Supplement to "Robust linear least squares regression"

Jean-Yves Audibert, Olivier Catoni

SUPPLEMENT TO "ROBUST LINEAR LEAST SQUARES REGRESSION"

By Jean-Yves Audibert[*,†], and Olivier Catoni[‡,§]

This supplementary material provides the proofs of Theorems 2.1, 2.2 and 3.1 of the article "Robust linear least squares regression".

CONTENTS

1. Proofs of Theorems 2.1 and 2.2. To shorten the formulae, we will write X for $\varphi(X)$, which is equivalent to considering without loss of generality that the input space is $\mathbb{R}^d$ and that the functions $\varphi_1,\dots,\varphi_d$ are the coordinate functions. Therefore, the function f_θ maps an input x to $\langle\theta,x\rangle$. With a slight abuse of notation, $R(\theta)$ will denote the risk of this prediction function.

Let us first assume that the matrix $Q_\lambda = Q+\lambda I$ is positive definite. This indeed does not restrict the generality of our study, even in the case when $\lambda = 0$, as we will discuss later (Remark 1.1).

Consider the change of coordinates

$$\overline{X} = Q_\lambda^{-1/2} X.$$

Let us introduce

$$\overline{R}(\theta) = \mathbb{E}\big[(\langle\theta,\overline{X}\rangle - Y)^2\big],$$

[*]Université Paris-Est, LIGM, Imagine, 6 avenue Blaise Pascal, 77455 Marne-la-Vallée, France, E-mail: audibert@imagine.enpc.fr

[†]CNRS/École Normale Supérieure/INRIA, LIENS, Sierra – UMR 8548, 23 avenue d'Italie, 75214 Paris cedex 13, France.

[‡]École Normale Supérieure, CNRS – UMR 8553, Département de Mathématiques et Applications, 45 rue d'Ulm, 75230 Paris cedex 05, France, E-mail: olivier.catoni@ens.fr

[§] INRIA Paris-Rocquencourt - CLASSIC team.

AMS 2000 subject classifications: 62J05, 62J07

Keywords and phrases: Linear regression, Generalization error, Shrinkage, PAC-Bayesian theorems, Risk bounds, Robust statistics, Resistant estimators, Gibbs posterior distributions, Randomized estimators, Statistical learning theory

so that

$$\overline{R}(Q_\lambda^{1/2}\theta) = R(\theta) = \mathbb{E}\big[(\langle\theta, X\rangle - Y)^2\big].$$

Let

$$\overline{\Theta} = \big\{Q_\lambda^{1/2}\theta; \theta \in \Theta\big\}.$$

Consider

$$r(\theta) = \frac{1}{n}\sum_{i=1}^n \big(\langle\theta, X_i\rangle - Y_i\big)^2, \tag{1.1}$$

$$\overline{r}(\theta) = \frac{1}{n}\sum_{i=1}^n \big(\langle\theta, \overline{X}_i\rangle - Y_i\big)^2, \tag{1.2}$$

$$\theta_0 = \arg\min_{\theta\in\overline{\Theta}} \overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2, \tag{1.3}$$

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2, \tag{1.4}$$

$$\theta_1 = Q_\lambda^{1/2}\hat{\theta} \in \arg\min_{\theta\in\overline{\Theta}} \overline{r}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2. \tag{1.5}$$

For $\alpha > 0$, let us introduce the notation

$$W_i(\theta) = \alpha\Big\{\big(\langle\theta, \overline{X}_i\rangle - Y_i\big)^2 - \big(\langle\theta_0, \overline{X}_i\rangle - Y_i\big)^2\Big\},$$
$$W(\theta) = \alpha\Big\{\big(\langle\theta, \overline{X}\rangle - Y\big)^2 - \big(\langle\theta_0, \overline{X}\rangle - Y\big)^2\Big\}.$$

For any $\theta_2 \in \mathbb{R}^d$ and $\beta > 0$, let us consider the Gaussian distribution centered at θ_2

$$\rho_{\theta_2}(d\theta) = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta}{2}\|\theta - \theta_2\|^2\right) d\theta.$$

