqrt{n} \lambda^\alpha) \) in the above lemma implies that

\[ I^2(f^0) = \mathcal{O} \left( (n \lambda^{2(1+\alpha)})^{-1} \right). \quad (6.1) \]

As a special case, suppose that we take \( \lambda = n^{-\frac{1}{2(1+\alpha)}} \). Then we find as in the previous subsection that \( s_0 = \mathcal{O}(n^{-\frac{1}{2(1+\alpha)}}) \) is valid. Bound (6.1) yields that \( I^2(f^0) = \mathcal{O}(1) \). Otherwise, we see that \( \tau(\hat{f}) \) concentrates on the boundary \( \tau_{\text{min}} \).

6.1.3. Case 3: Strictly Convex Penalty. Recall the general estimator

\[ \hat{g} := \arg \min_{g \in G} \left\{ -P_n g + \|g\|^2/2 + \lambda^q I^q(g) \right\}, \]

where \( 1 < q \leq 2 \) does not depend on \( n \).

**Corollary 6.3.** Suppose that:
- Condition 4.1 holds with \( K = \mathcal{O}(1) \),
○ The equivalence $r_0 \approx s_0$ holds. Then we have the deviation result
\[
\frac{\tau(\hat{f}) - s_0}{s_0} \leq O_P\left(\frac{\log n}{n s_0^2}\right)^{1/4}.
\]

Suppose moreover that there are constants $\alpha \in (0, 1)$ and $A$, independent of $n$, such that
\[
E(s) \leq J(s) = As^{1+(2/q-1)\alpha}/(\sqrt{n}\lambda^{2\alpha/q}) \quad \text{for } s \geq 2\tau_{\min}.
\]
Then we have $s_0 = O(\sqrt{n}\lambda^{2\alpha/q})^{q/(2-q)\alpha}$.

Note that the equivalence condition $r_0 \approx s_0$ is required because the proof is based on applying Lemma 5.1 with $q$ possibly less than 2.

6.2. Linearized Least Squares Regression

Now let us illustrate our general theory in application to a linearized form of least squares regression. Let $\{(X_i, Y_i)\}_{i=1}^n$ be independent samples taking values in $\mathbb{R}^p \times \mathbb{R}$, consider the model
\[
Y_i = g^0(X_i) + \epsilon_i, \quad \text{for } i = 1, \ldots, n,
\]
where each $\epsilon_i$ is zero-mean noise variable independent of $X_i$, and where the function $g^0$ belongs to a convex model class $\mathcal{G}$. Assume that $\mathcal{G}$ is a convex subset of a linear vector space $\tilde{\mathcal{G}}$, and moreover that
\[
K_X := \max_{g \in \mathcal{G}} \|g - g^0\|_\infty < \infty, \quad \text{and} \quad K_0 := \|g^0\|_\infty < \infty.
\]
Let $I$ be some pseudo-norm on $\tilde{\mathcal{G}}$, we consider the estimator
\[
\hat{g} := \arg\min_{g \in \tilde{\mathcal{G}}} \left\{-\frac{1}{n} \sum_{i=1}^n Y_i g(X_i) + \frac{Pg^2}{2} + \lambda^2 I^2(g)\right\}.
\]
(6.2)

Note that implementation of this estimator requires that $Pg^2$ is known for all $g \in \mathcal{G}$. In the special case of fixed design, $Pg^2 = \sum_{i=1}^n g^2(X_i)/n$, so that $\hat{g}$ is simply a penalized least squares estimator.

We present the case of fixed design below, so that the result is comparable to Chatterjee (2014), but now for the case of sub-Gaussian noise. Note that for $\hat{f}(x, y) = y\hat{g}(x) + Pg^2$, we have
\[
\tau^2(\hat{f}) = P(\hat{g} - g^0)^2 + \lambda^2 I^2(\hat{g}).
\]
The following corollary applies to the case of a fixed design $(X_1, \ldots, X_n)$ and the penalized least squares estimator
\[
\hat{g} := \min_{g \in \mathcal{G}} \|Y - g\|_n^2 + \lambda^2 I^2(g).
\]
Corollary 6.4. Assume that:

- There are constants $K_X$ and $K_0$ such that
  \[
  \max_{1 \leq i \leq n} \sup_{g \in \mathcal{G}} |g(X_i) - g^0(X_i)| \leq K_X \quad \text{and} \quad \|g^0\|_\infty \leq K_0.
  \]

- There are constants $c_\epsilon$ and $C_\epsilon$ such that
  \[
  \max_{1 \leq i \leq n} \mathbb{E} \exp\left[\frac{|c_i^2|}{C_\epsilon^2}\right] \leq c_\epsilon,
  \]
- The function $s \mapsto \mathbb{E}(s)$ is strictly increasing.

Then we have
\[
\frac{\left|\tau(\hat{f}) - s_0\right|}{s_0} = O_P\left(\frac{\log n}{n s_0^2}\right)^{1/4}.
\]

(6.3)

7 Exponential Families with Squared Norm Penalty

We now turn to some examples involving exponential families. Throughout this section, we specialize to the case of squared norm penalties, noting that more general penalties can be studied as in the previous section.

7.1. Density Estimation Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with distribution $P$ taking values in $X$. Given a sigma-finite measure $\nu$ on $X$, let us define the function class
\[
\tilde{\mathcal{G}} := \left\{ g : X \to \mathbb{R} \mid \int \exp[g(x)] d\nu(x) < \infty \right\},
\]

and note that it is convex. We define a functional on $\tilde{\mathcal{G}}$ via
\[
d(g) := \log\left(\int \exp[g(x)] d\nu(x)\right).
\]

Let $\mathcal{G} \subset \tilde{\mathcal{G}}$ be a convex subset and define the function $f_g = -g + d(g)$, for each $g \in \mathcal{G}$, along with the associated function class $\mathcal{F} := \{ f_g \mid g \in \mathcal{G} \}$. Letting $I$ be a pseudo-norm on $\tilde{\mathcal{G}}$ along with with a non-negative regularization weight $\lambda$, we consider the estimator
\[
\hat{g} := \arg\min_{g \in \mathcal{G}} \left\{ -P_n g + d(g) + \lambda^2 I^2(g) \right\},
\]

and define $\hat{f} := f_{\hat{g}}$.

For identification purposes, we take the functions in $\mathcal{G}$ to be centered—that is, such that $\int g d\nu = 0$ for each $g \in \mathcal{G}$. For simplicity and without loss of generality, we take $g^0 \equiv 0$ so that $f^0 \equiv 1$ and $\nu = P$. Since $P$ is unknown, the centering of the functions (in actual practice) is done with respect to some other measure. This difference does not alter the theory, our centering is merely to have a convenient notation.
Corollary 7.1. Suppose that:
- \( r_0 \approx s_0 \),
- \( \sup_{\tau(g) \leq r_0} \|g\|_{\infty} = o(1) \).
Then
\[
\left| \frac{\tau(\hat{f}) - s_0}{s_0} \right| = O_P \left( \frac{\log n}{ns_0^2} \right)^{1/4}.
\]

The above corollary requires the condition \( r_0 \approx s_0 \) because the proof is based on applying Lemma 5.2.

7.2. Log-Linear Regression with Fixed Design
Let \( \{(X_i, Y_i)\}_{i=1}^n \) be independent observations taking values in the Cartesian product space \( \mathcal{X} \times \mathbb{R} \).
We assume the design \( \{X_i\}_{i=1}^n \) is fixed, and that for some given sigma-finite measure \( \nu \), the log-density of \( Y_i \) given \( X_i \) takes the log-linear form
\[
-Y_i g^0(X_i) + d(g^0(X_i)),
\]
where the function \( d(\xi) := \log \left( \int \exp[y\xi]d\nu \right) \) has domain
\[
\Xi := \{\xi \in \mathbb{R} \mid \int \exp[y\xi]d\nu < \infty\}.
\]
We define \( g := (g(X_1), \ldots, g(X_n))^T \in \mathbb{R}^n \), let \( \mathcal{G} \) be a convex subset of \( \mathbb{R}^n \) and recall the shorthand notation \( \|v\|_n^2 := \|v\|_2^2/n \). Letting \( I \) be a pseudo-norm on \( \mathbb{R}^n \), we consider the estimator
\[
\hat{g} := \arg\max_{g \in \mathcal{G}} \left\{ -Y^T g/n + \sum_{i=1}^n d(g(X_i))/n + \lambda^2 I^2(g) \right\}.
\]
In this case, the effective function class takes the form
\[
\mathcal{F} = \{f_g(x, y) = -yg(x) + d(g(x)) \mid g \in \mathcal{G}\},
\]
and we have \( \hat{f} = f_{\hat{g}} \).

The following result is along the lines of Corollary 7.1 in the previous subsection. We define here \( \epsilon_i := Y_i - \mathbb{E}Y_i \) for \( i = 1, \ldots, n \).

