Empirical risk minimization for heavy-tailed losses

Christian Brownlees Emilien Joly Gábor Lugosi

June 8, 2014

Abstract

The purpose of this paper is to discuss empirical risk minimization when the losses are not necessarily bounded and may have a distribution with heavy tails. In such situations usual empirical averages may fail to provide reliable estimates and empirical risk minimization may provide large excess risk. However, some robust mean estimators proposed in the literature may be used to replace empirical means. In this paper we investigate empirical risk minimization based on a robust estimate proposed by Catoni. We develop performance bounds based on chaining arguments tailored to Catoni's mean estimator.

1 Introduction

One of the basic principles of statistical learning is empirical risk minimization that has been routinely used in a great variety of problems such as regression function estimation, classification, and clustering. The general model may be described as follows. Let X be a random variable taking values in some measurable space $\mathcal{X}$ and let $\mathcal{F}$ be a set of non-negative functions defined on $\mathcal{X}$. For each $f \in \mathcal{F}$, define the *risk* $m_f = \mathbb{E}f(X)$ and let $m^* = \inf_{f\in\mathcal{F}} m_f$ denote the optimal risk. In statistical learning n independent random variables $X_1, \ldots, X_n$ are available, all distributed as X, and one aims at finding a function with small risk. To this end, one may define the *empirical risk minimizer*

$$f_{\mathrm{ERM}} = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

where, for the simplicity of the discussion and essentially without loss of generality, we implicitly assume that the minimizer exists. If the minimum is achieved by more than one function, one may pick one of them arbitrarily.

Remark. (LOSS FUNCTIONS AND RISKS.) The main motivation and terminology may be explained by the following general prediction problem in statistical learning. Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ be independent identically distributed pairs of random variables representing "training data" where the Z_i take their values in, say, $\mathbb{R}^d$ and the Y_i are real-valued. In classification problems the Y_i take discrete values. Given a new observation Z,

one is interested in predicting the value of the corresponding response variable Y where the pair (Z, Y) has the same distribution as that of the (Z_i, Y_i). A predictor is a function $g : \mathbb{R}^d \to \mathbb{R}$ whose quality is measured with the help of a *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. The *risk* of g is then $\mathbb{E}\ell(g(Z), Y)$. Given a class $\mathcal{G}$ of functions $g : \mathbb{R}^d \to \mathbb{R}$, empirical risk minimization chooses one that minimizes the *empirical risk* $(1/n)\sum_{i=1}^n \ell(g(Z_i), Y_i)$ over all $g \in \mathcal{G}$. In the simplified notation followed in this paper, X_i corresponds to the pair (Z_i, Y_i), the function f represents $\ell(g(\cdot), \cdot)$, and m_f substitutes $\mathbb{E}\ell(g(Z), Y)$.

The performance of empirical risk minimization is measured by the *risk* of the selected function,

$$m_{\mathrm{ERM}} = \mathbb{E}\left[f_{\mathrm{ERM}}(X)|X_1, \ldots, X_n\right] \ .$$

In particular, the main object of interest for this paper is the *excess risk* $m_{\mathrm{ERM}} - m^*$. The performance of empirical risk minimization has been thoroughly studied and well understood using tools of empirical process theory. In particular, the simple observation that

$$m_{\mathrm{ERM}} - m^* \leq 2 \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - m_f \right|$$

allows one to apply the rich theory on the suprema of empirical processes to obtain upper performance bounds. The interested reader is referred to Bartlett and Mendelson [6], Boucheron, Bousquet, and Lugosi [8], Koltchinskii [14], Massart [18], Mendelson [21], van de Geer [29] for references and recent results in this area. Essentially all of the theory of empirical minimization assumes either that the functions f are uniformly bounded or that the random variables $f(X)$ have sub-Gaussian tails for all $f \in \mathcal{F}$. For example, when all $f \in \mathcal{F}$ take their values in the interval $[0, 1]$, Dudley's [12] classical metric-entropy bound, together with standard symmetrization arguments, imply that there exists a universal constant c such that

$$\mathbb{E}m_{\mathrm{ERM}} - m^* \leq \frac{c}{\sqrt{n}} \mathbb{E} \int_0^1 \sqrt{\log N_{\mathbb{X}}(\mathcal{F}, \epsilon)} d\epsilon \ , \tag{1}$$

where for any $\epsilon > 0$, $N_{\mathbb{X}}(\mathcal{F}, \epsilon)$ is the *ϵ-covering number* of the class $\mathcal{F}$ under the empirical quadratic distance $d_{\mathbb{X}}(f, g) = \left(\frac{1}{n}\sum_{i=1}^n (f(X_i) - g(X_i))^2\right)^{1/2}$, defined as the minimal cardinality N of any set $\{f_1, \ldots, f_N\} \subset \mathcal{F}$ such that for all $f \in \mathcal{F}$ there exists an $f_j \in \{f_1, \ldots, f_N\}$ with $d_{\mathbb{X}}(f, f_j) \leq \epsilon$. Of course, this is one of the most basic bounds and many important refinements have been established.

A tighter bound may be established by the so-called *generic chaining* method, see Talagrand [27]. Recall the following definition (see, e.g., [27, Definition 1.2.3]). Let T be a (pseudo) metric space. An increasing sequence $(\mathcal{A}_n)$ of partitions of T is called *admissible* if for all $n = 0, 1, 2, \ldots$, $\#\mathcal{A}_n \leq 2^{2^n}$. For any $t \in T$, denote by $A_n(t)$ the unique element of $\mathcal{A}_n$ that contains t. Let $\Delta(A)$ denote the diameter of the set $A \subset T$. Define, for $\alpha = 1, 2$,

$$\gamma_\alpha(T, d) = \inf_{\mathcal{A}_n} \sup_{t \in T} \sum_{n \geq 0} 2^{n/\alpha} \Delta(A_n(t)) \ ,$$

where the infimum is taken over all admissible sequences. Then one has

$$\mathbb{E}m_{\mathrm{ERM}} - m^* \leq \frac{c}{\sqrt{n}}\mathbb{E}\gamma_2(\mathcal{F}, d_{\mathbb{X}}) \tag{2}$$

for some universal constant c. This bound implies (1) as $\gamma_2(\mathcal{F}, d_{\mathbb{X}})$ is bounded by a constant multiple of the entropy integral $\int_0^1 \sqrt{\log N_{\mathbb{X}}(\mathcal{F}, \epsilon)}d\epsilon$.

However, when the functions f are no longer uniformly bounded and the random variables $f(X)$ may have a heavy tail, empirical risk minimization may have a much poorer performance. This is simply due to the fact that empirical averages become poor estimates of expected values. Indeed, for heavy-tailed distributions, several estimators of the mean are known to outperform simple empirical averages. It is a natural idea to define a robust version of empirical risk minimization based on minimizing such robust estimators.

In this paper we focus on an elegant and powerful estimator proposed and analyzed by Catoni [11]. (A version of) Catoni's estimator may be defined as follows.

Introduce the non-decreasing differentiable *truncation function*

$$\phi(x) = -\mathbb{1}_{\{x<0\}} \log\left(1 - x + \frac{x^2}{2}\right) + \mathbb{1}_{\{x\geq 0\}} \log\left(1 + x + \frac{x^2}{2}\right) \ . \tag{3}$$

To estimate $m_f = \mathbb{E}f(X)$ for some $f \in \mathcal{F}$, define, for all $\mu \in \mathbb{R}$,

$$\widehat{r}_f(\mu) = \frac{1}{n\alpha}\sum_{i=1}^{n} \phi(\alpha(f(X_i) - \mu))$$

where $\alpha > 0$ is a parameter of the estimator to be specified below. Catoni's estimator of m_f is defined as the unique value $\widehat{\mu}_f$ for which $\widehat{r}_f(\widehat{\mu}_f) = 0$. (Uniqueness is ensured by the monotonicity of $\mu \mapsto \widehat{r}_f(\mu)$). Catoni proves that for any fixed $f \in \mathcal{F}$ and $\delta \in [0, 1]$ such that $n > 2\log(1/\delta)$, under the only assumption that $\mathrm{Var}\left(f(X)\right) \leq v$, the estimator above with

$$\alpha = \sqrt{\frac{2\log(1/\delta)}{n\left(v + \frac{2v\log(1/\delta)}{n(1-(2/n)\log(1/\delta))}\right)}}$$

satisfies that, with probability at least $1 - 2\delta$,

$$|m_f - \widehat{\mu}_f| \leq \sqrt{\frac{2v\log(1/\delta)}{n(1 - (2/n)\log(1/\delta))}} \ . \tag{4}$$

In other words, the deviations of the estimate exhibit a sub-Gaussian behavior. The price to pay is that the estimator depends both on the upper bound v for the variance and on the prescribed confidence δ via the parameter α.

Catoni also shows that for any $n > 4(1 + \log(1/\delta))$, if $\mathrm{Var}\left(f(X)\right) \leq v$, the choice

$$\alpha = \sqrt{\frac{2}{nv}}$$

guarantees that, with probability at least $1 - 2\delta$,

$$|m_f - \widehat{\mu}_f| \le (1 + \log(1/\delta)) \sqrt{\frac{v}{n}} \; . \tag{5}$$

Even though we lose the sub-Gaussian tail behavior, the estimator is independent of the required confidence level.

Given such a powerful mean estimator, it is natural to propose an empirical risk minimizer that selects a function from the class $\mathcal{F}$ that minimizes Catoni's mean estimator. Formally, define

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mu}_f$$

where again, for the sake of simplicity we assume that the minimizer exists. (Otherwise one may select an appropriate approximate minimizer and all arguments go through in a trivial way.)