Lemma 1.1. *For any $\eta > 0$ and $\alpha > 0$, with probability at least $1 - \exp(-\eta)$, for any $\theta_2 \in \mathbb{R}^d$,*

$$-n\int \log\Big\{1 - \mathbb{E}\big[W(\theta)\big] + \mathbb{E}\big[W(\theta)^2\big]/2\Big\}\,\rho_{\theta_2}(d\theta)$$
$$\leq -\sum_{i=1}^n \int \log\Big\{1 - W_i(\theta) + W_i(\theta)^2/2\Big\}\,\rho_{\theta_2}(d\theta) + \mathcal{K}(\rho_{\theta_2}, \rho_{\theta_0}) + \eta,$$

where $\mathcal{K}(\rho_{\theta_2}, \rho_{\theta_0})$ is the Kullback-Leibler divergence function :

$$\mathcal{K}(\rho_{\theta_2}, \rho_{\theta_0}) = \int \log\left[\frac{d\rho_{\theta_2}}{d\rho_{\theta_0}}(\theta)\right]\rho_{\theta_2}(d\theta).$$

PROOF. Since

$$\mathbb{E}\left(\int \rho_{\theta_0}(d\theta)\prod_{i=1}^{n}\frac{1-W_i(\theta)+W_i(\theta)^2/2}{1-\mathbb{E}\big[W(\theta)\big]+\mathbb{E}\big[W(\theta)^2\big]/2}\right)\leq 1,$$

with probability at least $1-\exp(-\eta)$

$$\log\left(\int \rho_{\theta_0}(d\theta)\prod_{i=1}^{n}\frac{1-W_i(\theta)+W_i(\theta)^2/2}{1-\mathbb{E}\big[W(\theta)\big]+\mathbb{E}\big[W(\theta)^2\big]/2}\right)\leq \eta.$$

We conclude the proof using the convex inequality (see [2], [3, Proposition 1.4.2] or [1, page 159])

$$\log\Big(\int \rho_{\theta_0}(d\theta)\exp\big[h(\theta)\big]\Big)\geq \int \rho_{\theta_2}(d\theta)h(\theta)-\mathcal{K}(\rho_{\theta_2},\rho_{\theta_0}).$$

□

Let us compute some useful quantities

$$\mathcal{K}(\rho_{\theta_2},\rho_{\theta_0})=\frac{\beta}{2}\|\theta_2-\theta_0\|^2, \tag{1.6}$$

$$\begin{aligned}\int \rho_{\theta_2}(d\theta)\big[W(\theta)\big]&=\alpha\int \rho_{\theta_2}(d\theta)\langle\theta-\theta_2,\overline{X}\rangle^2+W(\theta_2)\\&=W(\theta_2)+\frac{\alpha\|\overline{X}\|^2}{\beta},\end{aligned} \tag{1.7}$$

$$\int \rho_{\theta_2}(d\theta)\langle\theta-\theta_2,\overline{X}\rangle^4=\frac{3\|\overline{X}\|^4}{\beta^2}, \tag{1.8}$$

$$\begin{aligned}\int \rho_{\theta_2}(d\theta)\big[W(\theta)^2\big]&=\alpha^2\int \rho_{\theta_2}(d\theta)\langle\theta-\theta_0,\overline{X}\rangle^2\big(\langle\theta+\theta_0,\overline{X}\rangle-2Y\big)^2\\&=\alpha^2\int \rho_{\theta_2}(d\theta)\Big[\langle\theta-\theta_2+\theta_2-\theta_0,\overline{X}\rangle\big(\langle\theta-\theta_2+\theta_2+\theta_0,\overline{X}\rangle-2Y\big)\Big]^2\\&=\int \rho_{\theta_2}(d\theta)\Big[\alpha\langle\theta-\theta_2,\overline{X}\rangle^2+2\alpha\langle\theta-\theta_2,\overline{X}\rangle\big(\langle\theta_2,\overline{X}\rangle-Y\big)+W(\theta_2)\Big]^2\\&=\int \rho_{\theta_2}(d\theta)\Big[\alpha^2\langle\theta-\theta_2,\overline{X}\rangle^4+4\alpha^2\langle\theta-\theta_2,\overline{X}\rangle^2\big(\langle\theta_2,\overline{X}\rangle-Y\big)^2+W(\theta_2)^2\\&\qquad+2\alpha\langle\theta-\theta_2,\overline{X}\rangle^2W(\theta_2)\Big]\\&=\frac{3\alpha^2\|\overline{X}\|^4}{\beta^2}+\frac{2\alpha\|\overline{X}\|^2}{\beta}\Big[2\alpha\big(\langle\theta_2,\overline{X}\rangle-Y\big)^2+W(\theta_2)\Big]+W(\theta_2)^2.\end{aligned} \tag{1.9}$$

Using the fact that

$$2\alpha\big(\langle\theta_2,\overline{X}\rangle-Y\big)^2+W(\theta_2)=2\alpha\big(\langle\theta_0,\overline{X}\rangle-Y\big)^2+3W(\theta_2),$$

and that for any real numbers a and b, $6ab\leq 9a^2+b^2$, we get

LEMMA 1.2.