Corollary 7.2. Assume that:
- We have \( r_0 \approx s_0 \) and \( E(s) = O(s_0^2) \) for \( s = O(s_0) \).
- There is a constant \( c \) such that
\[
\sup_{\tau^2(f_g) \leq \eta} \frac{\sum_{i=1}^n d(g(X_i)) - d(g^0(X_i))}{n\|g - g^0\|_n^2} = c^2(1 + o(1)) \quad \text{as } \eta \downarrow 0.
\]
○ $\sup_{\tau^2(g) \leq r_0} \|g - g^0\|_\infty = O(1)$,
○ There are constants $c_\epsilon$ and $C_\epsilon$ such that $\max_{1 \leq i \leq n} \mathbb{E}\exp[|\epsilon_i|/C_\epsilon] \leq c_\epsilon$. Then we have

$$\left| \frac{\tau(\hat{f}) - s_0}{s_0} \right| = O_p \left( \frac{(\log n)^3}{ns_0^2} \right)^{1/4}.$$

Corollary 7.2 is similar to Corollary 6.4 for least squares regression, which corresponds to the special case $d(g) = g^2/2$. However, Corollary 7.2 includes one additional log factor because $s \mapsto \mathbb{E}_n(s)$ is in general not concave, and another additional log factor because we assume only sub-exponential errors.

8 Concentration for Maxima of Empirical Processes

The following result due to Klein (2002) (see also Klein and Rio (2005)) is our main tool. It involves the maximum deviation $K := \max_{f \in F_s} \|f - f^0\|_\infty$ and maximal variance $\sigma^2_s := \max_{f \in F_s} \sigma^2(f - f^0)$.

**Theorem 8.1.** For all $t \geq 0$, we have

$$\hat{E}_n(s) \geq E(s) - \sqrt{8KE(s)} + 2\sigma^2_s \sqrt{t/n} - \frac{Kt}{n}, \quad \text{and}$$

$$\hat{E}_n(s) \leq E(s) + \sqrt{8KE(s)} + 2\sigma^2_s \sqrt{t/n} + \frac{2Kt}{3n},$$

where each bound holds with probability at least $1 - \exp[-t]$.

Under suitable conditions, we can we replace the quantity $E(s)$ appearing in the square-root of Theorem 8.1 by a suitable upper bound. We summarize in the following:

**Lemma 8.1.** Under Conditions 4.1 and 4.2, we have

$$\hat{E}_n(s) \geq E(s) - 2C(s + r_0)\sqrt{t/n} - \frac{Kt}{n}, \quad \text{and}$$

$$\hat{E}_n(s) \leq E(s) + 2C(s + r_0)\sqrt{t/n} + \frac{2Kt}{3n},$$

where each bound holds with probability at least $1 - \exp[-t]$.

We next present a consequence for the case where the functions in $F_s$ are not uniformly bounded, but have a (sub-Gaussian) envelope function. This result is invoked in the analysis of Section 6.2.
Lemma 8.2. Assume that for some constants $c_F \geq 1$ and $C_F \geq 1$, the envelope function $F(\cdot) := \max_{f \in F} |f(\cdot) - f^0(\cdot)|$ satisfies the bounds

$$PP^2\{ F > t \} \leq c_F^2 \exp[-t^2/C_F^2], \quad t > 0. \tag{8.1}$$

Then for all $t > 0$ with probability at least $1 - 2\exp[-t] - 2t^2$, we have

$$\hat{E}_n(s) \geq E(s) - \sqrt{8C_F \sqrt{\log n[E(s) + 2c_F(c_F + t)/n] + 2s^2 / t/n}} - C_F t \sqrt{\log n/n - c_F(4c_F + t)/n}, \quad \text{and} \quad \tag{8.2}$$

$$\hat{E}_n(s) \leq E(s) - \sqrt{8C_F \sqrt{\log n[E(s) + 2c_F^2/n] + 2s^2 / t/n}} - 2C_F t \sqrt{\log n/(3n) - c_F(4c_F + t)/n}. \tag{8.3}$$

9 The Shifted Version

For a scalar $\tau_*^2 \geq \tau_{\min}^2$ to be chosen, we study in this section the “shifted” function

$$F(s) := \max_{\tau^2(f) \leq \tau_*^2 + s^2} (P_n - P)(f^0 - f), \quad \text{defined for } s^2 \geq \tau_*^2 - \tau_{\min}^2.$$  

This shifted version may be of interest when $\tau^2(\hat{f})$ is of larger order than $P(\hat{f} - f^0)$. The idea is then to replace $g^0$ in the previous sections by the function $g^* := \arg\min_{g \in G} \tau^2(g)$. One then needs curvature conditions on $R(g) - R(g^*)$ instead of $R(g) - R(g^0)$. This we handle here by the notion of an oracle potential, as defined in Definition 9.1 below.

Lemma 9.1 shows that curvature conditions on the function $Q(s; E) := s^2 - E(s)$ are weaker than those on the function $Q(s; F) := s^2 - F(s)$. Using the shorthand $s_*^2 = s_0^2 - \tau_*^2$, the following lemma summarizes this fact:

Lemma 9.1. For any $s \geq \tau_{\min}$ and $\tilde{s}^2 = s^2 - \tau_*^2$, we have

$$Q(s; E) - Q(s_0; E) = Q(\tilde{s}; F) - Q(s_*; F)$$

and $|\tilde{s} - s_*| \geq |s - s_0|$.

We define for $s^2 \geq \tau_*^2 - \tau_{\min}^2$

$$\kappa_{\tilde{s}}^2 := \max\left\{ P(f - f^0) : f \in F, \ P(f - f^0) + \pen(f) \leq \tau_*^2 + s^2 \right\}.$$
Definition 9.1. We say that the oracle potential holds if
\[ \Gamma := \sup_{s > 0} \frac{\kappa_s}{s} < \infty. \] (9.1)

For the shifted version, the counterpart of Condition 4.2 replaces \( E(\cdot) \) by \( F(\cdot) \).

Condition 9.1. There is a strictly increasing function \( J(\cdot) \) such that \( s \mapsto J(s)/s^2 \) is decreasing and such that the bound
\[ F(s) \leq J(s) \] (9.2)
holds for all \( s \geq 0 \).

Theorem 9.1. Suppose that:
- Conditions 4.1 and 9.1, as well as the quadratic curvature condition with constant \( C \) hold.
- The shifted mapping \( Q(\cdot; F) \) satisfies the right-sided \( G \)-margin condition over the interval \( (s_* + \delta, \sqrt{\tau_{\max}^2 - \tau_*^2}] \).
- The oracle potential condition (9.1) holds.

Let \( r_*^2 \geq 4K\Gamma J(r_*)/C^2 \). There is a constant \( c_0 \) depending on \( C, K \) and \( \Gamma \), such that for all \( t > 0 \) and for a constant \( \delta(t) \) such that \( G(\delta(t)) \) is not larger than
\[
G^*(c_0 \sqrt{[t + \log(1 + \sqrt{n(\tau_{\max}^2 - \tau_*^2})]}/n)
+ c_0 (s_* + r_*) \sqrt{[t + \log(1 + \sqrt{n(\tau_{\max}^2 - \tau_*^2})]}/n
+ [t + \log(1 + \sqrt{n(\tau_{\max}^2 - \tau_*^2})]}/n
\]
one has the deviation inequality
\[
P\left( \sqrt{\tau_{\max}^2 - \tau_*^2} \geq \sqrt{\tau^2(\hat{f}) - \tau_*^2} > s_* + \max\{\delta(t), \delta\} \right) \leq \exp[-t]. \] (9.3)
Moreover, if \( s_* > \tau_{\min}^2 - \tau_*^2 \) and in fact the two-sided \( G \)-margin condition holds for \( Q(\cdot; F) \) in the range \( [\sqrt{\tau_{\min}^2 - \tau_*^2}, s_*) \cup (s_* + \delta, \sqrt{\tau_{\max}^2 - \tau_*^2}] \), then one has the concentration inequality
\[
P\left( \sqrt{\tau^2(\hat{f}) - \tau_*^2} > s_* \right) \leq \exp[-t]. \]
10 Upper Bounds for $\tau(\hat{f})$

We first present a localization result that shows that in the convex case one may without loss of generality assume the parameter space is bounded in some appropriate norm $\Upsilon$ (say). If the norm $\Upsilon$ is stronger than the sup-norm (up to constants), this means that the condition that $\{f - f^0 \mid f \in \mathcal{F}\}$ is a uniformly bounded class, so that Condition 4.1 can be removed.

Let $\mathcal{G}$ be a convex subset of a normed vector space $(\mathcal{G}, \Upsilon(\cdot))$. Let $\mathcal{L}_G : \mathcal{G} \to \mathbb{R}$ be some convex (possibly penalized) loss function.