Once again, as a first step of understanding the excess risk $m_{\widehat{f}} - m^*$, we may use the simple bound

$$m_{\widehat{f}} - m^* = \left(m_{\widehat{f}} - \widehat{\mu}_{\widehat{f}} \right) + \left(\widehat{\mu}_{\widehat{f}} - m^* \right) \le 2 \sup_{f \in \mathcal{F}} |m_f - \widehat{\mu}_f| \; .$$

When $\mathcal{F}$ is a finite class of cardinality, say $|\mathcal{F}| = N$, Catoni's bound may be combined, in a straightforward way, with the union-of-events bound. Indeed, if the estimators $\widehat{\mu}_f$ are defined with parameter

$$\alpha = \sqrt{\frac{2 \log(N/\delta)}{n \left(v + \frac{2v \log(N/\delta)}{n(1 - (2/n) \log(N/\delta))} \right)}} \; ,$$

then, with probability at least $1 - 2\delta$,

$$\sup_{f \in \mathcal{F}} |m_f - \widehat{\mu}_f| \le \sqrt{\frac{2v \log(N/\delta)}{n(1 - (2/n) \log(N/\delta))}} \; .$$

Note that this bound requires that $\sup_{f \in \mathcal{F}} \operatorname{Var}(f(X)) \le v$, that is, the variances are uniformly bounded by a *known* value v. Throughout the paper we work with this assumption. However, this bound does not take into account the structure of the class $\mathcal{F}$ and it is useless when $\mathcal{F}$ is an infinite class. Our strategy to obtain meaningful bounds is to use *chaining* arguments. However, the extension is nontrivial and the argument becomes more involved. The main results of the paper present performance bounds for empirical minimization of Catoni's estimator based on generic chaining.

Remark. (MEDIAN-OF-MEANS ESTIMATOR.) Catoni's estimator is not the only one with sub-Gaussian deviations for heavy-tailed distributions. Indeed, the *median-of-means* estimator, proposed by Nemirovsky and Yudin [23] (and also independently by Alon, Matias,

and Szegedy [2]) has similar performance guarantees as (4). This estimate is obtained by dividing the data in several small blocks, calculating the sample mean within each block, and then taking the median of these means. Hsu and Sabato [13] and Minsker [22] introduce multivariate generalizations of the median-of-means estimator and use it to define an analyze certain statistical learning procedures in the presence of heavy-tailed data. The sub-Gaussian behavior is achieved under various assumptions on the loss function. Such conditions can be avoided here. As an example, we detail applications of our results theorems in Section 4 for three different classes of loss functions. An important advantage of the median-of-means estimate over Catoni's estimate is that the parameter of the estimate (i.e., the number of blocks) only depends on the confidence level δ but not on v and therefore no prior upper bound of the variance v is required to compute this estimate. Also, the median-of-means estimate is useful even when the variance is infinite and only a moment of order $1+\epsilon$ exists for some $\epsilon > 0$ (see Bubeck, Cesa-Bianchi, and Lugosi [10]). Lerasle and Oliveira [15] consider empirical minimization of the median-of-means estimator and obtain interesting results in various statistical learning problems. However, to establish metric-entropy bounds for minimization of this mean estimate remains to be a challenge.

The rest of the paper is organized as follows. In Section 2 we state and discuss the main results of the paper. Section 3 is dedicated to the proofs. In Section 4 we describe some applications to regression under the absolute and squared losses and k-means clustering. Finally, in Section 5 we present some simulation results both for regression and k-means clustering. Some of the more technical arguments are relegated to the Appendix.

2 Main results

The bounds we establish for the excess risk depend on the geometric structure of the class $\mathcal{F}$ under different distances. The $L_2(P)$ distance is defined, for $f, f' \in \mathcal{F}$, by

$$d(f,f') = \left(\mathbb{E}\left[\left(f(X)-f'(X)\right)^2\right]\right)^{1/2}$$

and the L_∞ distance is

$$D(f,f') = \sup_{x\in\mathcal{X}} \left|f(x)-f'(x)\right| \ .$$

We also work with the (random) empirical quadratic distance

$$d_{\mathbb{X}}(f,f') = \left(\frac{1}{n}\sum_{i=1}^n (f(X_i)-f'(X_i))^2\right)^{1/2} \ .$$

Denote by f^* a function with minimal expectation

$$f^* = \operatorname*{argmin}_{f\in\mathcal{F}} m_f \ .$$

Next we present two results that bound the excess risk $m_{\widehat{f}} - m_{f^*}$ of the minimizer $\widehat{f}$ of Catoni's risk estimate in terms of metric properties of the class $\mathcal{F}$. The first result involves a combination of terms involving the γ_2 and γ_1 functionals under the metrics d and D while the second is in terms of quantiles of γ_2 under the empirical metric $d_{\mathbb{X}}$.

Theorem 1. *Let $\mathcal{F}$ be a class of non-negative functions defined on a set $\mathcal{X}$ and let $X, X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\mathcal{X}$. Assume that there exists $v > 0$ such that $\sup_{f\in\mathcal{F}} Var(f(X)) \leq v$. Let $\delta \in (0, 1/3)$. Suppose that $\widehat{f}$ is selected from $\mathcal{F}$ by minimizing Catoni's mean estimator with parameter $\alpha < 1$. Then there exist a universal constant $L \leq 384\log(2)$ such that, with probability at least $1 - 3\delta$ and for n large enough, the risk of $\widehat{f}$ satisfies*

$$m_{\widehat{f}} - m_{f^*} \leq 6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + \log(\delta^{-1})\left(\frac{8L}{\sqrt{n}}\gamma_2(\mathcal{F}, d) + \frac{4L}{3n}\gamma_1(\mathcal{F}, D)\right) .$$

Theorem 2. *Assume the hypotheses of Theorem 1. Set Γ_δ such that $\mathbb{P}\{\gamma_2(T, d_{\mathbb{X}}) > \Gamma_\delta\} \leq \frac{\delta}{2}$. Then there exist a universal constant K such that, with probability at least $1 - 3\delta$ and for n large enough, the risk of $\widehat{f}$ satisfies*

$$m_{\widehat{f}} - m_{f^*} \leq 6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + K\Gamma_\delta\sqrt{\frac{\log(\frac{2}{\delta})}{n}} .$$

In both theorems above, the choice of α only influences the term $\alpha v + 2\log(\delta^{-1})/(n\alpha)$. By taking $\alpha = \sqrt{2\log(\delta^{-1})/(nv)}$, this term equals

$$2\sqrt{\frac{2v\log(\delta^{-1})}{n}} .$$

This choice has the disadvantage that the estimator depends on the confidence level. By taking $\alpha = \sqrt{2/(nv)}$, one obtains the term

$$\sqrt{\frac{2v}{n}}(1 + \log(\delta^{-1})) .$$

Observe that the main term in the second part of the bound of Theorem 1 is

$$\left(\log\frac{1}{\delta}\right)\frac{L}{\sqrt{n}}\gamma_2(\mathcal{F}, d)$$

which is comparable to the bound (2) obtained under the strong condition of $f(X)$ being uniformly bounded. All other terms are of smaller order. Note that this part of the bound depends on the "weak" distribution-dependent $L_2(P)$ metric d. The quantity $\gamma_1(\mathcal{F}, D) \geq \gamma_2(\mathcal{F}, d)$ also enters the bound of Theorem 1 though only multiplied by $1/n$. The presence of this term requires that $\mathcal{F}$ is bounded in the L_∞ distance D which limits the usefulness

of the bound. In Section 4 we illustrate the bounds on two applications to regression and k-means clustering. In these applications, in spite of the presence of heavy tails, the covering numbers under the distance D may be bounded in a meaningful way. Note that no such bound can hold for "ordinary" empirical risk minimization that minimizes the usual empirical means $(1/n)\sum_{i=1}^n f(X_i)$ because of the poor performance of empirical averages in the presence of heavy tails.

The main merit of the bound of Theorem 2 is that it does not require that the class $\mathcal{F}$ has a finite diameter under the supremum norm. Instead, the quantiles of $\gamma_2(\mathcal{F}, d_{\mathbb{X}})$ enter the picture. In Section 4 we show it through the example of L_2 regression how these quantiles may be estimated.

3 Proofs

The proofs of Theorems 1 and 2 are based on showing that the excess risk can be bounded as soon as the supremum of the empirical process $\{X_f(\mu) : f \in \mathcal{F}\}$ is bounded for any fixed $\mu \in \mathbb{R}$, where for any $f \in \mathcal{F}$ and $\mu \in \mathbb{R}$, we define $X_f(\mu) = \widehat{r}_f(\mu) - \overline{r}_f(\mu)$ with

$$\overline{r}_f(\mu) = \frac{1}{\alpha}\mathbb{E}\left[\phi(\alpha(f(X) - \mu))\right]$$

and

$$\widehat{r}_f(\mu) = \frac{1}{n\alpha}\sum_{i=1}^n \phi(\alpha(f(X_i) - \mu)) \ .$$

The two theorems differ in the way the supremum of this empirical process is bounded.

Note first that, by the definition of Catoni's estimator, $\widehat{\mu}_f \geq \inf_i f(X_i)$. In particular, $\widehat{\mu}_f \geq 0$ for all $f \in \mathcal{F}$. Let $A_\alpha(\delta) = \alpha v + 2\log(\delta^{-1})/(n\alpha)$.

Once again, we may assume, essentially without loss of generality, that the minimum exists. In case of multiple minimizers we may choose one arbitrarily. The main result in [11] states that for any $\delta > 0$ such that $\alpha^2 v + 2\log(\delta^{-1})/n \leq 1$, with probability at least $1 - 2\delta$,

$$|\widehat{\mu}_{f^*} - m_{f^*}| \leq A_\alpha(\delta) \ . \tag{6}$$

3.1 A deterministic version of $\widehat{\mu}_f$

We begin with a variant of the argument of Catoni [11]. It involves a deterministic version $\overline{\mu}_f$ of the estimator defined, for each $f \in \mathcal{F}$, as the unique solution of the equation $\overline{r}_f(\mu) = 0$.

In Lemma 4 below we show that $\overline{\mu}_f$ is in a small (deterministic) interval centered at m_f. First we recall a fact from [11] in the next proposition. For any $f \in \mathcal{F}$, $\mu \in \mathbb{R}$, and

$\varepsilon > 0$, define

$$B_f^+(\mu,\varepsilon) = (m_f - \mu) + \frac{\alpha}{2}(m_f-\mu)^2 + \frac{\alpha}{2}v + \varepsilon \ ,$$
$$B_f^-(\mu,\varepsilon) = (m_f - \mu) - \frac{\alpha}{2}(m_f-\mu)^2 - \frac{\alpha}{2}v - \varepsilon$$

and let

$$\mu_f^+(\varepsilon) = m_f + \alpha v + 2\varepsilon \ , \qquad \mu_f^-(\varepsilon) = m_f - \alpha v - 2\varepsilon \ .$$

As a function of μ, $B_f^+(\mu,\varepsilon)$ is a quadratic polynomial such that $\mu_f^+(\varepsilon)$ is an upper bound of the smallest root of $B_f^+(\mu,\varepsilon)$. Similarly, $\mu_f^-(\varepsilon)$ is a lower bound of the largest root of $B_f^-(\mu,\varepsilon)$. Implicitly we assumed that these roots always exist. This is not always the case but a simple condition on α guarantees that these roots exists. In particular, $1 - \alpha^2 v - 2\alpha\varepsilon \geq 0$ guarantees that $B_f^+(\mu,\varepsilon) = 0$ and $B_f^-(\mu,\varepsilon) = 0$ have at least one solution. This condition will always be satisfied by our choice of ϵ and α.