(1.10)
$$\int \rho_{\theta_2}(d\theta)\big[W(\theta)\big] = W(\theta_2) + \frac{\alpha\|\overline{X}\|^2}{\beta},$$
$$\int \rho_{\theta_2}(d\theta)\big[W(\theta)^2\big] = W(\theta_2)^2 + \frac{2\alpha\|\overline{X}\|^2}{\beta}\Big[2\alpha\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2 + 3W(\theta_2)\Big]$$
(1.11)
$$+ \frac{3\alpha^2\|\overline{X}\|^4}{\beta^2}$$
(1.12)
$$\leq 10W(\theta_2)^2 + \frac{4\alpha^2\|\overline{X}\|^2}{\beta}\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2 + \frac{4\alpha^2\|\overline{X}\|^4}{\beta^2},$$

and the same holds true when W is replaced with W_i and $(\overline{X}, Y)$ with $(\overline{X}_i, Y_i)$.

Another important thing to realize is that

$$\begin{aligned}
\mathbb{E}\big[\|\overline{X}\|^2\big] &= \mathbb{E}\big[\mathrm{Tr}\big(\overline{X}\,\overline{X}^T\big)\big] && = \mathbb{E}\big[\mathrm{Tr}\big(Q_\lambda^{-1/2} X X^T Q_\lambda^{-1/2}\big)\big] \\
&= \mathbb{E}\big[\mathrm{Tr}\big(Q_\lambda^{-1} X X^T\big)\big] && = \mathrm{Tr}\big[Q_\lambda^{-1}\mathbb{E}(XX^T)\big] \\
&= \mathrm{Tr}\big(Q_\lambda^{-1}(Q_\lambda - \lambda I)\big) && = d - \lambda\,\mathrm{Tr}(Q_\lambda^{-1}) \qquad = \; D\,. \qquad (1.13)
\end{aligned}$$

We can weaken Lemma 1.1 (page 2) noticing that for any real number x,

$$x - \frac{x^2}{2} \leq -\log\left(1 - x + \frac{x^2}{2}\right) = \log\left(\frac{1 + x + x^2/2}{1 + x^4/4}\right)$$
$$\leq \log\left(1 + x + \frac{x^2}{2}\right) \leq x + \frac{x^2}{2}.$$

We obtain with probability at least $1 - \exp(-\eta)$

$$n\mathbb{E}\big[W(\theta_2)\big] + \frac{n\alpha}{\beta}\mathbb{E}\big[\|\overline{X}\|^2\big] - 5n\mathbb{E}\big[W(\theta_2)^2\big]$$
$$- \mathbb{E}\left\{\frac{2n\alpha^2\|\overline{X}\|^2}{\beta}\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2 + \frac{2n\alpha^2\|\overline{X}\|^4}{\beta^2}\right\}$$
$$\leq \sum_{i=1}^{n}\Bigg\{W_i(\theta_2) + 5W_i(\theta_2)^2$$
$$+ \frac{\alpha\|\overline{X}_i\|^2}{\beta} + \frac{2\alpha^2\|\overline{X}_i\|^2}{\beta}\big(\langle\theta_0,\overline{X}_i\rangle - Y\big)^2 + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2}\Bigg\}$$

$$+\frac{\beta}{2}\|\theta_2-\theta_0\|^2+\eta.$$

Noticing that for any real numbers a and b, $4ab\le a^2+4b^2$, we can then bound

$$\begin{aligned}\alpha^{-2}W(\theta_2)^2&=\langle\theta_2-\theta_0,\overline{X}\rangle^2\big(\langle\theta_2+\theta_0,\overline{X}\rangle-2Y\big)^2\\&=\langle\theta_2-\theta_0,\overline{X}\rangle^2\Big[\langle\theta_2-\theta_0,\overline{X}\rangle+2\big(\langle\theta_0,\overline{X}\rangle-Y\big)\Big]^2\\&=\langle\theta_2-\theta_0,\overline{X}\rangle^4+4\langle\theta_2-\theta_0,\overline{X}\rangle^3\big(\langle\theta_0,\overline{X}\rangle-Y\big)\\&\qquad+4\langle\theta_2-\theta_0,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle-Y\big)^2\\&\qquad\le 2\langle\theta_2-\theta_0,\overline{X}\rangle^4+8\langle\theta_2-\theta_0,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle-Y\big)^2.\end{aligned}$$