Lemma 10.1. For any fixed $g^* \in \mathcal{G}$ and $\epsilon > 0$, the inequality
\[
\inf \left\{ \mathcal{L}_G(g) \mid g \in \mathcal{G}, \epsilon < \Upsilon(g - g^*) \leq 2\epsilon \right\} > \mathcal{L}_G(g^*)
\]
implies that
\[
\inf \left\{ \mathcal{L}_G(g) \mid g \in \mathcal{G}, \Upsilon(g - g^*) > \epsilon \right\} > \mathcal{L}_G(g^*).
\]

Lemma 10.2. Let $\mathcal{F} := \{f_g \mid g \in \mathcal{G}\}$ where $\mathcal{G}$ is a convex subset of a normed vector space $(\mathcal{G}, \Upsilon(\cdot))$. Let $\mathcal{L}_g : \mathcal{G} \to \mathbb{R}$ be some convex (again possibly penalized) loss function of the form $\mathcal{L}_g(g) = \mathcal{L}_F(f_g), g \in \mathcal{G}$. For any fixed $g^* \in \mathcal{G}$ and $\epsilon > 0$, suppose that
\[
\sup \left\{ \|f_g - f^*\|_\infty \mid g \in \mathcal{G}, \Upsilon(g - g^*) \leq 2\epsilon \right\} \leq K \quad \text{where } f^* := f_{g^*}.
\]
Then the inequality
\[
\inf \left\{ \mathcal{L}_F(f_g) \mid g \in \mathcal{G}, \Upsilon(g - g^*) > \epsilon, \|f_g - f^*\|_\infty \leq K \right\} > \mathcal{L}_F(f^*)
\]
implies
\[
\inf \left\{ \mathcal{L}_F(f_g) \mid g \in \mathcal{G}, \Upsilon(g - g^*) > \epsilon \right\} > \mathcal{L}_F(f^*).
\]

The above lemma can be applied to the case where
\[
\mathcal{L}_F(f) := P_n f + \text{pen}(f), \quad f \in \mathcal{F},
\]
where $\text{pen}(f_g) = \lambda^2 I^q(g), g \in \mathcal{G}$, with $I(\cdot)$ some pseudo-norm on $\mathcal{G} \in L_2(\nu)$.

An appropriate metric $\Upsilon$ is then typically
\[
\Upsilon(g) = \|g\| + \mu I(g) \quad \text{for } g \in \mathcal{G},
\]
with $\|\cdot\|$ corresponding to the $L_2(\nu)$-norm and $\mu$ an appropriate constant possibly depending on $g^*$.

We now provide a high-probability upper bound for $\tau(\hat{f})$. Results of this type are relatively standard; see for example (Koltchinskii, 2011) and references therein. Letting $f^*$ be a function such that $\tau^2_{\min} = P(f^* - f^0) + \text{pen}(f^*)$, the proof of Theorem 10.1 is based on the inequality

$$\mathcal{L}_F(\hat{f}) \leq \mathcal{L}_F(f^*).$$

which follows from the feasibility of $f^*$ and optimality of $\hat{f}$.

In the context of the concentration result of Klein (2002) from Theorem 8.1, and assuming the quadratic curvature condition with constant $C$ (see Definition 4.1) the assumptions (10.1) and (10.2) are true for appropriate constants $c_1$ and $c_2$ depending on $C$ and $K$ and for $J(\cdot) \geq 2\mathbb{E}(\cdot)$ for example. Thus, Theorem 10.1 requires a subset of the conditions of the concentration result formulated in Theorem 4.1: the latter needs in addition a margin condition.

**Theorem 10.1.** Let $J : [\tau_{\min}, \infty) \to \mathbb{R}^+$ is an increasing function such that the function $s \mapsto J(s)/(s^2 - \tau^2_{\min})$ is decreasing over the interval $(\tau_{\min}, \infty)$. Suppose there are constants $c_1$ and $c_2$ such that

$$P\left(\hat{E}_n(\tau_{\min}) < -c_1\tau_{\min}\sqrt{\frac{t}{n} - \frac{c_2^2t}{n}}\right) \leq \exp[-t] \quad \text{for all } t > 0, \quad (10.1)$$

and

$$P\left(\hat{E}_n(s) > J(s) + c_1s\sqrt{\frac{t}{n} + \frac{c_2^2t}{n}}\right) \leq \exp[-t] \quad \text{for all } s \geq \tau_{\min}. \quad (10.2)$$

Then for any scalar $r_0 \geq 2\tau_{\min}$ that satisfies the critical inequality

$$r_0^2 \geq 8J(r_0) + \tau^2_{\min},$$

we have

$$P(\hat{r} > r_0) \leq \frac{2\exp[-nr_0^2/c_3^2]}{1 - \exp[-nr_0^2/c_3^2]},$$

where $c_3 = 4 \max\{8c_1, c_2\}$.

**Remark 10.1.** Note that if the function $s \mapsto J(s)$ is increasing and the function $s \mapsto J(s)/s^2$ is decreasing, then the function $s \mapsto J(s)/(s^2 - \tau^2_{\min})$ is decreasing.
We now present a lemma that provides sufficient conditions for the equivalence $r_0 \asymp s_0$. Recall that given a convex function $\Phi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, its convex conjugate is defined as $\Phi^*(v) := \max_{u>0}\{uv - \Phi(u)\}$.

**Lemma 10.3.** Assume that the function $s \mapsto E(u)$ is strictly increasing and that $u \mapsto \Phi(u) := (E^{-1}(u))^2 - \tau_{min}^2$ is increasing and strictly convex with $\Phi(0) = 0$. Then the function $s \mapsto E(s)/(s^2 - \tau_{min}^2)$ is decreasing, and for any $0 < \alpha < 1$, we have

$$s_0^2 \leq \tau_{min}^2 + \frac{\alpha \Phi^*(1/\alpha) - \Phi^*(1)}{1 - \alpha}, \quad \text{and} \quad s_0^2 \geq \tau_{min}^2 + \frac{\alpha \Phi^*(1) - \Phi^*(\alpha)}{1 - \alpha}.$$  

Moreover, if we take $r_0 := \sqrt{4\tau_{min}^2 + \Phi^*(16)} \geq 2\tau_{min}$, then we have

$$r_0^2 - \tau_{min}^2 \geq 8E(r_0).$$

Observe that strict convexity ensures that $\Phi^*(\alpha) < \alpha \Phi^*(1)$ for all $\alpha \in (0,1)$, so that the bounds for $s_0^2 - \tau_{min}^2$ in the above lemma are strictly positive. The lemma shows that with the bounds (10.1) and (10.2) derived from Theorem 8.1, and assuming the quadratic curvature condition, Theorem 10.1 can be applied with $r_0^2 = 4\tau_{min}^2 + \Phi^*(16)$, and the bounding function $J(s) = 2E(s)$. Then, if (for instance), we have $\Phi^*(16) \asymp \Phi^*(1/16)$—as an instance of the stability condition outlined in the beginning of Section 6—then the equivalence $r_0 \asymp s_0$ follows. Lemma 10.3 involves the condition that $s^2 \mapsto E(\sqrt{s^2 - \tau_{min}^2})$ is strictly concave. One way of verifying this condition is via Lemma 5.3.

## 11 Proofs

This section is devoted to the proofs of our results.

### 11.1. Proof of Theorem 2.1

After some simple algebra, we may write

$$\hat{g} = \arg\min_{g \in \mathbb{R}^n} \left\{ \frac{1}{2} \tau^2(g) - \frac{\epsilon}{n} g^T g / n \right\},$$

where $\tau^2(g) := \|g - g^0\|_n^2 + \text{pen}(g)$. The function $\tau^2$ is convex, so that it has a sub-differential, denoted by $\partial \tau^2(g)$. With this notation, the minimizing argument $\hat{g}$ must satisfy the relation

$$\epsilon / n \in \partial \tau^2(\hat{g}) / 2.$$  

(11.1)
Due to the strong convexity and coercivity of \( g \mapsto \tau^2(g) \), the inclusions (11.1) always have a unique solution \( \hat{g} \).

We now use a classical fact from convex analysis (Rockafellar, 1970): since the function \( g \mapsto \tau^2(g)/2 \) is \((1/n)\)-strongly convex, the sub-differential mapping \( g \mapsto \partial \tau^2(g)/2 \) is \((1/n)\)-strongly monotone, which means that for any pair of vectors \( u, u' \in \mathbb{R}^n \)

\[
\|u - u'\|^2/n \leq (v - v')^T(u - u'),
\]

(11.2)

where \( v, v' \) denote any members of \( \partial \tau^2(u)/2 \) and \( \partial \tau^2(u')/2 \) respectively.