In our notation, Proposition 2.2 in [11] is equivalent to the following.

Proposition 3. *Let $\delta > 0$ and $\mu \in \mathbb{R}$. For any $f \in \mathcal{F}$, the events*

$$\Omega_f^-(\mu,\delta) = \left\{ B_f^-\left(\mu, \frac{\log \delta^{-1}}{n\alpha}\right) \leq \widehat{r}_f(\mu) \right\}$$
$$\Omega_f^+(\mu,\delta) = \left\{ \widehat{r}_f(\mu) \leq B_f^+\left(\mu, \frac{\log \delta^{-1}}{n\alpha}\right) \right\}$$

both hold with probability at least $1-\delta$.

Let $\varepsilon = \frac{\log \delta^{-1}}{n\alpha}$ and define

$$\Omega_{f^*}(\delta) \stackrel{\text{def}}{=} \Omega_f^-(\mu_{f^*}^-(\varepsilon),\delta) \ \cap \ \Omega_f^+(\mu_{f^*}^+(\varepsilon),\delta) \ .$$

If $\alpha^2 v + \frac{2\log\delta^{-1}}{n} \leq 1$, (6) holds on the event $\Omega_{f^*}(\delta)$. (Just replace ε by $\frac{\log\delta^{-1}}{n\alpha}$ in the expression of $\mu_{f^*}^+(\varepsilon)$ and $\mu_{f^*}^-(\varepsilon)$.) Since $\widehat{\mu}_{f^*}$ is the unique zero of $\widehat{r}_{f^*}(\mu)$, it is squeezed into the interval $[\mu_{f^*}^-(\varepsilon) \ , \ \mu_{f^*}^+(\varepsilon)]$ centered at m_{f^*} and of size $2A_\alpha(\delta)$. Note that $\mathbb{P}\{\Omega_{f^*}(\delta)\} \geq 1 - 2\delta$.

Still following ideas of [11], the next lemma bounds $\overline{r}_f(\mu)$ by the quadratic polynomials B^+ and B^-. The lemma will help us compare the zero of $\overline{r}_f(\mu)$ to the zeros of these quadratic functions.

Lemma 4. *For any fixed $f \in \mathcal{F}$ and $\mu \in \mathbb{R}$,*

$$B_f^-(\mu,0) \leq \overline{r}_f(\mu) \leq B_f^+(\mu,0) \ , \tag{7}$$

and therefore $m_f - \alpha v \leq \overline{\mu}_f \leq m_f + \alpha v$. In particular,

$$B_{\widehat{f}}^-(\mu,0) \leq \overline{r}_{\widehat{f}}(\mu) \leq B_{\widehat{f}}^+(\mu,0) \ .$$

For any μ such that $\overline{r}_{\widehat{f}}(\mu) \leq \varepsilon$, if $1 - \alpha^2 v - 2\alpha\varepsilon \geq 0$, then

$$m_{\widehat{f}} \leq \mu + \alpha v + 2\varepsilon \ . \tag{8}$$

Proof. Writing Y for $\alpha(f(X) - \mu)$ and using the fact that $\phi(x) \leq \log(1 + x + x^2/2)$ for all $x \in \mathbb{R}$,

$$\begin{aligned}
\exp\left(\alpha \overline{r}_f(\mu)\right) &\leq \exp\left(\mathbb{E}\left[\log(1 + Y + \frac{Y^2}{2})\right]\right) \\
&\leq \mathbb{E}\left[1 + Y + \frac{Y^2}{2}\right] \\
&\leq 1 + \alpha(m_f - \mu) + \frac{\alpha^2}{2}[v + (m_f - \mu)^2] \\
&\leq \exp(\alpha B_f^+(\mu, 0)) \ .
\end{aligned}$$

Thus, we have $\overline{r}_f(\mu) - B_f^+(\mu) \leq 0$. Since this last inequality is true for any f, $\sup_f(\overline{r}_f(\mu) - B_f^+(\mu)) \leq 0$ and the second inequality of (7) is proved. The other part can be treated with the same argument.
If $\overline{r}_{\widehat{f}}(\mu) \leq \varepsilon$ then $B_{\widehat{f}}^-(\mu, 0) \leq \varepsilon$ which is equivalent to $B_{\widehat{f}}^-(\mu, \varepsilon) \leq 0$. If $1 - \alpha^2 v - 2\alpha\varepsilon \geq 0$ then a solution of $B_{\widehat{f}}^-(\mu, \varepsilon) = 0$ exists and since $\overline{r}_{\widehat{f}}(\mu)$ is a non-increasing function, μ is above the largest of these two solutions. This implies $\mu_{\widehat{f}}^-(\varepsilon) \leq \mu$ which gives inequality (8). □

The last inequality (8) is the key tool to ensure that the risk $m_{\widehat{f}}$ of the minimizer $\widehat{f}$ can be upper bounded as soon as $\bar{r}_{\widehat{f}}$ is. It remains to find the smallest μ and ε such that $\bar{r}_f(\mu)$ is bounded uniformly on $\mathcal{F}$.

3.2 Bounding the excess risk in terms of the supremum of an empirical process

The key to all proofs is that we link the excess risk to the supremum of the empirical process $X_f(\mu) = \widehat{r}_f(\mu) - \overline{r}_f(\mu)$ as f ranges through $\mathcal{F}$ for a suitably chosen value of μ. For fixed $\mu \in \mathbb{R}$ and $\delta \in (0, 1)$, define the $1 - \delta$ quantile of $\sup_{f \in \mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)|$ by $Q(\mu, \delta)$, that is, the infimum of all positive numbers such that

$$\mathbb{P}\left\{\sup_{f \in \mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)| \leq Q(\mu, \delta)\right\} \geq 1 - \delta \ .$$

First we need a few simple facts summarized in the next lemma.

Lemma 5. *Let $\mu_0 = m_{f^*} + A_\alpha(\delta)$. Then on the event $\Omega_{f^*}(\delta)$, the following inequalities hold:*

1. $\widehat{r}_{\widehat{f}}(\mu_0) \leq 0$
2. $\overline{r}_{f^*}(\mu_0) \leq 0$
3. $-\widehat{r}_{f^*}(\mu_0) \leq 2A_{f^*}(\delta)$

Proof. We prove each inequality separately.

1. First note that on $\Omega_{f^*}(\delta)$ equation (6) holds and we have $\hat{\mu}_{\widehat{f}} \leq \hat{\mu}_{f^*} \leq \mu_0$ and $\widehat{\mu}_{f^*} \leq \mu_0$. By definition $\widehat{\mu}_{f^*} \geq \widehat{\mu}_{\widehat{f}}$. Since $\widehat{r}_{\widehat{f}}$ is a non-increasing function of μ, $\widehat{r}_{\widehat{f}}(\mu_0) \leq \widehat{r}_{\widehat{f}}(\widehat{\mu}_{\widehat{f}}) = 0$.

2. By (7), $\overline{\mu}_{f^*} \leq m_{f^*} + \alpha v \leq m_{f^*} + \alpha v + \frac{2\log(\delta^{-1})}{n\alpha} = \mu_0$. Since $\overline{r}_{f^*}$ is a non-increasing function, $\overline{r}_{f^*}(\mu_0) \leq \overline{r}_{f^*}(\overline{\mu}_{f^*}) = 0$.

3. $\widehat{r}_{f^*}$ is a 1-Lipschitz function and therefore

$$\begin{aligned} |\widehat{r}_{f^*}(\mu_0)| = |\widehat{r}_{f^*}(\widehat{\mu}_{f^*}) - \widehat{r}_{f^*}(\mu_0)| &\leq |\widehat{\mu}_{f^*} - \mu_0| \\ &\leq |\widehat{\mu}_{f^*} - m_{f^*}| + |m_{f^*} - \mu_0| \\ &\leq 2A_{f^*}(\delta) \end{aligned}$$

   which gives $-\widehat{r}_{f^*}(\mu_0) \leq 2A_{f^*}(\delta)$.

□

We will use Lemma 4 with μ_0 introduced in Lemma 5.
With the notation introduced above, we see that with probability at least $1 - \delta$,

$$\begin{aligned} \overline{r}_{\widehat{f}}(\mu_0) &= \widehat{r}_{\widehat{f}}(\mu_0) + \overline{r}_{f^*}(\mu_0) - \widehat{r}_{f^*}(\mu_0) + \left|\overline{r}_{\widehat{f}}(\mu_0) - \widehat{r}_{\widehat{f}}(\mu_0) - \overline{r}_{f^*}(\mu_0) + \widehat{r}_{f^*}(\mu_0)\right| \\ &\leq \widehat{r}_{\widehat{f}}(\mu_0) + \overline{r}_{f^*}(\mu_0) - \widehat{r}_{f^*}(\mu_0) + \sup_{f\in\mathcal{F}} |\overline{r}_f(\mu_0) - \widehat{r}_f(\mu_0) - \overline{r}_{f^*}(\mu_0) + \widehat{r}_{f^*}(\mu_0)| \\ &\leq \widehat{r}_{\widehat{f}}(\mu_0) + \overline{r}_{f^*}(\mu_0) - \widehat{r}_{f^*}(\mu_0) + Q(\mu,\delta) \ . \end{aligned}$$

This inequality, together with Lemma 5, implies that with probability at least $1 - 3\delta$,

$$\overline{r}_{\widehat{f}}(\mu_0) \leq 2A_{f^*}(\delta) + Q(\mu,\delta) \ .$$

Now using Lemma 4 under the condition $1 - \alpha^2 v - 4\alpha A_{f^*}(\delta) - 2\alpha Q(\mu,\delta) \geq 0$ we have

$$\begin{aligned} m_{\widehat{f}} - m_{f^*} &\leq \alpha v + 5A_{f^*}(\delta) + 2Q(\mu,\delta) \\ &\leq 6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + 2Q(\mu,\delta) \ , \end{aligned} \tag{9}$$

with probability at least $1 - 3\delta$. The condition $1 - \alpha^2 v - 4\alpha A_{f^*}(\delta) - 2\alpha Q(\mu,\delta) \geq 0$ is implied (since $\alpha \leq 1$) by $6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + 2Q(\mu,\delta) \leq 1$ which will be seen to hold for sufficiently large n.