THEOREM 1.3. *Let us put*

$$\widehat{D}=\frac{1}{n}\sum_{i=1}^n\|\overline{X}_i\|^2\qquad\textit{(let us remind that } D=\mathbb{E}\big[\|\overline{X}\|^2\big]\textit{ from (1.13))},$$

$$B_1=2\mathbb{E}\Big[\|\overline{X}\|^2\big(\langle\theta_0,\overline{X}\rangle-Y\big)^2\Big],$$

$$\widehat{B}_1=\frac{2}{n}\sum_{i=1}^n\Big[\|\overline{X}_i\|^2\big(\langle\theta_0,\overline{X}_i\rangle-Y_i\big)^2\Big],$$

$$B_2=2\mathbb{E}\Big[\|\overline{X}\|^4\Big],$$

$$\widehat{B}_2=\frac{2}{n}\sum_{i=1}^n\|\overline{X}_i\|^4,$$

$$B_3=40\sup\Big\{\mathbb{E}\big[\langle u,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle-Y\big)^2\big]:u\in\mathbb{R}^d,\|u\|=1\Big\},$$

$$\widehat{B}_3=\sup\Big\{\frac{40}{n}\sum_{i=1}^n\langle u,\overline{X}_i\rangle^2\big(\langle\theta_0,\overline{X}_i\rangle-Y_i\big)^2:u\in\mathbb{R}^d,\|u\|=1\Big\},$$

$$B_4=10\sup\Big\{\mathbb{E}\Big[\langle u,\overline{X}\rangle^4\Big]:u\in\mathbb{R}^d,\|u\|=1\Big\},$$

$$\widehat{B}_4=\sup\Big\{\frac{10}{n}\sum_{i=1}^n\langle u,\overline{X}_i\rangle^4:u\in\mathbb{R}^d,\|u\|=1\Big\}.$$

With probability at least $1-\exp(-\eta)$, *for any* $\theta_2\in\mathbb{R}^d$,

$$\begin{aligned}n\mathbb{E}\big[W(\theta_2)\big]-\Big[n\alpha^2(B_3+\widehat{B}_3)+\frac{\beta}{2}\Big]\|\theta_2-\theta_0\|^2\\-n\alpha^2(B_4+\widehat{B}_4)\|\theta_2-\theta_0\|^4\end{aligned}$$

$$\leq \sum_{i=1}^{n} W_i(\theta_2) + \frac{n\alpha}{\beta}(\widehat{D} - D) + \frac{n\alpha^2}{\beta}(B_1 + \widehat{B}_1) + \frac{n\alpha^2}{\beta^2}(B_2 + \widehat{B}_2) + \eta.$$

Let us now assume that $\theta_2 \in \overline{\Theta}$ and let us use the fact that $\overline{\Theta}$ is a convex set and that $\theta_0 = \arg\min_{\theta\in\overline{\Theta}} \overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2$. Introduce $\theta_* = \arg\min_{\theta\in\mathbb{R}^d} \overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2$. As we have

$$\overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2 = \|\theta - \theta_*\|^2 + \overline{R}(\theta_*) + \lambda\|Q_\lambda^{-1/2}\theta_*\|^2,$$

the vector θ_0 is uniquely defined as the projection of θ_* on $\overline{\Theta}$ for the Euclidean distance, and for any $\theta_2 \in \overline{\Theta}$

$$\begin{aligned}(1.14)\quad &\alpha^{-1}\mathbb{E}\big[W(\theta_2)\big] + \lambda\|Q_\lambda^{-1/2}\theta_2\|^2 - \lambda\|Q_\lambda^{-1/2}\theta_0\|^2\\ &\quad= \overline{R}(\theta_2) - \overline{R}(\theta_0) + \lambda\|Q_\lambda^{-1/2}\theta_2\|^2 - \lambda\|Q_\lambda^{-1/2}\theta_0\|^2\\ &\quad= \|\theta_2 - \theta_*\|^2 - \|\theta_0 - \theta_*\|^2\\ &\quad= \|\theta_2 - \theta_0\|^2 + 2\langle\theta_2 - \theta_0, \theta_0 - \theta_*\rangle \geq \|\theta_2 - \theta_0\|^2.\end{aligned}$$