By Borell’s theorem (1975) on the concentration of Lipschitz functions of Gaussian vectors, it suffices to show that the mapping \( \epsilon \mapsto \hat{m} := \|\hat{g} - g^0\|_n \) is Lipschitz with parameter \( 1/\sqrt{n} \). Let \( \epsilon \) and \( \epsilon' \) be two realizations of the noise vector, with corresponding solutions \( \hat{g} \) and \( \hat{g}' \), along with their associated errors \( \hat{m} = \|\hat{g} - g^0\|_n \) and \( \hat{m}' = \|\hat{g}' - g^0\|_n \). By the triangle inequality, we have

\[
|\hat{m} - \hat{m}'| = \|\hat{g} - g^0\|_n - \|\hat{g}' - g^0\|_n \leq \|\hat{g} - \hat{g}'\|_n,
\]

so that it suffices to prove that

\[
\|\hat{g} - \hat{g}'\|_n \leq \|\epsilon - \epsilon'\|_n = \frac{1}{\sqrt{n}} \|\epsilon - \epsilon'\|_2.
\]

Now consider the pair \( u = \hat{g} \) and \( u' = \hat{g}' \), along with the corresponding elements \( v = \epsilon/n \) and \( v' = \epsilon'/n \). Applying the monotone property (11.2) to these pairs yields the inequality

\[
\|\hat{g} - \hat{g}'\|^2/n \leq (\epsilon - \epsilon')^T(\hat{g} - \hat{g}')/n \leq \|\epsilon - \epsilon'\|_n \|\hat{g} - \hat{g}'\|_n,
\]

where the final step follows from the Cauchy-Schwarz inequality. Cancelling terms completes the proof.

11.2. Proofs for Section 3 In this section, we collect the proofs of all results stated in Section 3.

Proof of Lemma 3.1:. For any scalar \( s \) and \( f \in \mathcal{F} \) such that \( \tau(f) \leq s \), we have

\[
P_n(f - f^0) + \text{pen}(f) = \tau^2(f) - (P_n - P)(f^0 - f) \leq s^2 - (P_n - P)(f^0 - f).
\]
Consequently, we have
\[
\tau^2(\hat{f}) - (P_n - P)(f^0 - \hat{f}) = \min_{f \in \mathcal{F}} \left\{ \tau^2(f) - (P_n - P)(f^0 - f) \right\}
\]
\[
= \min_{s \geq \tau_{\min}} \min_{\tau(f) \leq s} \left\{ \tau^2(f) - (P_n - P)(f^0 - f) \right\}
\]
\[
\leq \min_{s \geq \tau_{\min}} \min_{\tau(f) \leq s} \left\{ s^2 - (P_n - P)(f^0 - f) \right\}
\]
\[
= \min_{s \geq \tau_{\min}} \left\{ s^2 - \hat{\mathcal{E}}_n(s) \right\}.
\]

On the other hand, for any \( f \in \mathcal{F} \), we have the lower bound
\[
\tau^2(f) - (P_n - P)(f^0 - f) \geq \tau^2(\hat{f}) - \hat{\mathcal{E}}_n(\tau(\hat{f})) = \tau^2(f) - \hat{\mathcal{E}}_n(\tau(f)),
\]
which implies that
\[
\tau^2(\hat{f}) - (P_n - P)(f^0 - \hat{f}) \geq \tau^2(\hat{f}) - \hat{\mathcal{E}}_n(\hat{f}) \geq \min_{s \geq \tau_{\min}} \{ s^2 - \hat{\mathcal{E}}_n(s) \}.
\]

Since the minimizing argument \( \hat{s} = \arg \min_{s \geq \tau_{\min}} \{ s^2 - \hat{\mathcal{E}}_n(s) \} \) is unique by assumption, we conclude that \( \tau(\hat{f}) = \hat{s} \), as claimed.

11.3. Proofs for Section 4
We first state and prove an auxiliary lemma that serves as a tool in the proof of Theorem 4.1.

**Lemma 11.1.** Let \( G \) be a real-valued function with convex conjugate \( G^* \). Then all for positive scalars \( a, b \) and \( c \) such that \( G(a) \geq G^*(2b) + 2c \), we have \( G(a) - ab - c \geq 0 \).

**Proof of Lemma 11.1.** By the Fenchel-Young inequality, we have
\[
ab = a \frac{2b}{2} \leq \frac{G(a)}{2} + \frac{G^*(2b)}{2},
\]
and consequently,
\[
G(a) - ab - c \geq G(a) - \left\{ \frac{G(a)}{2} + \frac{G^*(2b)}{2} \right\} - c = \frac{G(a)}{2} - \frac{G^*(2b)}{2} - c \geq 0,
\]

as claimed.
We now turn the proof of the main theorem.

**Proof of Theorem 4.1.**

Our strategy is to apply a “peeling argument” so as to transition from the fixed $s$-result of Theorem 8.1 to a result that holds uniformly in $s$. That is, we show that $\hat{Q}(s) := s^2 - \hat{E}_n(s)$ is uniformly in $s$ close to $Q(s) := s^2 - E(s)$. Once this is established we invoke that for constants $a, b \geq 0$,

\[
\hat{Q}(s) - Q(s_0) \geq G(|s - s_0|) \quad \forall s
\]

\[
\hat{Q}(s) - Q(s) \leq a|s - s_0| + b \quad \forall s \Rightarrow G(|\hat{s} - s_0|) \leq G^*(2a) + 2b,
\]

which can be verified straightforwardly.

Let $t > 0$ be arbitrary, and define

\[
z(t) := 2C(s_0 + r_0)\sqrt{\frac{t}{n} + \frac{Kt}{3n}}, \quad \text{and} \quad \bar{z}(t) := 2C(s_0 + r_0)\sqrt{\frac{t}{n} + \frac{Kt}{n}}.
\]

Assume first that the right-sided margin condition holds with $\delta = 0$. For a parameter $\epsilon > 0$ to be chosen later, define the integer $J := \lceil \tau_{\max} / \epsilon \rceil$, the collection of intervals

\[
I_j := [(j-1)\epsilon + \delta + s_0, s_0 + \delta + j\epsilon], \quad \text{for } j = 1, \ldots, J,
\]

as well as the associated probabilities

\[
P_j := P\left(\exists s \in I_j \text{ such that } s^2 - \hat{E}_n(s) \leq s_0^2 - E(s_0) + \bar{z}(t)\right).
\]

Then for a parameter $\delta > 0$ to be chosen later (and leading to the $\delta(t)$ in the theorem statement), we have

\[
P\left(s_0 + \delta < \hat{s} \leq \tau_{\max}, \hat{E}_n(s_0) > E(s_0) + \bar{z}(t)\right)
\]

\[
\leq P\left(\exists s \in (s_0 + \delta, \tau_{\max}] \text{ such that } s^2 - \hat{E}_n(s) \leq s_0^2 - E(s_0) + \bar{z}(t)\right)
\]

\[
\leq \sum_{j=1}^{J} P_j.
\]

For each index $j$ and for all $s \in I_j$, we have

\[
s^2 - \hat{E}_n(s) \geq ((j-1)\epsilon + \delta + s_0)^2 - \hat{E}_n(s_0 + \delta + j\epsilon).
\]

Moreover, for all $u > 0$, we have by Theorem 8.1

\[
\hat{E}_n(s_0 + \delta + j\epsilon) \leq E(s_0 + \delta + j\epsilon) + 2C(\delta + j\epsilon)\sqrt{u/n} + \bar{z}(u)
\]
with probability at least $1 - \exp[-u]$. Furthermore, by the one-sided form of margin condition, we have the lower bound

$$(s_0 + \delta + (j-1)\epsilon)^2 - \mathbf{E}(s_0 + \delta + j\epsilon) \geq s_0^2 - \mathbf{E}(s_0) + G(\delta + j\epsilon)$$

$$+ (s_0 + \delta(j-1)\epsilon)^2 - (s_0 + \delta + j\epsilon)^2$$

$$= s_0^2 - \mathbf{E}(s_0) + G(\delta + j\epsilon) - 2\epsilon(s_0 + \delta + j\epsilon) + \epsilon^2.$$

Putting together the pieces, for all $s \in I_j$, we have

$$s^2 - \hat{\mathbf{E}}_n(s) \geq s_0^2 - \mathbf{E}(s_0) + G(\delta + j\epsilon) - 2 (C\sqrt{\frac{u}{n}} + \epsilon)(\delta + j\epsilon) - 2\epsilon s_0 + \epsilon^2 - \bar{z}(u),$$

with probability at least $1 - \exp[-u]$. We now apply Lemma 11.1 with the choices $a := \delta + j\epsilon$, $b := 2(C\sqrt{u/n} + \epsilon)$, and $c := 2\epsilon s_0 - \epsilon^2 + \bar{z}(u) + \bar{z}(t)$. In order to be able to do so, we require that

$$G(\delta) \geq G^*(4(C\sqrt{u/n} + \epsilon)) + 2 \left(2\epsilon s_0 - \epsilon^2 + \bar{z}(u) + \bar{z}(t)\right).$$

(11.3)

We now settle the choice of $\epsilon$ and $u$. Taking $\epsilon = 1/\sqrt{n}$, we are then guaranteed that

$$J \leq 1 + \frac{\tau_{\max}}{\epsilon} = 1 + \sqrt{n\tau_{\max}^2}.$$

Moreover, recalling the arbitrary $t > 0$ introduced at the start of the proof, we set $u = t + \log(1 + \sqrt{n\tau_{\max}^2})$, and then the condition (11.3) on $\delta := \delta(t)$ becomes

$$G(\delta(t)) := G^* \left(4(C\sqrt{[t + \log(1 + \sqrt{n\tau_{\max}^2})]/n + 1/\sqrt{n}})\right)$$

$$+ 2 \left(2s_0/\sqrt{n} - 1/\sqrt{n} + \bar{z}([t + \log(1 + \sqrt{n\tau_{\max}^2})]) + \bar{z}(t)\right).$$