3.3 Bounding the supremum of the empirical process

Theorems 1 and 2 both follow from (9) by two different ways of bounding the quantile $Q(\mu, \delta)$ of $\sup_{f \in \mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)|$. He we present these two inequalities. Both of them use basic results of "generic chaining", see Talagrand [27]. Theorem 1 follows from (9) and the next inequality:

Proposition 6. *Let $\mu \in R$ and $\alpha > 0$. There exist a universal constant $L < 384 \log 2$ such that for any $\delta \in (0,1)$,*

$$Q(\mu, \delta) \leq \log(\delta^{-1}) \left(\frac{4L}{\sqrt{n}} \gamma_2(\mathcal{F}, d) + \frac{2L}{3n} \gamma_1(\mathcal{F}, D) \right) .$$

The proof is an immediate consequence of Theorem 13 and (14) in the Appendix and the following lemma.

Lemma 7. *For any $\mu \in \mathbb{R}$, $\alpha > 0$, $f, f' \in \mathcal{F}$, and $t > 0$,*

$$\mathbb{P}\left\{ |X_f(\mu) - X_{f'}(\mu)| > t \right\} \leq \exp\left(- \frac{nt^2}{2(4d(f,f')^2 + \frac{2D(f,f')t}{3})} \right)$$

where the distances d, D are defined at the beginning of Section 2.

Proof. Observe that $n(X_f - X_{f'})$ is the sum of the independent zero-mean random variables

$$C_i(f,f') = \frac{1}{\alpha}\phi(\alpha(f(X_i)-\mu)) - \frac{1}{\alpha}\phi(\alpha(f'(X_i)-\mu)) - \left[\frac{1}{\alpha}\mathbb{E}\left[\phi(\alpha(f(X)-\mu))\right] - \frac{1}{\alpha}\mathbb{E}\left[\phi(\alpha(f'(X)-\mu))\right] \right] .$$

Note that since the truncation function ϕ is 1-Lipschitz, we have $C_i(f,f') \leq 2D(f,f')$. Also,

$$\sum_{i=1}^n \mathbb{E}[C_i(f,f')^2] \leq 4 \sum_{i=1}^n \mathbb{E}\left[\left((f(X_i)-\mu) - (f'(X_i)-\mu) \right)^2 \right] = 4nd(f,f')^2$$

The lemma follows from Bernstein's inequality (see, e.g., [9, Theorem 2.10]). □

Similarly, Theorem 2 is implied by (9) and the following. Recall the notation of Theorem 2.

Theorem 8. *Let $\mu \in \mathbb{R}$, $\alpha > 0$, and $\delta \in (0,1)$. There exists a universal constant $K \leq 384\sqrt{32}\log(2)$ such that*

$$Q(\mu, \delta) \leq K \Gamma_\delta \sqrt{\frac{\log(\frac{2}{\delta})}{n}} .$$

Proof. The proof is based on a standard symmetrization argument. Let $(X_1', \dots, X_n')$ be independent copies of $(X_1, \dots, X_n)$ and define

$$Z_i(f) = \frac{1}{\alpha}\phi(\alpha(f(X_i) - \mu)) - \frac{1}{\alpha}\phi(\alpha(f(X_i') - \mu)) \ .$$

Introduce also independent Rademacher random variables $(\varepsilon_1, \dots, \varepsilon_n)$. For any $f \in \mathcal{F}$, denote by $Z(f) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i Z_i(f)$. Then by Hoeffding's inequality, for all $f, g \in \mathcal{F}$ and for every $t > 0$,

$$\mathbb{P}_{(\varepsilon_1,\dots,\varepsilon_n)}\left\{|Z(f) - Z(g)| > t\right\} \leq 2\exp\left(-\frac{nt^2}{2d_{X,X'}(f,g)^2}\right) \tag{10}$$

where $\mathbb{P}_{(\varepsilon_1,\dots,\varepsilon_n)}$ denotes probability with respect to the Rademacher variables only (i.e., conditional on the X_i and X_i') and $d_{\mathbb{X},\mathbb{X}'}(f,g) = \sqrt{\frac{1}{n}\sum_{i=1}^n (Z_i(f) - Z_i(g))^2}$ is a random distance.

Denote by $\widehat{r}_f'(\mu)$ the independent copy of $\widehat{r}_f(\mu)$ that depends only on the random vector $(X_1', \dots, X_n')$. Let $\lambda > 0$ be a parameter that we optimize later.

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|X_f - X_{f^*}| \geq t\right\} \\
&\leq \mathbb{E}\left[e^{\lambda\sup_{f\in\mathcal{F}}\left|(\widehat{r}_f(\mu)-\mathbb{E}\left[\widehat{r}_f(\mu)\right])-(\widehat{r}_{f^*}(\mu)-\mathbb{E}\left[\widehat{r}_{f^*}(\mu)\right])\right|}\right]e^{-\lambda t} \\
&\leq \mathbb{E}_{\mathbb{X}}\left[e^{\lambda\mathbb{E}_{\mathbb{X}'}\left[\sup_{f\in\mathcal{F}}\left|(\widehat{r}_f(\mu)-\widehat{r}_f'(\mu))-(\widehat{r}_{f^*}(\mu)-\widehat{r}_{f^*}'(\mu))\right|\right]}\right]e^{-\lambda t} \\
&\leq \mathbb{E}_{\mathbb{X},\mathbb{X}'}\left[e^{\lambda\sup_{f\in\mathcal{F}}\left|(\widehat{r}_f(\mu)-\widehat{r}_f'(\mu))-(\widehat{r}_{f^*}(\mu)-\widehat{r}_{f^*}'(\mu))\right|}\right]e^{-\lambda t} \\
&= \mathbb{E}_{\mathbb{X},\mathbb{X}'}\left[\mathbb{E}_{\varepsilon}\left[e^{\lambda\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i[Z_i(f)-Z_i(f^*)]\right|}\right]\right]e^{-\lambda t}
\end{aligned}$$

Using (15) in the Appendix with distance $\frac{d_{\mathbb{X},\mathbb{X}'}}{\sqrt{n}}$ and (10), we get

$$\begin{aligned}
\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|X_f - X_{f^*}| \geq t\right\} &\leq \mathbb{E}_{\mathbb{X},\mathbb{X}'}\left[e^{\lambda^2 2L^2\gamma_2(T,\frac{d_{\mathbb{X},\mathbb{X}'}}{\sqrt{n}})^2}\right]e^{-\lambda t} \\
&\leq \mathbb{E}_{\mathbb{X},\mathbb{X}'}\left[e^{\frac{\lambda^2 2L^2}{n}\gamma_2(T,d_{\mathbb{X},\mathbb{X}'})^2}\right]e^{-\lambda t} \ .
\end{aligned}$$

A few more calculations are able to reduce the random entropy on the couple $(\mathbb{X}, \mathbb{X}')$ to

the random entropy only on $\mathbb{X}$. Since $x \mapsto \frac{1}{\alpha}\phi(\alpha x)$ is Lipschitz with constant 1,

$$\begin{aligned} & d_{\mathbb{X},\mathbb{X}'}(f,g) \\ = & \left(\frac{1}{n\alpha}\sum_{i=1}^{n}\left(\phi(\alpha(f(X_i)-\mu)) - \phi(\alpha(f(X_i')-\mu)) - \phi(\alpha(g(X_i)-\mu)) + \phi(\alpha(g(X_i')-\mu))\right)^2\right)^{1/2} \\ \leq & \sqrt{2}\left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-g(X_i))^2\right)^{1/2} + \sqrt{2}\left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i')-g(X_i'))^2\right)^{1/2} \end{aligned}$$

This implies

$$\gamma_2(T, d_{\mathbb{X},\mathbb{X}'}) \leq \sqrt{2}(\gamma_2(T, d_{\mathbb{X}}) + \gamma_2(T, d_{\mathbb{X}'}))$$

and therefore

$$\mathbb{E}_{\mathbb{X},\mathbb{X}'}\left[e^{\frac{\lambda^2 2L^2}{n}\gamma_2(T,d_{\mathbb{X},\mathbb{X}'})^2}\right] \leq \mathbb{E}_{\mathbb{X}}\left[e^{\frac{\lambda^2 16L^2}{n}\gamma_2(T,d_{\mathbb{X}})^2}\right] .$$

Hence,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|X_f - X_{f^*}| \geq t\right\} \leq \mathbb{E}_{\mathbb{X},\mathbb{X}'}\left[e^{\frac{\lambda^2 16L^2}{n}\gamma_2(T,d_{\mathbb{X}})^2}\right]e^{-\lambda t} \tag{11}$$

Recall that, by definition, Γ_δ is such that $\mathbb{P}\{\gamma_2(T, d_{\mathbb{X}}) > \Gamma_\delta\} \leq \frac{\delta}{2}$. Thus,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|X_f - X_{f^*}| \geq t\right\} \leq \frac{\delta}{2} + e^{\frac{\lambda^2 16L^2}{n}\Gamma_\delta^2}e^{-\lambda t}$$

Optimization in λ with $t = 4\sqrt{2}L\Gamma_\delta\sqrt{\frac{\log(\frac{2}{\delta})}{n}}$ gives

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|X_f - X_{f^*}| \geq t\right\} \leq \delta$$

as desired. □

4 Applications

In this section we describe two applications of Theorems 1 and 2 to simple statistical learning problems. The first is a regression estimation problem in which we distinguish between L_1 and L_2 risks and the second is k-means clustering.