This and the inequality

$$\alpha^{-1}\sum_{i=1}^{n} W_i(\theta_1) + n\lambda\|Q_\lambda^{-1/2}\theta_1\|^2 - n\lambda\|Q_\lambda^{-1/2}\theta_0\|^2 \leq 0$$

leads to the following result.

THEOREM 1.4. *With probability at least* $1 - \exp(-\eta)$,

$$\begin{aligned}&R(\hat{\theta}) + \lambda\|\hat{\theta}\|^2 - \inf_{\theta\in\Theta}\big[R(\theta) + \lambda\|\theta\|^2\big]\\ &\qquad= \alpha^{-1}\mathbb{E}\big[W(\theta_1)\big] + \lambda\|Q_\lambda^{-1/2}\theta_1\|^2 - \lambda\|Q_\lambda^{-1/2}\theta_0\|^2\end{aligned}$$

is not greater than the smallest positive non degenerate root of the following polynomial equation as soon as it has one

$$\begin{aligned}&\Big\{1 - \big[\alpha(B_3 + \widehat{B}_3) + \tfrac{\beta}{2n\alpha}\big]\Big\}x - \alpha(B_4 + \widehat{B}_4)x^2\\ &\qquad= \frac{1}{\beta}\max(\widehat{D} - D, 0) + \frac{\alpha}{\beta}(B_1 + \widehat{B}_1) + \frac{\alpha}{\beta^2}(B_2 + \widehat{B}_2) + \frac{\eta}{n\alpha}.\end{aligned}$$

PROOF. Let us remark first that when the polynomial appearing in the theorem has two distinct roots, they are of the same sign, due to the sign of its constant coefficient. Let $\widehat{\Omega}$ be the event of probability at least $1-\exp(-\eta)$

described in Theorem 1.3 (page 5). For any realization of this event for which the polynomial described in Theorem 1.4 does not have two distinct positive roots, the statement of Theorem 1.4 is void, and therefore fulfilled. Let us consider now the case when the polynomial in question has two distinct positive roots $x_1 < x_2$. Consider in this case the random (trivially nonempty) closed convex set

$$\widehat{\Theta} = \Big\{\theta \in \Theta : R(\theta) + \lambda\|\theta\|^2 \leq \inf_{\theta' \in \Theta}\big[R(\theta') + \lambda\|\theta'\|^2\big] + \tfrac{x_1+x_2}{2}\Big\}.$$

Let $\theta_3 \in \arg\min_{\theta\in\widehat{\Theta}} r(\theta) + \lambda\|\theta\|^2$ and $\theta_4 \in \arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2$. We see from Theorem 1.3 that

$$R(\theta_3) + \lambda\|\theta_3\|^2 < R(\theta_0) + \lambda\|\theta_0\|^2 + \frac{x_1 + x_2}{2}, \tag{1.15}$$

because it cannot be larger from the construction of $\widehat{\Theta}$. On the other hand, since $\widehat{\Theta} \subset \Theta$, the line segment $[\theta_3, \theta_4]$ is such that $[\theta_3, \theta_4] \cap \widehat{\Theta} \subset \arg\min_{\theta\in\widehat{\Theta}} r(\theta) + \lambda\|\theta\|^2$. We can therefore apply equation (1.15) to any point of $[\theta_3, \theta_4] \cap \widehat{\Theta}$, which proves that $[\theta_3, \theta_4] \cap \widehat{\Theta}$ is an open subset of $[\theta_3, \theta_4]$. But it is also a closed subset by construction, and therefore, as it is non empty and $[\theta_3, \theta_4]$ is connected, it proves that $[\theta_3, \theta_4] \cap \widehat{\Theta} = [\theta_3, \theta_4]$, and thus that $\theta_4 \in \widehat{\Theta}$. This can be applied to any choice of $\theta_3 \in \arg\min_{\theta\in\widehat{\Theta}} r(\theta) + \lambda\|\theta\|^2$ and $\theta_4 \in \arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2$, proving that $\arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2 \subset \arg\min_{\theta\in\widehat{\Theta}} r(\theta) + \lambda\|\theta\|^2$ and therefore that any $\theta_4 \in \arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2$ is such that