We are then guaranteed that for each $j \in \{1, \ldots, J\}$, with probability at least $1 - \exp[-(t + \log(1 + \sqrt{n\tau_{\max}^2})]]$, for all $s \in I_j$ it holds that

$$s^2 - \hat{\mathbf{E}}_n(s) \geq s_0^2 - \mathbf{E}(s_0) + \bar{z}(t).$$

It follows that $\mathbb{P}_j \leq \exp \left\{-[t + \log(1 + \sqrt{n\tau_{\max}^2})]\right\}$, and hence

$$\sum_{j=1}^{J} \mathbb{P}_j \leq J \exp \left[-[t + \log(1 + \sqrt{n\tau_{\max}^2})]\right] \leq \exp[-t].$$
One easily verifies that for some constant $c_0$ depending on $C$ and $K$, we have
\[
G(\delta(t)) \leq G^*(c_0 \sqrt{[t + \log(1 + \sqrt{n\tau_{\max}^2})/n])}
+ c_0 \left((s_0 + r_0)\sqrt{[t + \log(1 + \sqrt{n\tau_{\max}^2})/n] + [t + \log(1 + \sqrt{n\tau_{\max}^2})/n]}\right).
\]

Overall, we have established that $\mathbb{P}(\delta > s_0 + \delta(t)) \leq 2 \exp[-t]$.

In stating the bound in the theorem, we removed the pre-factor of 2 for cosmetic reasons. This can be done by replacing $t$ by $t + \log 2$. In order to prove the lower bound, one may follow the same argument, instead using the left-sided version of the margin condition.

In our argument thus far, we assumed $\delta = 0$. If the margin condition only holds at distance $\tilde{\delta} > 0$, it is clear that one simply can take the maximum of $\delta(t)$ and $\tilde{\delta}$ in the bounds.

**Proof of Theorem 4.2.** The proof is based on the argument that if $h : \mathbb{R} \to \mathbb{R}$ is a strictly convex function, and there are scalars $s_2 > s_1$ and a constant $c_0 > \min_s h(s)$ such that $h(s_k) \geq c_0$ for $k = 1, 2$, then we must have the inclusion $s_1 \leq \arg\min_s h(s) \leq s_2$.

For each $t > 0$, define the scalars
\[
\bar{z}(t) := 2C(s_0 + r_0)\sqrt{\frac{t}{n} + \frac{2Kt}{3n}}, \quad \text{and} \quad z(t) := 2C(s_0 + r_0)\sqrt{\frac{t}{n} + \frac{Kt}{n}}.
\]
By an application of Lemma 8.1, for each $s$, we have
\[
\hat{E}_n(s) \leq E(s) + 2C(s - s_0)\sqrt{\frac{t}{n} + \bar{z}(t)}, \quad \text{and} \quad \hat{E}_n(s) \geq E(s) - z(t),
\]
where each inequality hold probability at least $1 - \exp[-t]$. Apply the above upper bound with the pair $s_1 > s_0 + \delta(t)$ and $s_2 < s_0 - \delta(t)$, followed by the lower bound with $s = s_0$. Combining these three bounds guarantees that for each $k = 1, 2$, with probability at least $1 - 3 \exp[-t]$, we have the lower bound
\[
s_k^2 - \hat{E}_n(s_k) \geq s_k^2 - E(s_k) - 2C(s_k - s_0)\sqrt{\frac{t}{n} + \bar{z}(t)}
\geq s_0^2 - E(s_0) + \frac{(s_k - s_0)^2}{C^2} - 2C(s_k - s_0)\sqrt{\frac{t}{n} - \bar{z}(t)}
= s_0^2 - E(s_0) + \left(\frac{s_k - s_0}{C} - cC\sqrt{\frac{t}{n}}\right)^2 - \frac{c^2C^2t}{n} - \bar{z}(t),
\]
as well as
\[ s_0^2 - \hat{E}_n(s_0) \leq s_0^2 - \mathbf{E}(s_0) + z(t). \]

But by the definition of \( \delta(t) \), we have
\[
\left( \frac{s_k - s_0}{c} - cC \sqrt{\frac{t}{n}} \right)^2 - \frac{c^2C_2 t}{n} - \bar{z}(t) > z(t) \quad \text{for } k = 1, 2.
\]

11.4. Proofs for Section 5
We now turn to the proofs of results from Section 5.

**Proof of Lemma 5.1.** Let us introduce the shorthand notation \( \tilde{s} := s^{2/q} \) and \( \tilde{s}_0 = s_0^{2/q} \), let \( \tilde{s}_1, \tilde{s}_2 \in [\tau_{\min}^{2/q}, \infty) \) be arbitrary, and define
\[
\hat{f}_1 := f_{\hat{g}_1} := \arg \max_{\tau^{2/q}(f) \leq \tilde{s}_1} (P_n - P)(f^0 - f),
\]
as well as
\[
\hat{f}_2 := f_{\hat{g}_2} := \arg \max_{\tau^{2/q}(f) \leq \tilde{s}_2} (P_n - P)(f^0 - f).
\]

With these choices, we have
\[
\tau^{2/q}(f_{\hat{g}_1} + (1 - t)\hat{g}_2) \leq t\tau^{2/q}(\hat{f}_1) + (1 - t)\tau^{2/q}(\hat{f}_2) \leq t\tilde{s}_1 + (1 - t)\tilde{s}_2 \quad \forall \ t \in [0, 1].
\]

Moreover, we have the lower bound
\[
\hat{E}_n((t\tilde{s}_1 + (1 - t)\tilde{s}_2)^{q/2}) \geq (P_n - P)(f^0 - f_{\hat{g}_1} + (1 - t)\hat{g}_2) = t(P_n - P)(f^0 - \hat{f}_1) + (1 - t)(P_n - P)(f^0 - \hat{f}_2) = t\hat{E}_n(\tilde{s}_1^{q/2}) + (1 - t)\hat{E}_n(\tilde{s}_2^{q/2}).
\]

Taking expectations yields the lower bound
\[
\mathbf{E}((t\tilde{s}_1 + (1 - t)\tilde{s}_2)^{q/2}) \geq t\mathbf{E}(\tilde{s}_1^{q/2}) + (1 - t)\mathbf{E}(\tilde{s}_2^{q/2}).
\]

Using the fact that
\[
[(t\tilde{s}_1 + (1 - t)\tilde{s}_2)^q - \mathbf{E}((t\tilde{s}_1 + (1 - t)\tilde{s}_2)^{q/2})] - [\tilde{s}_0^q - \mathbf{E}(\tilde{s}_0^{q/2})] \geq 0,
\]
we have
\[
[\tilde{s}^q - \mathbf{E}(\tilde{s}^{q/2})] - [\tilde{s}_0^q - \mathbf{E}(\tilde{s}_0^{q/2})] \geq \tilde{s}^q - \tilde{s}_0^q + \frac{1}{t}\tilde{s}_0^q - (t\tilde{s} + (1 - t)\tilde{s}_0)^q.
\]
Taking $t \downarrow 0$ then gives

$$[\tilde{s}^q - \mathbf{E}(\tilde{s}^{q/2})] - [\tilde{s}_0^q - \mathbf{E}(\tilde{s}_0^{q/2})] \geq \tilde{s}^q - \tilde{s}_0^q - q\tilde{s}_0^{q-1}(\tilde{s} - \tilde{s}_0)$$

$$\geq q(q-1)(M+1)\frac{2(2-q)}{q} \tilde{s}_0^{-(2-q)}(\tilde{s} - \tilde{s}_0)^2/2,$$

valid when $q \in (1, 2]$ and $\tilde{s} > (M+1)^{2/q}\tilde{s}_0$. Furthermore, for $1 < q \leq 2$ we get for some $s_0 < \tilde{s} \leq s$

$$\tilde{s} - \tilde{s}_0 = s^{2/q}_0 - s_0^{2/q} = 2\tilde{s} \frac{2-q}{q} (s - s_0)/q \geq 2s_0^{2/q} (s - s_0)/q.$$

Consequently, for all $q \in (1, 2]$, we have

$$[\tilde{s}^q - \mathbf{E}(\tilde{s}^{q/2})] - [\tilde{s}_0^q - \mathbf{E}(\tilde{s}_0^{q/2})] \geq 2(q-1)(M+1)\frac{2(2-q)}{q} (s - s_0)^2/q,$$

as claimed.