4.1 Empirical risk minimization for regression

4.1.1 L_1 regression

Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ be independent identically taking values in $\mathcal{Z} \times \mathbb{R}$ where $\mathcal{Z}$ a bounded subset of (say) $\mathbb{R}^d$. Suppose $\mathcal{G}$ is a class of functions $\mathcal{Z} \to \mathbb{R}$ bounded in the L_∞ norm, that is, $\Delta \stackrel{\text{def}}{=} \sup_{g,g' \in \mathcal{G}} \sup_{z \in \mathcal{Z}} |g(z) - g'(z)| < \infty$. First we consider the setup when the *risk* of each $g \in \mathcal{G}$ is defined by the L_1 loss

$$R(g) = \mathbb{E}|g(Z) - Y|$$

where the pair (Z, Y) has the same distribution of the (Z_i, Y_i) and is independent of them. Let $g^* = \operatorname{argmin}_{g \in \mathcal{G}} R(g)$ be a minimizer of the risk (which, without loss of generality, is assumed to exist). The statistical learning problem we consider here consists of choosing a function $\widehat{g}$ from the class $\mathcal{G}$ that has a risk $R(\widehat{g})$ not much larger than $R(g^*)$.

The standard procedure is to pick $\widehat{g}$ by minimizing the empirical risk $(1/n) \sum_{i=1} |g(Z_i) - Y_i|$ over $g \in \mathcal{G}$. However, if the response variable Y is unbounded and may have a heavy tail, ordinary empirical risk minimization is may fail to provide a good predictor of Y as the empirical risk is an unreliable estimate of the true risk.

Here we propose choosing $\widehat{g}$ by minimizing Catoni's estimate. To this end, we only need to assume that the second moment of Y is bounded by a known constant. More precisely, assume that $\mathbb{E}Y^2 \le \sigma^2$ for some $\sigma > 0$. Then $\sup_{g \in \mathcal{G}} \operatorname{Var}(|g(Z) - Y|) \le \sigma^2 + \sup_{g \in \mathcal{G}} \sup_{z \in \mathcal{Z}} |g(z)|^2 \stackrel{\text{def}}{=} v$ is a known and finite constant.

Now for all $g \in \mathcal{G}$ and $\mu \in \mathbb{R}$, define

$$\widehat{r}_g(\mu) = \frac{1}{n\alpha} \sum_{i=1}^{n} \phi(\alpha(|g(X_i) - Y_i| - \mu))$$

where ϕ is the truncation function defined in (3). Define $\widehat{R}(g)$ as the unique value for which $\widehat{r}_f(\widehat{R}(g)) = 0$. The empirical risk minimizer based on Catoni's risk estimate is then

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \widehat{R}(g) \ .$$

By Theorem 1, the performance of $\widehat{g}$ may be bounded in terms of covering numbers of the class of functions $\mathcal{F} = \{f(z, y) = |g(z) - y| : g \in \mathcal{G}\}$ based on the distance

$$D(f, f') = \sup_{z \in \mathcal{Z}, y \in \mathbb{R}} \big| |g(z) - y| - |g'(z) - y| \big| \le \sup_{z \in \mathcal{Z}} \big| g(z) - g'(z) \big| \ .$$

Thus, the covering numbers of $\mathcal{F}$ under the distance D may be bounded in terms of the covering numbers of $\mathcal{G}$ under the L_∞ distance. We obtain the following.

Corollary 9. *Consider the setup described above. Let $\alpha > 0$, $\delta \in (0,1)$ and $A_{f^*}(\delta) = \alpha v + \frac{2\log\delta^{-1}}{n\alpha}$. There exists a universal constant C such that, with probability at least $1-3\delta$,*

$$R(\widehat{g}) - R(g^*) \le 6A_{f^*}(\delta) + C\left(\log\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}}\int_0^{\Delta}\sqrt{\log N_\infty(\mathcal{G},\epsilon)}d\epsilon + O\left(\frac{1}{n}\right)\right) .$$

Note that the bound essentially has the same form as (1) but to apply (1) it is crucial that the response variable Y is bounded or at least has sub-Gaussian tails. We get this under the weak assumption that Y has a bounded second moment (with a known upper bound). The price we pay is that covering numbers under the distance $d_{\mathbb{X}}$ are now replaced by covering numbers under the supremum norm.

4.1.2 L_2 regression

Here we consider the same setup as in Section 4.1.1 but now the risk is measured by the L_2 loss. The *risk* of each $g \in \mathcal{G}$ is defined by the L_2 loss

$$R(g) = \mathbb{E}(g(Z) - Y)^2 .$$

Note that Theorem 1 is useless here as the difference $|R(g) - R(g')|$ is not bounded by the L_∞ distance of g and g' anymore and the covering numbers of $\mathcal{F}$ under the metric D are infinite. However, Theorem 2 gives meaningful bounds. Let $g^* = \operatorname{argmin}_{g\in\mathcal{G}} R(g)$ and again we choose $\widehat{g}$ by minimizing Catoni's estimate.

Here we need to assume that $\mathbb{E}Y^4 \le \sigma^2$ for some $\sigma > 0$. Then $\sup_{g\in\mathcal{G}} \operatorname{Var}\left((g(Z)-Y)^2\right) \le \sigma^2 + \sup_{g\in\mathcal{G}}\sup_{z\in\mathcal{Z}}|g(z)|^4 \stackrel{\text{def}}{=} v$ is a known and finite constant.

By Theorem 2, the performance of $\widehat{g}$ may be bounded in terms of covering numbers of the class of functions $\mathcal{F} = \{f(z,y) = (g(z)-y)^2 : g \in \mathcal{G}\}$ based on the distance

$$d_{\mathbb{X}}(f,f') = \left(\frac{1}{n}\sum_{i=1}^n\left((g(Z_i)-Y_i)^2 - (g'(Z_i)-Y_i)^2\right)^2\right)^{1/2}$$

Note that

$$\begin{aligned}\left|(g(Z_i)-Y_i)^2 - (g'(Z_i)-Y_i)^2\right| &= |g(Z_i)-g'(Z_i)||2Y_i - g(Z_i) - g'(Z_i)| \\ &\le 2|g(Z_i)-g'(Z_i)|(|Y_i|+\Delta) \\ &\le 2d_\infty(g,g')(|Y_i|+\Delta) ,\end{aligned}$$

and therefore

$$\begin{aligned}d_{\mathbb{X}}(f,f') &\le 2d_\infty(g,g')\sqrt{\frac{1}{n}\sum_{i=1}^n(|Y_i|+\Delta)^2} \\ &\le 2\sqrt{2}d_\infty(g,g')\sqrt{\Delta^2 + \frac{1}{n}\sum_{i=1}^n Y_i^2} .\end{aligned}$$

By Chebyshev's inequality,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \mathbb{E}\left[Y^2\right] > t\right\} \leq \frac{\mathrm{Var}\left(Y^2\right)}{nt^2} \leq \frac{\sigma^2}{nt^2}$$

thus $\frac{1}{n}\sum_{i=1}^{n} Y_i^2 > \mathbb{E}\left[Y^2\right] + \sqrt{\frac{2\sigma^2}{n\delta}}$ with probability at most $\frac{\delta}{2}$ and

$$d_{\mathbb{X}}(f, f') > 2\sqrt{2} d_\infty(g, g') \sqrt{\Delta^2 + \mathbb{E}\left[Y^2\right] + \sqrt{\frac{2\sigma^2}{n\delta}}}$$

occurs with a probability bounded by $\frac{\delta}{2}$. Then Theorem 2 applies with

$$\Gamma_\delta = 2\sqrt{2}\sqrt{\Delta^2 + \mathbb{E}\left[Y^2\right] + \sqrt{\frac{2\sigma^2}{n\delta}}}\gamma_2(\mathcal{G}, d_\infty) \ .$$

Corollary 10. *Consider the setup described above. Let $\alpha > 0$, $\delta \in (0,1)$ and $A_{f^*}(\delta) = \alpha v + \frac{2\log\delta^{-1}}{n\alpha}$. There exists a universal constant C such that, with probability at least $1-3\delta$,*

$$R(\widehat{g}) - R(g^*) \leq 6A_{f^*}(\delta) + C\sqrt{\log\left(\frac{2}{\delta}\right)}\sqrt{\frac{\Delta^2 + \mathbb{E}\left[Y^2\right] + 2\sigma^2/(n\delta)}{n}} \int_0^\Delta \sqrt{\log N_\infty(\mathcal{G}, \epsilon)} d\epsilon \ .$$

4.2 k-means clustering under heavy tailed distribution

In *k-means clustering*—or *vector quantization*—one wishes to represent a distribution by a finite number of points. Formally, let X be a random vector taking values in $\mathbb{R}^d$ and let P denote the distribution of X. Let $k \geq 2$ be a positive integer that we fix for the rest of the section. A clustering scheme is given by a set of k cluster centers $C = \{y_1, \ldots, y_k\} \subset \mathbb{R}^d$ and a *quantizer* $q : \mathbb{R}^d \to C$. Given a *distortion measure* $\ell : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, one wishes to find C and q such that the expected distortion

$$D_k(P, q) = \mathbb{E}\ell(X, q(X))$$

is as small as possible. The minimization problem is meaningful whenever $\mathbb{E}\ell(X, 0) < \infty$ which we assume throughout. Typical distortion measures are of the form $\ell(x, y) = \|x - y\|^\alpha$ where $\|\cdot\|$ is a norm on $\mathbb{R}^d$ and $\alpha > 0$ (typically α equals 1 or 2). Here, for concreteness and simplicity, we assume that ℓ is the Euclidean distance $\ell(x, y) = \|x - y\|$ though the results may be generalized in a straightforward manner to other norms. In a way equivalent to the arguments of Section 4.1.2, the results may be generalized to the case of the quadratic distortion $\ell(x, y) = \|x - y\|^2$. In order to avoid repetition of arguments, the details are omitted.

It is not difficult to see that if $\mathbb{E}\|X\| < \infty$, then there exists a (not necessarily unique) quantizer q^* that is optimal, that is, q^* is such that for all clustering schemes q,

$$D_k(P, q) \geq D_k(P, q^*) \stackrel{\text{def}}{=} D_k^*(P) \ .$$

It is also clear that q^* is a *nearest neighbor quantizer*, that is,

$$\|x - q^*(x)\| = \min_{y_i \in C} \|x - y_i\| \ .$$

Thus, nearest neighbor quantizers are determined by their cluster centers $C = \{y_1, \ldots, y_k\}$. In fact, for all quantizers with a particular set C of cluster centers, the corresponding nearest neighbor quantizer has minimal distortion and therefore it suffices to restrict our attention to nearest neighbor quantizers.