$$R(\theta_4) + \lambda\|\theta_4\|^2 \leq \inf_{\theta\in\Theta}\big[R(\theta) + \lambda\|\theta\|^2\big] + x_1.$$

because the values between x_1 and x_2 are excluded by Theorem 1.3. □

The actual convergence speed of the least squares estimator $\hat{\theta}$ on Θ will depend on the speed of convergence of the "empirical bounds" $\widehat{B}_k$ towards their expectations. We can rephrase the previous theorem in the following more practical way:

THEOREM 1.5. *Let $\eta_0, \eta_1, \ldots, \eta_5$ be positive real numbers. With probability at least*

$$1 - \mathbb{P}\big(\widehat{D} > D + \eta_0\big) - \sum_{k=1}^{4} \mathbb{P}\big(\widehat{B}_k - B_k > \eta_k\big) - \exp(-\eta_5),$$

$R(\hat{\theta}) + \lambda\|\hat{\theta}\|^2 - \inf_{\theta\in\Theta}\big[R(\theta) + \lambda\|\theta\|^2\big]$ is smaller than the smallest non degenerate positive root of

$$\Big\{1-\big[\alpha(2B_3+\eta_3)+\tfrac{\beta}{2n\alpha}\big]\Big\}x-\alpha(2B_4+\eta_4)x^2 = \frac{\eta_0}{\beta}+\frac{\alpha}{\beta}(2B_1+\eta_1)+\frac{\alpha}{\beta^2}(2B_2+\eta_2)+\frac{\eta_5}{n\alpha}, \tag{1.16}$$

where we can optimize the values of $\alpha>0$ and $\beta>0$, since this equation has non random coefficients. For example, taking for simplicity

$$\alpha=\frac{1}{8B_3+4\eta_3},$$
$$\beta=\frac{n\alpha}{2},$$

we obtain

$$x-\frac{2B_4+\eta_4}{4B_3+2\eta_3}x^2=\frac{16\eta_0(2B_3+\eta_3)}{n}+\frac{8B_1+4\eta_1}{n}+\frac{32(2B_3+\eta_3)(2B_2+\eta_2)}{n^2}+\frac{8\eta_5(2B_3+\eta_3)}{n}.$$

1.1. *Proof of Theorem 2.1.* Let us now deduce Theorem 2.1 from Theorem 1.5. Let us first remark that with probability at least $1-\varepsilon/2$

$$\widehat{D}\le D+\sqrt{\frac{B_2}{\varepsilon n}},$$

because the variance of $\widehat{D}$ is less than $\dfrac{B_2}{2n}$. For a given $\varepsilon>0$, let us take $\eta_0=\sqrt{\dfrac{B_2}{\varepsilon n}}$, $\eta_1=B_1$, $\eta_2=B_2$, $\eta_3=B_3$ and $\eta_4=B_4$. We get that $R_\lambda(\hat\theta)-\inf_{\theta\in\Theta}R_\lambda(\theta)$ is smaller than the smallest positive non degenerate root of

$$x-\frac{B_4}{2B_3}x^2=\frac{48B_3}{n}\sqrt{\frac{B_2}{n\varepsilon}}+\frac{12B_1}{n}+\frac{288B_2B_3}{n^2}+\frac{24\log(3/\varepsilon)B_3}{n},$$

with probability at least

$$1-\frac{5\,\varepsilon}{6}-\sum_{k=1}^{4}\mathbb{P}\big(\widehat{B}_k>B_k+\eta_k\big).$$

According to the weak law of large numbers, there is n_ε such that for any $n\ge n_\varepsilon$,

$$\sum_{k=1}^{4}\mathbb{P}\big(\widehat{B}_k>B_k+\eta_k\big)\le\varepsilon/6.$$

Thus, increasing n_ε and the constants to absorb the second order terms, we see that for some n_ε and any $n \geq n_\varepsilon$, with probability at least $1-\varepsilon$, the excess risk is less than the smallest positive root of

$$x - \frac{B_4}{2B_3}x^2 = \frac{13B_1}{n} + \frac{24\log(3/\varepsilon)B_3}{n}.$$

Now, as soon as $ac < 1/4$, the smallest positive root of $x - ax^2 = c$ is $\dfrac{2c}{1+\sqrt{1-4ac}}$. This means that for n large enough, with probability at least $1-\varepsilon$,

$$R_\lambda(\hat{\theta}) - \inf_\theta R_\lambda(\theta) \leq \frac{15B_1}{n} + \frac{25\log(3/\varepsilon)B_3}{n},$$

which is precisely the statement of Theorem 2.1, up to some change of notation.