**Proof of Lemma 5.2.** For all $s \geq \tau_{\min}$, we have

$$\mathbf{E}^T(As) \geq \mathbf{E}^c(A^{1/2}s) \geq \mathbf{E}^T(s).$$

For $ts + (1-t)s_0 \geq \tau_{\min}$, the concavity of $\mathbf{E}^c$ yields

$$\mathbf{E}^T(A(ts + (1-t)s_0)) \geq \mathbf{E}^c(A^{1/2}(ts + (1-t)s_0))$$

$$\geq t\mathbf{E}^c(A^{1/2}s) + (1-t)\mathbf{E}^c(A^{1/2}s_0)$$

$$\geq t\mathbf{E}^T(s) + (1-t)\mathbf{E}^T(s_0)$$

$$= t(\mathbf{E}(s) - \mathbf{E}(s_0)) + \mathbf{E}(s_0).$$

Since $s_0$ is the minimizer of $s^2 - \mathbf{E}(s)$, we have

$$A^2(ts + (1-t)s_0)^2 - \mathbf{E}(A^2(ts + (1-t)s_0)) \geq s_0^2 - \mathbf{E}(s_0),$$

and consequently

$$\mathbf{E}(A(ts + (1-t)s_0)) \leq A^2(ts + (1-t)s_0)^2 - s_0^2 + \mathbf{E}(s_0).$$

It follows that

$$t(\mathbf{E}(s) - \mathbf{E}(s_0)) \leq A^2(ts + (1-t)s_0)^2 - s_0^2$$

$$= A^2(t^2s^2 + s_0^2 - 2ts_0^2 + t^2s_0^2 + 2tss_0 - 2t^2s_0^2) - s_0^2$$

$$= A^2(t^2(s - s_0)^2 + s_0^2 - 2ts_0^2 + 2tss_0) - s_0^2$$

$$= t^2(s - s_0)^2 - 2ts_0^2 + 2tss_0$$

$$+ \epsilon(t^2(s - s_0)^2 - 2ts_0^2 + 2tss_0) + \epsilon s_0^2.$$
Putting together the pieces, we have
\[
[s^2 - \mathbf{E}(s)] - [s_0^2 - \mathbf{E}(s_0)] = s^2 - s_0^2 - \frac{t(\mathbf{E}(s) - \mathbf{E}(s_0))}{t}
\]
\[
\geq s^2 - s_0^2 - t(s - s_0)^2 + 2s_0^2 - 2ss_0 - \epsilon \left( t(s - s_0)^2 - 2s_0^2 + 2ss_0 \right) - \epsilon s_0^2/t
\]
\[
= (s - s_0)^2 - t(s - s_0)^2 - \epsilon \left( t(s - s_0)^2 + 2s_0(s - s_0) \right) - \epsilon s_0^2/t.
\]

We now choose \( t = \sqrt{\epsilon} \) to find that
\[
[s^2 - \mathbf{E}(s)] - [s_0^2 - \mathbf{E}(s_0)] \geq (s - s_0)^2 - \sqrt{\epsilon}(s - s_0)^2
\]
\[
- \epsilon \left( \sqrt{\epsilon}(s - s_0)^2 + 2s_0(s - s_0) \right) - \sqrt{\epsilon}s_0^2.
\]

Next we use our assumption that \( s \leq \tau_{\text{max}} = (M + 1)s_0 \). We then get
\[
[s^2 - \mathbf{E}(s)] - [s_0^2 - \mathbf{E}(s_0)] \geq \left( 1 - \sqrt{\epsilon}(1 + \epsilon) \right)(s - s_0)^2 - \epsilon \right( 2s_0^2 M \right) - \sqrt{\epsilon} s_0^2
\]
\[
= (1 - \sqrt{\epsilon}(1 + \epsilon))(s - s_0)^2 - \sqrt{\epsilon}(2\sqrt{\epsilon} M + 1)s_0^2
\]
\[
\geq (s - s_0)^2/2 - \sqrt{\epsilon}(2\sqrt{\epsilon} M + 1)s_0^2.
\]

Now we take \( |s - s_0| \geq 2[\sqrt{\epsilon}(2\sqrt{\epsilon} M + 1)]^{1/2}s_0 \), and conclude that
\[
[s^2 - \mathbf{E}(s)] - [s_0^2 - \mathbf{E}(s_0)] \geq (s - s_0)^2/4.
\]

**Proof of Lemma 5.3.** Let \( s_1 \geq \varsigma_{\text{min}} \) and \( s_2 \geq \varsigma_{\text{min}} \) be arbitrary, and define
\[
\hat{f}_1 := f_{\hat{g}_1} := \arg \max_{\varsigma(f) \leq s_1} (P_n - P)(f^0 - f),
\]
and
\[
\hat{f}_2 := f_{\hat{g}_2} := \arg \max_{\varsigma(f) \leq s_2} (P_n - P)(f^0 - f).
\]

For all \( t \in [0, 1] \), we have
\[
\varsigma(f_{t\hat{g}_1 + (1-t)\hat{g}_2}) \leq t\varsigma(\hat{f}_1) + (1-t)\varsigma(\hat{f}_2) \leq ts_1 + (1-t)s_2.
\]

In addition, we have
\[
\hat{\mathbf{E}}_n^\varsigma(ts_1 + (1-t)s_2) \geq (P_n - P)(f^0 - f_{t\hat{g}_1 + (1-t)\hat{g}_2})
\]
\[
= t(P_n - P)(f^0 - f_{\hat{g}_1}) + (1-t)(P_n - P)(f^0 - f_{\hat{g}_2})
\]
\[
= t\hat{\mathbf{E}}_n^\varsigma(s_1) + (1-t)\hat{\mathbf{E}}_n^\varsigma(s_2),
\]

which completes the proof.
11.5. Proofs for Section 6

Proof of Corollary 6.1. First, note that the assumption that \( s \mapsto E(s) \) is strictly increasing ensures that \( s^2 \mapsto E(s) \) is strictly concave. We choose \( J(\cdot) \propto E(\cdot) \). Then it follows from Theorem 10.1 that \( r_0 \asymp s_0 \) by the stability assumption on \( J \). We then combine Theorem 4.2 with Lemma 5.1. The parameter \( q \) is in this case \( q = 2 \).

Proof of Corollary 6.2. This follows from the same arguments as in the proof of Corollary 6.1.

Proof of Corollary 6.3. This follows from Theorem 4.1 combined with Lemma 5.1.

Proof of Corollary 6.4. Note that the function \( s^2 \mapsto E(s) \) is strictly concave. Setting \( J(s) \propto E(s) \), a standard adjustment of Theorem 10.1 and the stability assumption on \( J \) imply that \( \tau(\hat{f}) = O(s_0) \).

Given the form of the estimator, we have

\[
\mathcal{F} = \left\{ f_g(x, y) = -yg(x) + Pg^2/2 \mid g \in \mathcal{G} \right\}.
\]

We claim that this class has an envelope function \( F \) satisfying the conditions of Lemma 8.2. In order to verify this claim, note the bounds

\[
\max_{g \in \mathcal{G}} |\epsilon_i(g(X_i) - g^0(X_i))| \leq |\epsilon_i| K_X,
\]

\[
\max_{g \in \mathcal{G}} |g^0(X_i)(g(X_i) - g^0(X_i))| \leq K_X K_0, \quad \text{and}
\]

\[
\max_{g \in \mathcal{G}} |Y_i(g(X_i) - g^0(X_0))| \leq (|\epsilon_i| + K_0) K_X.
\]

The concentration result of Lemma 8.2 can thus be used. The class \( \mathcal{F} \) is however not uniformly bounded in this case so that one cannot apply Lemma 8.1. However, since \( \tau(\hat{f}) = O_P(s_0) \) and \( E(s) = O(s_0^2) \) for \( s = O(s_0) \) (see also Lemma 10.3), we have an alternative way to deal with the \( E(s) \) term under the square-root in Lemma 8.2. The result then follows from Lemma 5.1 with \( q = 2 \) and from Theorem 4.2 with the minor modification that one invokes Lemma 8.2 instead of Theorem 8.1 as concentration tool.

11.6. Proofs for Section 7

The following lemma captures some useful properties of the function \( d \):

**Lemma 11.2.** Suppose that \( K := \max_{g \in \mathcal{G}} \|g\|_{\infty} < \infty \). Then we have

\[
\max_{g \in \mathcal{G}} \left| \frac{d(tg)}{t^2Pg^2} - \frac{1}{2} \right| = O(t), \quad \text{valid as } t \downarrow 0. \quad (11.4)
\]
Moreover, for each $g \in G_{\infty}(\eta) := \{g \mid \|g\|_{\infty} \leq \eta\}$, we have
\[
d(g) = \frac{1}{2}Pg^2(1 + O(\eta)), \quad \text{valid as } \eta \downarrow 0.
\] (11.5)

**Proof of Lemma 11.2.** Throughout this proof, we let $0 \leq \tilde{t} \leq t$ be some intermediate point, not the same at each appearance. The function $h(t) = d(tg)$ is infinitely differentiable with $h(0) = 0$ and $h'(0) = 0$, so a second-order Taylor series expansion yields $d(tg) = \frac{1}{2}t^2h''(t)$. Consequently, it suffices to show that
\[
h''(t) = Pg^2(1 + O(t)).
\] (11.6)