In the problem of empirical quantizer design, one is given an i.i.d. sample $X_1, \ldots, X_n$ drawn from the distribution P and one's aim is to find a quantizer q_n whose distortion

$$D_k(P, q_n) = \mathbb{E}\left[\|X - q_n(X)\| | X_1, \ldots, X_n\right]$$

is as close to $D_k^*(P)$ as possible. A natural strategy is to choose a quantizer–or equivalently, a set C of cluster centers–by minimizing the *empirical distortion*

$$D_k(P_n, q) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - q(X_i)\| = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|X_i - y_j\| \ ,$$

where P_n denotes the standard empirical distribution based on $X_1, \ldots, X_n$. If $\mathbb{E}\|X\| < \infty$, then the empirically optimal quantizer asymptotically minimizes the distortion. More precisely, if q_n denotes the empirically optimal quantizer (i.e., $q_n = \operatorname{argmin}_q D_k(P_n, q)$), then

$$\lim_{n \to \infty} D_k(P, q_n) = D_k^*(P) \quad \text{with probability 1,}$$

see Pollard [24, 26] and Abaya and Wise [1] (see also Linder [17]). The rate of convergence of $D_k(P, q_n)$ to $D_k^*(P)$ has drawn considerable attention, see, e.g., Pollard [25], Bartlett, Linder, and Lugosi [5], Antos [3], Antos, Györfi, and György [4], Biau, Devroye, and Lugosi [7], Maurer and Pontil [20], and Levrard [16]. Such rates are typically studied under the assumption that X is almost surely bounded. Under such assumptions one can show that

$$\mathbb{E}D_k(P, q_n) - D_k^*(P) \leq C(P, k, d) n^{-1/2}$$

where the constant $C(P, k, d)$ depends on $\operatorname{esssup} \|X\|$, k, and the dimension d. (The value of the constant has mostly be investigated in the case of quadratic loss $\ell(x, y) = \|x - y\|^2$ but most proofs may be modified for the case studied here.

However, little is known about the finite-sample performance of empirically designed quantizers under possibly heavy-tailed distributions. In fact, there is no hope to extend the

results cited above for distributions with finite second moment simply because empirical averages are poor estimators of means under such general conditions.

In the recent paper of Telgarsky and Dasgupta [28], bounds on the excess risk under conditions on higher moments have been developed. They prove a bound of $\mathcal{O}(n^{-1/2+2/p})$ for the excess distortion where p is the number of moments of $\|X\|$ that are assumed to be finite. Here we show that there exists an empirical quantizer $\hat{q}_n$ whose excess distortion $D_k(P, \widehat{q}_n) - D_k^*(P)$ is of the order of $n^{-1/2}$ (with high probability) under the only assumption that $\mathbb{E}\left[\|X\|^2\right]$ is finite. This may be achieved by choosing a quantizer that minimizes Catoni's estimate of the distortion.

The proposed empirical quantizer uses two parameters that depend on the (unknown) distribution of X. For simplicity, we assume that upper bounds for these two parameters are available. (Otherwise either one may try to estimate them or, as the sample size grows, use increasing values for these parameters. The details go beyond the scope of this paper.)

One of these parameters is the second moment $\mathrm{Var}\,(\|X\|)$ and let V be an upper bound. The other parameter $\rho > 0$ is an upper bound for the norm of the possible cluster centers. The next lemma offers an estimate.

Lemma 11. *(Linder [17].) Let $2 \le m \le k$ be the unique integer such that $D_k^* = \cdots = D_m^* < D_{m-1}^*$ and define $\varepsilon = (D_{k-1}^* - D_k^*)/2$. Let $(y_1, \ldots, y_m)$ be a set of cluster centers such that the distortion of the corresponding quantizer is less than $D_m^* + \varepsilon$. Let $B_r = \{x : \|x\| \le r\}$ denote the closed ball of radius $r > 0$ centered at the origin. If $\rho > 0$ is such that*

- $\frac{\rho}{10} P(B_{\frac{\rho}{10}}) > 2\mathbb{E}\,\|X\|$
- $P(B_{2\rho/5}) > 1 - \frac{\varepsilon^2}{4\mathbb{E}\left[\|X\|^2\right]}$

then for all $1 \le j \le k$, $\|y_j\| \le \rho$.

Now we are prepared to describe the proposed empirical quantizer. Let $\mathcal{C}_\rho$ be the set of all collections $C = \{y_1, \ldots, y_k\} \in (\mathbb{R}^d)^k$ of cluster centers with $\|y_j\| \le \rho$ for all $j = 1, \ldots, k$. For each $C \in \mathcal{C}_\rho$, denote by q_C the corresponding quantizer. Now for all $C \in \mathcal{C}_\rho$, we may calculate Catoni's mean estimator of the distortion $D(P, q_C) = \mathbb{E}\|X - q_C(X)\| = \mathbb{E}\min_{j=1,\ldots,k} \|X_i - y_j\|$ defined as the unique value $\mu \in \mathbb{R}$ for which

$$\frac{1}{n\alpha} \sum_{i=1}^{n} \phi\left(\alpha\left(\min_{j=1,\ldots,k} \|X_i - y_j\| - \mu\right)\right) = 0$$

where we use the parameter value $\alpha = \sqrt{2/nkV}$. Denote this estimator by $\widehat{D}(P_n, q_C)$ and let $\widehat{q}_n$ be any quantizer minimizing the estimated distortion. An easy compactness argument shows that such a minimizer exists.

The main result of this section is the following bound for the distortion of the chosen quantizer.

Theorem 12. *Assume that* $Var(\|X\|) \leq V < \infty$ *and* $n \geq d$. *Then, with probability at least* $1 - \delta$,

$$D(P, \widehat{q}_n) - D(P, q^*) \leq C \left(\log \frac{1}{\delta} \right) \left(\sqrt{\frac{Vk}{n}} + \sqrt{\frac{dk}{n}} \right) + O\left(\frac{1}{n} \right) \ ,$$

where the constant C *only depends on* ρ.

Proof. The result follows from Theorem 1. All we need to check is that $\mathrm{Var}\left(\min_{j=1,\ldots,k} \|X - y_j\|\right)$ is bounded by $2kV$ and estimate the covering numbers of the class of functions

$$\mathcal{F}_\rho = \left\{ f_C(x) = \min_{y \in C} \|x - y\| : C \in \mathcal{C}_\rho \right\} \ .$$

The variance bound follows simply by the fact that for all $C \in \mathcal{C}$,

$$\mathrm{Var}\left(\min_{j=1,\ldots,k} \|X - y_j\| \right) \leq \sum_{i=1}^{k} \mathrm{Var}\left(\|X - y_i\| \right) \leq \sum_{i=1}^{k} 2\mathrm{Var}\left(\|X\| \right) \leq 2kV \ .$$

In order to use the bound of Theorem 1, we need to bound the covering numbers of the class $\mathcal{F}_\rho$ under both metrics d and D. We begin with the metric

$$D(f_C, f_{C'}) = \sup_{x \in \mathbb{R}^d} |f_C(x) - f_{C'}(x)| \ .$$

$B_z(\epsilon, d)$ refers to the ball under the metric d of radius ϵ centered at z. Let Z be a subset of B_ρ such that

$$\mathcal{B}_{B_\rho} := \{ B_z(\epsilon, d_2) : z \in Z \}$$

is a covering of the set B_ρ by balls of radius ϵ under the Euclidean norm. Let $C \in \mathcal{C}_\rho$ and associate to any $y_i \in C$ one of the centers in Z such that $\|y_i - z_i\| \leq \epsilon$. If there is more than one possible choice for z_i, we pick one of them arbitrarily. We denote by $q_{C'}$ the nearest neighbor quantizer with codebook $C' = (z_i)_i$. Finally, let $S_i = q_{C'}^{-1}(z_i)$. Now clearly, $\forall i, \ \forall x \in S_i$

$$\begin{aligned} f_C(x) - f_{C'}(x) &= \min_{1 \leq j \leq k} \|x - y_j\| - \min_{1 \leq j \leq k} \|x - z_j\| \\ &= \min_{1 \leq j \leq k} \|x - y_j\| - \|x - z_i\| \\ &\leq \|x - y_i\| - \|x - z_i\| \leq \epsilon \end{aligned}$$

and symmetrically for $f_{C'}(x) - f_C(x)$. Then $f_C \in B_{f_{C'}}(\epsilon, D)$ and

$$\mathcal{B}_{\mathcal{F}_\rho} := \{ B_{f_C}(\epsilon, D) : C \in Z^k \}$$

is a covering of $\mathcal{F}_\rho$. Since Z can be taken such that $|Z| = N_{d_2}(B_\rho, \epsilon)$ we end with

$$N_d(\mathcal{F}_\rho, \epsilon) \leq N_D(\mathcal{F}_\rho, \epsilon) \leq N_{d_2}(B_\rho, \epsilon)^k \ .$$

By standard estimates on the covering numbers of the ball B_ρ by balls of size ϵ under the Euclidean metric,

$$N_{d_2}(B_\rho, \epsilon) \leq \left(\frac{4\rho}{\epsilon}\right)^d$$

(see, e.g., Matousek [19]). In other words, there exists a constant C_ρ that depends only on ρ such that

$$\gamma_2(\mathcal{F}_\rho, d) \leq \int_0^{2\rho} \sqrt{\log N_d(\mathcal{F}_\rho, \epsilon)} d\epsilon \leq C_\rho \sqrt{kd}$$
$$\text{and} \quad \gamma_1(\mathcal{F}_\rho, D) \leq \int_0^{2\rho} \log N_D(\mathcal{F}_\rho, \epsilon) d\epsilon \leq C'_\rho kd$$

Theorem 1 may now be applied to the class $\mathcal{F}_\rho$. □

5 Simulation Study

In this closing section we present the results of two simulation exercises to assess the performance of the estimators developed in this work.

5.1 L_2 Regression

The first application is an L_2 regression exercise. Data are simulated from a linear model with heavy-tailed errors and the L_2 regression procedure based on Catoni's risk minimizer introduced in Section 4.1.2 is used for estimation. The procedure is benchmarked against regular ("vanilla") L_2 regression based on the minimization of the empirical L_2 loss.