1.2. *Proof of Theorem 2.2.* Let us now weaken Theorem 1.4 in order to make a more explicit non asymptotic result and obtain Theorem 2.2. From now on, we will assume that $\lambda = 0$. We start by giving bounds on the quantity defined in Theorem 1.3 in terms of

$$B = \sup_{f \in \mathrm{span}\{\varphi_1,\dots,\varphi_d\}-\{0\}} \|f\|_\infty^2 / \mathbb{E}[f(X)]^2.$$

Since we have

$$\|\overline{X}\|^2 = \|Q_\lambda^{-1/2}X\|^2 \leq dB,$$

we get

$$\begin{aligned}
\widehat{d} &= \frac{1}{n}\sum_{i=1}^n \|\overline{X}_i\|^2 \leq dB,\\
B_1 &= 2\mathbb{E}\Big[\|\overline{X}\|^2\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2\Big] \leq 2dB\, R(f^*),\\
\widehat{B}_1 &= \frac{2}{n}\sum_{i=1}^n \Big[\|\overline{X}_i\|^2\big(\langle\theta_0,\overline{X}_i\rangle - Y_i\big)^2\Big] \leq 2dB\, r(f^*),\\
B_2 &= 2\mathbb{E}\Big[\|\overline{X}\|^4\Big] \leq 2d^2B^2,\\
\widehat{B}_2 &= \frac{2}{n}\sum_{i=1}^n \|\overline{X}_i\|^4 \leq 2d^2B^2,\\
B_3 &= 40\sup\Big\{\mathbb{E}\big[\langle u,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2\big] : u\in\mathbb{R}^d, \|u\|=1\Big\} \leq 40B\, R(f^*),\\
\widehat{B}_3 &= \sup\Big\{\frac{40}{n}\sum_{i=1}^n \langle u,\overline{X}_i\rangle^2\big(\langle\theta_0,\overline{X}_i\rangle - Y_i\big)^2 : u\in\mathbb{R}^d, \|u\|=1\Big\} \leq 40B\, r(f^*),
\end{aligned}$$

$$B_4 = 10 \sup\Big\{ \mathbb{E}\Big[\langle u, \overline{X}\rangle^4\Big] : u \in \mathbb{R}^d, \|u\| = 1 \Big\} \le 10B^2,$$
$$\widehat{B}_4 = \sup\Big\{ \frac{10}{n} \sum_{i=1}^n \langle u, \overline{X}_i\rangle^4 : u \in \mathbb{R}^d, \|u\| = 1 \Big\} \le 10B^2.$$

Let us put

$$a_0 = \frac{2dB + 4dB\alpha[R(f^*) + r(f^*)] + \eta}{\alpha n} + \frac{16B^2d^2}{\alpha n^2},$$
$$a_1 = 3/4 - 40\alpha B[R(f^*) + r(f^*)],$$

and

$$a_2 = 20\alpha B^2.$$

Theorem 1.4 applied with $\beta = n\alpha/2$ implies that with probability at least $1-\eta$ the excess risk $R(\hat{f}^{(\mathrm{erm})}) - R(f^*)$ is upper bounded by the smallest positive root of $a_1 x - a_2 x^2 = a_0$ as soon as $a_1^2 > 4a_0a_2$. In particular, setting $\varepsilon = \exp(-\eta)$ when (1.17) holds, we have

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \le \frac{2a_0}{a_1 + \sqrt{a_1^2 - 4a_0a_2}} \le \frac{2a_0}{a_1}.$$