Computing derivatives, we have
\[
h'(t) = \left[ P \exp[tg] \right]^{-1}P(\exp[tg]g), \quad \text{and}
\]
\[
h''(t) = \left[ P \exp[tg] \right]^{-1}P(\exp[tg]g^2) - \left\{ \left[ P \exp[tg] \right]^{-1}P(\exp[tg]g) \right\}^2.
\]

By a Taylor series expansion of the exponential, we have
\[
P \exp[tg] = 1 + tPg + t^2P(\exp[\tilde{t}g]g^2)/2 = 1 + O(t^2)Pg^2 = 1 + O(t^2),
\]
and
\[
P \exp[tg]g = tPg^2 + \frac{t^2}{2}P(\exp[\tilde{t}g]g^3) = tPg^2 + O(t^2)Pg^2 = tPg^2(1 + O(t)).
\]

Combining the pieces, we find that
\[
\left[ P \exp[tg] \right]^{-1}P(\exp[tg]g) = [1 + O(t^2)]^{-1}\left[ tPg^2(1 + O(t)) \right] = tPg^2(1 + O(t)).
\]

It follows that
\[
\left\{ \left[ P \exp[tg] \right]^{-1}P(\exp[tg]g) \right\}^2 = t^2(Pg^2)^2(1 + O(t)) = O(t^2Pg^2).
\]

But $P(\exp[tg]g^2) = Pg^2 + tP(\exp[\tilde{t}g]g^3) = Pg^2(1 + O(t))$, and hence the bound (11.6) follows, which completes the proof.

**Proof of Corollary 7.1.** First we use Lemma 10.2 so see that we may assume the functions are uniformly bounded. Then we can apply Theorem 10.1 to conclude $\tau(\hat{f}) = O(r_0)$. The concentration now follows from Combining Theorem 4.1 with Lemmas 5.2 and 5.3, along with Eq. 11.5 from Lemma 11.2.

**Proof of Corollary 7.2.** The proof is based on the same arguments as used in the proof of Corollary 7.1. We need a slight modification of Theorem 10.1 because the $\epsilon_i$ ($i = 1, \ldots, n$) are not necessarily bounded. Then we conclude $\tau(\hat{f}) = O_P(r_0)$. Next, we apply Theorem 4.1 in conjunction with Lemmas 5.2 and 5.3. To deal with the fact that the $Y_i$ ($i = 1, \ldots, n$) are not necessarily bounded we usethe truncation technique as in Lemma 8.2 and the assumption $E(s) = O(s_0^2)$ for $s = O(s_0)$. 

11.7. Proofs for Section 8

Proof of Lemma 8.1. By Theorem 8.1 of Klein and Rio, for all \( t \geq 0 \), with probability at least \( 1 - \exp[-t] \),

\[
\hat{\mathbb{E}}_n(s) \geq \mathbb{E}(s) - \sqrt{8K\mathbb{E}(s)} + 2\sigma_s^2\sqrt{t/n} - Kt/n.
\]

But since \( J(s)/\left(s^2 - \tau_{\min}^2\right) \) is decreasing, we have for \( s \geq r_0 \)

\[
8K\mathbb{E}(s) \leq 8KJ(s) = 8KJ(r_0)(s^2 - \tau_{\min}^2)/(r_0^2 - \tau_{\min}^2) \leq 2C^2s^2
\]

For \( s \leq r_0 \) we have using that \( J \) is increasing, we find that

\[
8K\mathbb{E}(s) \leq 8KJ(s) \leq 8KJ(r_0) \leq 2C^2r_0^2,
\]

and hence

\[
8K\mathbb{E}(s) \leq 2C^2(s + r_0)^2 \quad \text{for all } s.
\]

Moreover, we have \( \sigma_s^2 \leq C^2s^2 \) by the quadratic curvature condition, and hence

\[
\sqrt{8K\mathbb{E}(s)} + 2\sigma_s^2\sqrt{t/n} \leq \sqrt{4C^2(s + r_0)^2}\sqrt{t/n} = 2C(s + r_0)\sqrt{t/n}.
\]

Proof of Lemma 8.2.

For each \( t > 0 \), we have

\[
(P_n - P)(f^0 - f) \leq (P_n - P)(f^0 - f)1\{F \leq t\} + (P_n - P)F1\{F > t\}
+ 2PF1\{F > t\},
\]

and also

\[
(P_n - P)(f^0 - f) \geq (P_n - P)(f^0 - f)1\{F \leq t\} - (P_n - P)F1\{F > t\}
- 2PF1\{F > t\}.
\]

Taking \( t \) here equal to \( t_0 := C_F\sqrt{\log n} \) (and assuming \( t_0 > 1 \)) we see that

\[
2PF1\{F > t_0\} \leq 2PF^21\{F > t_0\} \leq 2c_F^2/n.
\]

Moreover, for all \( t > 0 \), with probability at least \( 1 - 1/t^2 \), we have

\[
|(P_n - P)F1\{F > t_0\}| \leq t(PF^21\{F > t_0\})^{1/2}/\sqrt{n} \leq t c_F/n.
\]
Write the truncated versions as
\[
\hat{E}^{\text{trunc}}_n(s) := \max_{f \in F_s}(P_n - P)(f^0 - f)I\{F \leq t\}, \quad \text{and}
\]
\[
E^{\text{trunc}}(s) := E\left(\max_{f \in F_s}(P_n - P)(f^0 - f)I\{F \leq t\}\right).
\]
Then \(|E(s) - E^{\text{trunc}}(s)| \leq 2c_F^2/n\), and moreover, with probability at least \(1 - 1/t^2\), we have
\[
|\hat{E}_n(s) - \hat{E}^{\text{trunc}}_n(s)| \leq c_F(2c_F + t)/n.
\]
Now Theorem 8.1 ensures that, for all \(t \geq 0\),
\[
\hat{E}^{\text{trunc}}_n(s) \geq E^{\text{trunc}}(s) - \sqrt{8C_F \log n} E^{\text{trunc}}(s) + 2\sigma^2_s \sqrt{t/n} - C_F t \sqrt{\log n/n},
\]
and
\[
\hat{E}^{\text{trunc}}_n(s) \leq E^{\text{trunc}}(s) + \sqrt{8C_F \log n} E^{\text{trunc}}(s) + 2\sigma^2_s \sqrt{t/n} + 2C_F t \sqrt{\log n/(3n)},
\]
where each bound holds with probability at least \(1 - \exp[-t]\).

11.8. Proofs for Section 9
PROOF OF LEMMA 9.1. Since \(F(\tilde{s}) = E(s)\), it follows that
\[
\tilde{s} - F(\tilde{s}) = s^2 - E(s) - \tau^2_*.
\]
We also have
\[
|s - s_0| = \left|\sqrt{s^2 + \tau^2_*} - \sqrt{s^2 + \gamma^2_*}\right| = \left|\tilde{s} - s_*\right| \frac{\tilde{s} + s_*}{\sqrt{s^2 + \tau^2_*} + \sqrt{\gamma^2_* + \tau^2_*}}
\]
\[
\leq |\tilde{s} - s_*|
\]
PROOF OF THEOREM 9.1. By the oracle potential (see Definition 9.1), we know that for all scalars \(s\) such that \(s^2 \geq \tau^2_* - \tau^2_{\min}\), we have
\[
\sup_{f \in F: \tau^2(f) \leq \tau^2_* + s^2} P(f - f_0) \leq \Gamma^2 s^2,
\]
and hence \(\sup_{f \in F: \tau^2(f) \leq \tau^2_* + s^2} \sigma^2(f - f_0) \leq \Gamma^2 C^2 s^2\). We can thus proceed along the same lines as in the proof of Theorem 4.1, replacing \(C\) by \(\Gamma C\).
11.9. Proofs for Section 10

**Proof of Lemma 10.1.** Consider some \( g \in \mathcal{G} \) such that \( \mathcal{L}_G(g) \leq \mathcal{L}_G(g^*) \) and \( \Upsilon(g - g^*) > 2\epsilon \). It suffices to construct some \( \tilde{g} \in \mathcal{G} \) with \( \Upsilon(\tilde{g} - g^*) \in (\epsilon, 2\epsilon] \) such that \( \mathcal{L}_G(\tilde{g}) \leq \mathcal{L}_G(g^*) \).