The simulation exercise is designed as follows. We simulate $(Z_1, Y_1), (Z_2, Y_2), ..., (Z_n, Y_n)$ i.i.d. pairs of random variables in $\mathbb{R}^4 \times \mathbb{R}$. Each component of the Z_i vector is drawn from a uniform distribution with support $[-1, 1]$ while Y_i is generated as

$$Y_i = Z_i^T \theta + \epsilon_i \ ,$$

where the parameter vector θ is $(0.25, -0.25, 0.50, 0.70)'$ and ϵ_i is drawn from a Student's t distribution with d degrees of freedom. As it is well known, the degrees of freedom parameter determines the highest finite moment of the Student's t distribution. Moments of order $k \geq d$ do not exist. We are interested in finding the value of θ which minimizes the L_2 loss

$$\mathbb{E}\left|Y - Z_i^T \theta\right|^2 \ .$$

The parameter θ is estimated using the Catoni and the vanilla L_2 regressions. Let $\widehat{R}_C(\theta)$ denote the solution of the equation

$$\hat{r}_\theta(\mu) = \frac{1}{n\alpha}\sum_{i=1}^{n}\psi\left(\alpha\left(\left|Y_i - Z_i^T\theta\right|^2 - \mu\right)\right) = 0\,,$$

then the Catoni L_2 regression estimator is defined as

$$\hat{\theta}_C = \arg\min_\theta \widehat{R}_C(\theta)\,.$$

The vanilla L_2 regression estimator is defined as the minimizer of the empirical L_2 loss,

$$\hat{\theta}_V = \arg\min_\theta \widehat{R}_V(\theta) = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Z_i^T\theta\right|^2\,,$$

which is the classic least squares estimator. The estimated risk of the Catoni and vanilla estimators are denoted as $\widehat{R}_C(\hat{\theta}_C)$ and $\widehat{R}_V(\hat{\theta}_V)$ respectively.

Expected risk is the natural index to assess the precision of the estimators

$$\begin{array}{rcl} R_C & = & \mathbb{E}|Y - Z^T\hat{\theta}_n^C|^2 \\ R_V & = & \mathbb{E}|Y - Z^T\hat{\theta}_n^V|^2 \end{array}\,.$$

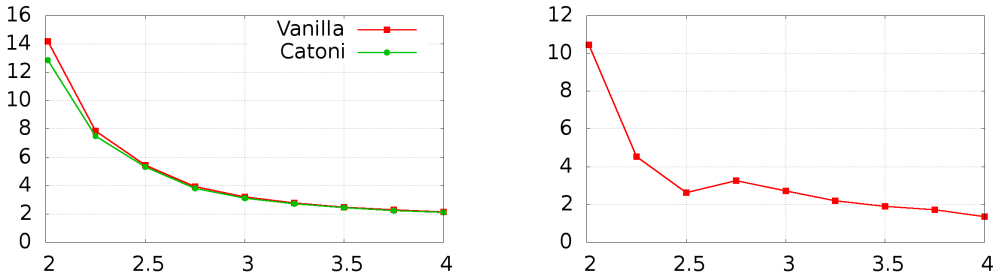
We estimate the expected risk by simulation. For each replication of the simulation exercise, we estimate the empirical risk of the estimators using an i.i.d. sample $(Z_1', Y_1'), ..., (Z_m', Y_m)$ that is independent of the one used for estimation,

$$\begin{array}{rcl} \widetilde{R}_C & = & \frac{1}{m}\sum_{i=1}^{m}|Y_i' - Z_i'^T\hat{\theta}_n^C|^2 \\ \widetilde{R}_V & = & \frac{1}{m}\sum_{i=1}^{m}|Y_i' - Z_i'^T\hat{\theta}_n^V|^2 \end{array}\,. \tag{12}$$

The simulation experiment is replicated for different values of the tail parameter d ranging from 2 to 4 and different values of the sample size n ranging from 25 to 200. For each combination of the degrees of freedom parameter d and sample size n the experiment is replicated 10‘000 times.

Figure 1 displays the Monte Carlo estimate of R_C and R_V as functions of the tail parameter d when the sample size n is equal to 50. The left panel reports the level of the indices while the right panel reports the percentage improvement of the Catoni procedure over the benchmark. When the tails are not excessively heavy (high values of d) the difference between the procedures is small. As the tails become heavier (small values of d) the risk of both procedures increases. Importantly, the Catoni estimator becomes progressively more efficient as the tails become heavier. The improvement is roughly 10% of the benchmark when the the tail parameter is close to 2. Detailed results for different values of n are reported in Table 1. The pattern documented in the pictures holds for different values of n but the advantages of the Catoni approach are stronger when the sample size n is smaller. Overall the Catoni L_2 regression estimator never performs significantly worse than the benchmark and it is substantially better when the tails of the data become heavier and data are scarce.

Figure 1: L_2 Regression Parameter Estimation.



(a) (b)

The figure plots the risk of the Catoni and vanilla L_2 regression parameter estimators (a) and the percentage improvement of the Catoni procedure relative to the vanilla (b) as a function of the tail parameter d for a sample size n equal to 50.

Table 1: Relative Performance of the Catoni L_2 Parameter Estimator.

d	**n=25**	**n=50**	**n=75**	**n=100**	**n=150**	**n=200**
2.01	15.50	10.50	4.40	3.70	3.40	1.90
2.25	9.80	4.50	3.30	4.00	1.70	1.10
2.50	8.20	2.60	2.40	2.50	1.00	1.00
2.75	7.20	3.30	2.10	1.80	1.10	0.80
3.00	5.40	2.70	2.30	1.40	0.80	0.70
3.25	4.90	2.20	1.60	1.20	0.80	0.60
3.50	3.60	1.90	1.30	1.00	0.70	0.50
3.75	2.90	1.70	1.20	0.80	0.60	0.40
4.00	2.90	1.40	1.00	0.70	0.50	0.30

The table reports the improvement of the Catoni L_2 parameter estimator relative to the vanilla procedure as a function of the tail parameter d and sample size n.

5.2 k-means

In the second experiment we carry out a k–means clustering exercise. Data are simulated from a heavy-tailed mixture distribution and then cluster centers are chosen by minimizing Catoni's estimate of the L_2 distortion. The performance of the algorithm is benchmarked against the vanilla k–means algorithm procedure where the distortion is estimated by simple empirical averages.

The simulation exercise is designed as follows. An i.i.d. sample of random vectors $X_1, ..., X_n$ in $\mathbb{R}^2$ is drawn from a four-component mixture distribution with equal weights. Each mixture component is a bivariate Student's t distribution with d degrees of freedom and independent coordinates. The k–means algorithm based on Catoni as well as the standard ("vanilla") k–means algorithm are used to estimate the cluster centers, which are denoted respectively as $\hat{q}_C$ and $\hat{q}_V$.

Analogously to the previous exercise, we summarize the performance of the clustering procedures using their expected distortion of the algorithms, that is

$$\begin{array}{rcl} R_C & = & D_k(P, \hat{q}_n^V) \\ R_V & = & D_k(P, \hat{q}_n^V) \end{array} .$$

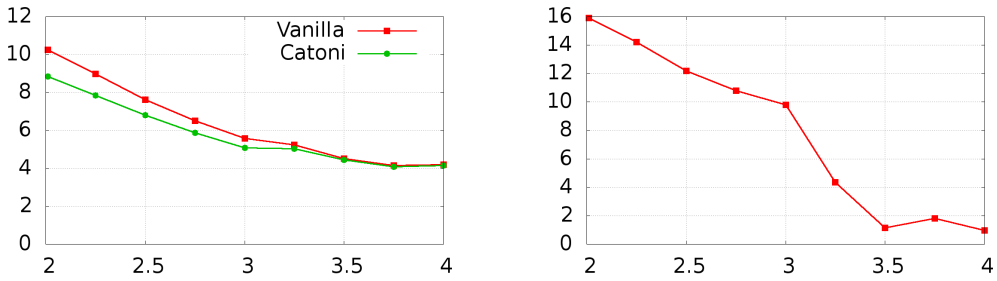
We estimate the expected distortion by simulation. We compute the empirical distortion of the quantizers using an i.i.d. sample $X'_1, ..., X'_m$ of vectors that is independent of the ones used for estimation, that is,

$$\begin{array}{rcl} \widetilde{R}_C = D_k(P'_m, \hat{q}_n^V) & = & \frac{1}{m} \sum_{i=1}^m \min_{j=1,...,k} \left|\left|X'_i - \hat{q}_n^V(X'_i)\right|\right|^2 \\ \widetilde{R}_V = D_k(P'_m, \hat{q}_n^C) & = & \frac{1}{m} \sum_{i=1}^m \min_{j=1,...,k} \left|\left|X'_i - \hat{q}_n^C(X'_i)\right|\right|^2 \end{array} . \tag{13}$$

The experiment is replicated for different values of the tail parameter d ranging from 2 to 4 and different values of the sample size n ranging from 25 to 200. For each combination of tail parameter d and sample size n the experiment is replicated 10'000 times.

Figures 2 displays the Monte Carlo estimate of R_C and R_V as a function of the degree of freedom d for $n = 50$. The left panel reports the absolute estimated risk while the right panel reports the percentage improvement of the Catoni procedure over the benchmark. The overall results are analogous to the ones of the L_2 regression application. When the tails of the mixture are not excessively heavy (high values of d) the difference in the procedures is small. As the tails become heavier (small values of d) the risk of both procedure increases, but the Catoni algorithm becomes progressively more efficient. The percentage gains for the Catoni procedure are above 15% of the benchmark when the tail parameter is close to 2. Tables 2 report detailed results for different values of n. Overall, the Catoni k–means algorithm never performs worse than the benchmark and it is substantially better when the tails of the mixture become heavier and the sample size is small.

Figure 2: k–means Quantizer Estimation.



(a) (b)

The figure plots the risk of the Catoni and vanilla k–means quantizer estimator (a) and the percentage improvement of the Catoni procedure relative to the vanilla (b) as a function of the tail parameter d for a sample size n equal to 100.

Table 2: Relative Performance of the Catoni k–means Quantizer Estimator.

d	**n=25**	**n=50**	**n=75**	**n=100**	**n=150**	**n=200**
2.01	21.30	15.90	11.30	9.50	8.20	8.40
2.25	15.90	14.20	10.40	9.80	8.70	7.80
2.50	13.20	12.20	8.80	8.50	6.90	6.30
2.75	11.50	10.80	5.50	5.00	6.00	6.20
3.00	10.40	9.80	6.40	5.80	4.90	3.50
3.25	9.20	4.40	1.30	2.50	0.70	3.30
3.50	7.50	1.10	0.90	0.80	1.10	0.60
3.75	8.20	1.80	1.60	0.90	0.40	0.50
4.00	4.70	1.00	1.00	0.80	0.60	0.40

The table reports the improvement of the Catoni k–means quantizer estimator relative to the vanilla procedure as a function of the tail parameter d and sample size n.

6 Appendix

6.1 A chaining theorem

The following result is a version of standard bounds based on "generic chaining", see Talagrand [27]. We include the proof for completeness.

Recall that if ψ is a non-negative increasing convex function defined on $\mathbb{R}_+$ with $\psi(0) = 0$, then the Orlicz norm of a random variable X is defined by

$$\|X\|_\psi = \inf\left\{c > 0 \ : \ \mathbb{E}\left[\psi\left(\frac{|X|}{c}\right)\right] \le 1\right\} \ .$$

We consider Orlicz norms defined by

$$\psi_1(x) = \exp(x) - 1 \quad \text{and} \quad \psi_2(x) = \exp(x^2) - 1 \ .$$

It is easy to see that $\|X\|_{\psi_1} \le \|X\|_{\psi_2}$ always holds. Also note that, by Markov's inequality, $\|X\|_{\psi_1} \le c$ implies that $\mathbb{P}\{|X| > t\} \le e^{-t/c}$ and similarly, if $\|X\|_{\psi_2} \le c$, then $\mathbb{P}\{|X| > t\} \le e^{-t^2/c^2}$. Then

$$\begin{aligned} X &\le \|X\|_{\psi_1} \log(\delta^{-1}) && \text{with probability at least } 1-\delta \ , \\ X &\le \|X\|_{\psi_2} \sqrt{\log(\delta^{-1})} && \text{with probability at least } 1-\delta \ . \end{aligned} \tag{14}$$

Recall the following definition (see, e.g., [27, Definition 1.2.3]). Let T be a (pseudo) metric space. An increasing sequence $(\mathcal{A}_n)$ of partitions of T is called *admissible* if for all $n = 0, 1, 2, \ldots$, $\#\mathcal{A}_n \le 2^{2^n}$. For any $t \in T$, denote by $A_n(t)$ the unique element of $\mathcal{A}_n$ that contains t. Let $\Delta(A)$ denote the diameter of the set $A \subset T$. Define, for $\alpha = 1, 2$,

$$\gamma_\alpha(T, d) = \inf_{\mathcal{A}_n} \sup_{t \in T} \sum_{n \ge 0} 2^{n/\alpha} \Delta(A_n(t)) \ ,$$

where the infimum is taken over all admissible sequences.

Theorem 13. *Let $(X_t)_{t \in T}$ be a stochastic process indexed by a set T on which two (pseudo) metrics, d_1 and d_2, are defined such that T is bounded with respect to both metrics. Assume that for any $s, t \in T$ and for all $x > 0$,*

$$\mathbb{P}\{|X_s - X_t| > x\} \le 2 \exp\left(-\frac{1}{2} \frac{x^2}{d_2^2 + d_1 x}\right) \ .$$

Then for all $t \in T$,

$$\left\| \sup_{s \in T} |X_s - X_t| \right\|_{\psi_1} \le L \left(\gamma_1(T, d_1) + \gamma_2(T, d_2)\right)$$

with $L \le 384 \log(2)$.

Corollary 14. *Assume that for any $s, t \in T$ and for all $x > 0$,*

$$\mathbb{P}\{|X_s - X_t| > x\} \leq 2 \exp\left(-\frac{x^2}{2d_2^2}\right) .$$

Then for all $t \in T$,

$$\left\| \sup_{s \in T} |X_s - X_t| \right\|_{\psi_2} \leq L\gamma_2(T, d_2)$$

with $L \leq 384 \log(2)$.

In particular, we obtain

$$\mathbb{E}\left[e^{\lambda \sup_{s \in T} |X_s - X_t|}\right] \leq e^{\lambda^2 2L^2 \gamma_2(T,d_2)^2} . \tag{15}$$

The proof of Theorem 13 uses the following lemma:

Lemma 15. ([30, LEMMA 2.2.10].) *Let $a, b > 0$ and assume that the random variables $X_1, \ldots, X_m$ satisfy, for all $x > 0$,*

$$\mathbb{P}\{|X_i| > x\} \leq 2 \exp\left(-\frac{1}{2} \frac{x^2}{b + ax}\right) .$$

Then

$$\left\| \max_{1 \leq i \leq m} X_i \right\|_{\psi_1} \leq 48 \left(a \log(1 + m) + \sqrt{b} \sqrt{\log(1 + m)}\right) .$$

Proof of Theorem 13: Consider an admissible sequence $(\mathcal{B}_n)_{n \geq 0}$ such that for all $t \in T$,

$$\sum_{n \geq 0} 2^n \Delta_1(B_n(t)) \leq 2\gamma_1(T, d_1)$$

and an admissible sequence $(\mathcal{C}_n)_{n \geq 0}$ such that for all $t \in T$,

$$\sum_{n \geq 0} 2^{n/2} \Delta_1(C_n(t)) \leq 2\gamma_2(T, d_2)$$

Now we may define an admissible sequence by intersection of the elements of $(\mathcal{B}_{n-1})_{n \geq 1}$ and $(\mathcal{C}_{n-1})_{n \geq 1}$: set $\mathcal{A}_0 = \{T\}$ and let

$$\mathcal{A}_n = \{B \cap C \ : \ B \in \mathcal{B}_{n-1} \ \& \ C \in \mathcal{C}_{n-1}\}$$

$(\mathcal{A}_n)_{n \geq 0}$ is an admissible sequence because each $\mathcal{A}_n$ is increasing and contains at most $(2^{2^{n-1}})^2 = 2^{2^n}$ sets. Define a sequence of finite sets $T_0 = \{t\} \subset T_1 \subset \cdots \subset T$ such that

T_n contains a single point in each set of $\mathcal{A}_n$. For any $s \in T$, denote by $\pi_n(s)$ the unique element of T_n in $A_n(s)$. Now for any $s \in T_{k+1}$, we write

$$X_s - X_t = \sum_{k=0}^{\infty} \left(X_{\pi_{k+1}(s)} - X_{\pi_k(s)} \right) \ .$$

Then, using the fact that $\|\cdot\|_{\psi_1}$ is a norm and Lemma 15,

$$\begin{aligned}
&\left\| \sup_{s \in T} |X_s - X_t| \right\|_{\psi_1} \\
&\quad \leq \quad \sum_{k=0}^{\infty} \left\| \max_{s \in T_{k+1}} |X_{\pi_{k+1}(s)} - X_{\pi_k(s)}| \right\|_{\psi_1} \\
&\quad \leq \quad 48 \sum_{k=0}^{\infty} \left(d_1(\pi_{k+1}(s), \pi_k(s)) \log(1 + 2^{2^{k+1}}) + d_2(\pi_{k+1}(s), \pi_k(s)) \sqrt{\log(1 + 2^{2^{k+1}})} \right) \ .
\end{aligned}$$

Since $(\mathcal{A}_n)_{n \geq 0}$ is an increasing sequence, $\pi_{k+1}(s)$ and $\pi_k(s)$ are both in $A_k(s)$. By construction, $A_k(s) \subset B_k(s)$, and therefore $d_1(\pi_{k+1}(s), \pi_k(s)) \leq \Delta_1(B_k(s))$. Similarly, $d_2(\pi_{k+1}(s), \pi_k(s)) \leq \Delta_2(C_k(s))$. Using $\log(1 + 2^{2^{k+1}}) \leq 4 \log(2) 2^k$, we get

$$\begin{aligned}
\left\| \max_{s \in T} |X_s - X_t| \right\|_{\psi_1} \quad &\leq \quad 192 \log(2) \left[\sum_{k=0}^{\infty} 2^k \Delta_1(B_k(s)) + \sum_{k=0}^{\infty} 2^{k/2} \Delta_1(C_k(s)) \right] \\
&\leq \quad 384 \log(2) \left[\gamma_1(T, d_1) + \gamma_2(T, d_2) \right] \ .
\end{aligned}$$

References

[1] E. A. Abaya and G. L. Wise. Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44:183–189, 1984.

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58:137–147, 2002.

[3] A. Antos. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory*, 51:4022–4032, 2005.

[4] A. Antos, L. Györfi, and A. György. Improved convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, 51:4013–4022, 2005.

[5] P. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44:1802–1813, Sep. 1998.

[6] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory Related Fields*, 135:311–334, 2006.

[7] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 200.

[8] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM. Probability and Statistics*, 9:323–375, 2005.

[9] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[10] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Badits with heavy tail. *manuscript*, 2013.

[11] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Arxiv preprint arXiv:1009.2048*, 2010.

[12] R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

[13] D. Hsu and S. Sabato. Approximate loss minimization with heavy tails. *Computing Research Repository*, abs/1307.1827, 2013.

[14] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 36:00–00, 2006.

[15] M. Lerasle and R.I. Oliveira. Robust empirical mean estimators. *manuscript*, 2012.

[16] C. Levrard. Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, pages 1716–1746, 2013.

[17] T. Linder. Learning-theoretic methods in vector quantization. In L. Györfi, editor, *Principles of nonparametric learning*, number 434 in CISM Courses and Lecture Notes. Springer-Verlag, New York, 2002.

[18] P. Massart. *Concentration inequalities and model selection*. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer, 2006.

[19] J. Matoušek. *Lectures on discrete geometry*. Springer, 2002.

[20] A. Maurer and M. Pontil. k-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56:5839–5846, 2010.

[21] S. Mendelson. Learning without concentration. *arXiv preprint arXiv:1401.0304*, 2014.

[22] S. Minsker. Geometric median and robust estimation in banach spaces. *arXiv preprint*, 2013.

[23] A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[24] D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9, no. 1:135–140, 1981.

[25] D. Pollard. A central limit theorem for k-means clustering. *Annals of Probability*, 10(4):919–926, 1982.

[26] D. Pollard. Quantization and the method of k-means. *IEEE Trans. Inform. Theory*, IT-28:199–205, 1982.

[27] M. Talagrand. *The generic chaining*. Springer, 2005.

[28] M. Telgarsky and S. Dasgupta. Moment-based uniform deviation bounds for k-means and friends. *arXiv preprint arXiv:1311.1903*, 2013.

[29] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK, 2000.

[30] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, New York, 1996.