Define the rescaled function \( \tilde{g} := \alpha g + (1 - \alpha) g^* \), where \( \alpha := \frac{2\epsilon}{2\epsilon + \Upsilon(g - g^*)} \in (0, 1) \). By convexity of the function class, we have \( \tilde{g} \in \mathcal{G} \), and by convexity of the loss function \( \mathcal{L} \), we have

\[
\mathcal{L}_G(\tilde{g}) \leq \alpha \mathcal{L}(g) + (1 - \alpha) \mathcal{L}_G(g^*) \leq \mathcal{L}_G(g^*).
\]

Moreover, we have the upper bound

\[
\Upsilon(\tilde{g} - g^*) = \frac{2\epsilon \Upsilon(g - g^*)}{2\epsilon + \Upsilon(g - g^*)} \leq 2\epsilon.
\]

Moreover, since the function \( t \mapsto \frac{2\epsilon t}{2\epsilon + t} \) is strictly increasing and \( \Upsilon(g - g^*) > 2\epsilon \), we have

\[
\Upsilon(\tilde{g} - g^*) = \frac{2\epsilon \Upsilon(g - g^*)}{2\epsilon + \Upsilon(g - g^*)} > \frac{4\epsilon^2}{4\epsilon} = \epsilon.
\]

**Proof of Theorem 10.1.** Since the function \( s \mapsto \mathcal{J}(s)/(s^2 - \tau_{\min}^2) \) is decreasing by assumption, for all \( j \in \mathbb{N} \), we have

\[
(2^{j-1}r_0)^2 - \tau_{\min}^2 \geq \frac{1}{4} \frac{r_0^2 - \tau_{\min}^2}{\mathcal{J}(r_0)} \mathcal{J}(2^j r_0) \geq 2 \mathcal{J}(2^j r_0).
\]

Setting \( c_3 = 4 \max\{8c_1, c_2\} \) we see that

\[
c_1(2^j r_0) \left( \frac{2^{j-1} r_0}{c_3} \right) + c_2 \left( \frac{2^{j-1} r_0}{c_3} \right)^2 \leq \frac{(2^{j-1} r_0)^2}{8},
\]

and we get that with probability at least \( 1 - \exp[-n(2^{j-1} r_0)^2/c_3^2] \)

\[
\hat{\mathcal{E}}_n(2^j r_0) \leq \mathcal{J}(2^j r_0) + (2^{j-1} r_0)^2/8.
\]

Therefore, with probability at least \( 1 - \exp[-n(2^{j-1} r_0)^2/c_3^2] \)

\[
(2^{j-1} r_0)^2 - \hat{\mathcal{E}}_n(2^j r_0) \geq (2^{j-1} r_0)^2 - \mathcal{J}(2^j r_0)/m_n - (2^{j-1} r_0)^2/8 \geq (2^{j-1} r_0)^2/2 - \mathcal{J}(2^j r_0)/m_n + r_0^2/4 + r_0^2/8 \geq \tau_{\min}^2 + r_0^2/8.
\]
S. van de Geer and M. J. Wainwright

We moreover know that $\tau^2(\hat{f}) = \arg\min_{s \geq \tau_{\min}} \{ s^2 - \hat{E}_n(s) \}$, and therefore

$$\tau^2(\hat{f}) - \hat{E}_n(\hat{r}) \leq \tau_{\min}^2 - \hat{E}_n(\tau_{\min}).$$

With probability at least $1 - \exp[-nr_0^2/c_3^2]$, we have

$$\hat{E}_n(\tau_{\min}) \geq -c_1 \tau_{\min} \frac{r_0}{c_3} - c_2 \frac{r_0^2}{c_3^2} \geq r_0^2/8,$$

and then

$$\tau^2(\hat{f}) - \hat{E}_n(\hat{r}) \leq \tau_{\min}^2 + r_0^2/8.$$

Thus, if for some $j \in \mathbb{N}$ it holds that $2^{j-1}r_0 < \hat{r} \leq 2^jr_0$ we must have

$$(2^{j-1}s_0)^2 - \hat{E}_n(2^jr_0) < \tau_{\min}^2 + r_0^2/8.$$ 

It follows that

$$\mathbb{P}(\tau(\hat{f}) > r_0) \leq \sum_{j=1}^{\infty} \exp[-nr(2^{j-1}r_0)^2/c_3^2] + \exp[-nr_0^2/c_3^2]$$

$$\leq \frac{2\exp[-nr_0^2/c_3^2]}{1 - \exp[-nr_0^2/c_3^2]}.$$

**Proof of Lemma 10.3.** By the convexity of $\Phi$, for each $u > 0$, we have

$$\frac{\Phi(\alpha u)}{\alpha u} \leq \frac{\alpha \Phi(u)}{\alpha u} = \frac{\Phi(u)}{u}$$

for all $\alpha \in (0, 1)$, so that the function $u \mapsto \Phi(u)/u$ is increasing. Hence also $v \mapsto v/\Phi^{-1}(v)$ is increasing. The mapping $s \mapsto s^2 - \tau_{\min}^2$ is increasing as well. Recalling that $\mathbb{E}(s) = \Phi^{-1}(s^2 - \tau_{\min}^2)$, we conclude that the function $s \mapsto \mathbb{E}(s)/(s^2 - \tau_{\min}^2)$ is decreasing.

Next we observe that for each $t > 0$, the invertibility of $\Phi$ ensures that

$$\Phi^*(t) = \max_u \{ ut - \Phi(u) \} = \max_s \left\{ \Phi^{-1}(s^2 - \tau_{\min}^2)t - (s^2 - \tau_{\min}^2) \right\}. \quad (11.7)$$
When $t = 1$, this maximum is achieved at $s = s_0$. Applying this fact with $t = 1$, we have, for each fixed $\alpha \in (0, 1)$

$$
\Phi^*(1) = \Phi^{-1}(s_0^2 - \tau_{\min}^2) - (s_0^2 - \tau_{\min}^2) \\
= \alpha \frac{\Phi^{-1}(s_0^2 - \tau_{\min}^2) - (s_0^2 - \tau_{\min}^2)}{\alpha} \\
\leq \alpha \left( s_0^2 - \tau_{\min}^2 + \Phi^*(1/\alpha) \right) - (s_0^2 - \tau_{\min}^2) \\
= \alpha \Phi^*(1/\alpha) - (1 - \alpha)(s_0^2 - \tau_{\min}^2).
$$

Re-arranging yields the inequality

$$
s_0^2 - \tau_{\min}^2 \leq \frac{\alpha \Phi^*(1/\alpha) - \Phi^*(1)}{1 - \alpha}.
$$

On the other hand, again using the representation (11.7), we have

$$
\Phi^*(\alpha) \geq \Phi^{-1}(s_0^2 - \tau_{\min}^2)\alpha - (s_0^2 - \tau_{\min}^2) \\
= \alpha \left( \Phi^{-1}(s_0^2 - \tau_{\min}^2) - (s_0^2 - \tau_{\min}^2) \right) - (1 - \alpha)(s_0^2 - \tau_{\min}^2) \\
= \alpha \Phi^*(1) - (1 - \alpha)(s_0^2 - \tau_{\min}^2),
$$

where the last step uses the equivalence (11.8). Now re-arranging implies that

$$
s_0^2 - \tau_{\min}^2 \geq \frac{\alpha \Phi^*(1) - \Phi^*(\alpha)}{1 - \alpha}.
$$

Thus, setting $r_0^2 = 4\tau_{\min}^2 + \Phi^*(16)$ guarantees $r_0 \geq 2\tau_{\min}$ and also

$$
E(r_0) = \Phi^{-1}(r_0^2 - \tau_{\min}^2) \leq \frac{1}{16} \left( r_0^2 - \tau_{\min}^2 + \Phi^*(16) \right) \leq \frac{1}{8} (r_0^2 - \tau_{\min}^2),
$$

as claimed.

**References**

BORELL, C. (1975). The brunn-Minkowski inequality in Gauss space. *Inventiones Mathematicae* **30**, 2, 207–216.

BOUCHERON, S. and MASSART, P. (2011). A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields* **150**, 3-4, 405–433.
BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). Concentration inequalities: A nonasymptotic theory of independence. OUP Oxford.

CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. The Annals of Statistics 42, 6, 2340–2381.

KLEIN, T. (2002). Une inégalité de concentration à gauche pour les processus empiriques. Comptes Rendus Mathematique 334, 6, 501–504.

KLEIN, T. and RIO, E. (2005). Concentration around the mean for maxima of empirical processes. The Annals of Probability 33, 3, 1060–1077.

KOLTCHINSKII, V. (2011). Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d’Esté de probabilités de Saint-Flour XXXVIII-2008, volume 38 Springer Science & Business Media.

LEDOUX, M. (2001). The concentration of measure phenomenon, volume 89. American Mathematical Society.

MASSART, P. (2000). Some applications of concentration inequalities to statistics. Annales de la faculté des sciences de toulouse: Mathématiques, volume 9, pages 245–303.

MURO, A. and VAN DE GEER, S. (2015). Concentration behavior of the penalized least squares estimator. arXiv:1511.08698.

ROCKAFELLAR, R.T. (1970). Convex analysis. Princeton University Press.

SAUMARD, A. (2012). Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. Electronic Journal of Statistics 6, 579–655.

TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. Publications Mathématiques de l’IHES 81, 73–205.

VAN DER VAART, A.W. and WELLNER, J.A. (1996). Weak Convergence and Empirical Processes. Springer Series in Statistics. Springer-Verlag, New York. ISBN 0-387-94640-3.

Sara van de Geer
Seminar for Statistics, ETH Zürich, Zurich, Switzerland
E-mail: geer@stat.math.ethz.ch

Martin J. Wainwright
Department of Statistics and
Department of EECS,
University of California, Berkeley,
CA 94720, USA
E-mail: wainwrig@Berkeley.EDU

Paper received: 25 January 2017.