We conclude that

THEOREM 1.6. *For any $\alpha > 0$ and $\varepsilon > 0$, with probability at least $1-\varepsilon$, if the inequality*

$$80\left(\frac{(2 + 4\alpha[R(f^*) + r(f^*)])Bd + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd}{n}\right)^2 \right) < \left(\frac{3}{4B} - 40\alpha[R(f^*) + r(f^*)] \right)^2 \tag{1.17}$$

holds, then we have

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \le \mathcal{J}\left(\frac{(2 + 4\alpha[R(f^*) + r(f^*)])Bd + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd}{n}\right)^2 \right), \tag{1.18}$$

where $\mathcal{J} = 8/(3\alpha - 160\alpha^2 B[R(f^*) + r(f^*)])$

Now, the Bienaymé-Chebyshev inequality implies

$$\mathbb{P}\big(r(f^*) - R(f^*) \ge t\big) \le \frac{\mathbb{E}\big(r(f^*) - R(f^*)\big)^2}{t^2} \le \mathbb{E}[Y - f^*(X)]^4/nt^2.$$

Under the finite moment assumption of Theorem 2.2, we obtain that for any $\varepsilon \geq 1/n$, with probability at least $1-\varepsilon$,

$$r(f^*) < R(f^*) + \sqrt{\mathbb{E}[Y - f^*(X)]^4}.$$

From Theorem 1.6 and a union bound, by taking

$$\alpha = \left(80B[2R(f^*) + \sqrt{\mathbb{E}[Y - f^*(X)]^4}\right)^{-1},$$

we get that with probability $1-2\varepsilon$,

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \leq \mathcal{J}_1 B\left(\frac{3Bd + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd}{n}\right)^2\right),$$

with $\mathcal{J}_1 = 640\left(2R(f^*) + \sqrt{\mathbb{E}\{[Y - f^*(X)]^4\}}\right)$. This concludes the proof of Theorem 2.2.

REMARK 1.1. *Let us indicate now how to handle the case when Q is degenerate. Let us consider the linear subspace S of $\mathbb{R}^d$ spanned by the eigenvectors of Q corresponding to positive eigenvalues. Then almost surely $Span\{X_i, i = 1, \ldots, n\} \subset S$. Indeed for any θ in the kernel of Q, $\mathbb{E}\big(\langle \theta, X\rangle^2\big) = 0$ implies that $\langle \theta, X\rangle = 0$ almost surely, and considering a basis of the kernel, we see that $X \in S$ almost surely, S being orthogonal to the kernel of Q. Thus we can restrict the problem to S, as soon as we choose*

$$\hat{\theta} \in \mathrm{span}\big\{X_1, \ldots, X_n\big\} \cap \arg\min_{\theta} \sum_{i=1}^{n} \big(\langle \theta, X_i\rangle - Y_i\big)^2,$$

or equivalently with the notation $\mathbf{X} = (\varphi_j(X_i))_{1\leq i\leq n, 1\leq j\leq d}$ and $\mathbf{Y} = [Y_j]_{j=1}^n$,

$$\hat{\theta} \in im\,\mathbf{X}^T \cap \arg\min_{\theta} \|\,\mathbf{X}\,\theta - Y\|^2$$

This proves that the results of this section apply to this special choice of the empirical least squares estimator. Since we have $\mathbb{R}^d = ker\,\mathbf{X} \oplus im\,\mathbf{X}^T$, this choice is unique. Finally, we also have that inequality (2.3) of the paper still holds by replacing d by $\mathrm{rank}(Q)$.

2. Proof of Theorem 3.1. We use the same notations as in Section 1. We write X for $\varphi(X)$, therefore, the function f_θ maps an input x to $\langle \theta, x\rangle$. We consider the change of coordinates

$$\overline{X} = Q_\lambda^{-1/2} X.$$

Thus, from (1.13), we have $\mathbb{E}\big[\|\overline{X}\|^2\big] = D$. We will use

$$\overline{R}(\theta) = \mathbb{E}\big[(\langle\theta, \overline{X}\rangle - Y)^2\big],$$

so that $\overline{R}(Q^{1/2}\theta) = \mathbb{E}\big[(\langle\theta, X\rangle - Y)^2\big] = R(f_\theta)$. Let

$$\overline{\Theta} = \big\{Q_\lambda^{1/2}\theta; \theta \in \Theta\big\},$$

and consider

$$\theta_0 = \arg\min_{\theta\in\overline{\Theta}} \Big\{\overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2\Big\}.$$

With these notations,

$$\begin{aligned}
\tilde{\theta} &= Q_\lambda^{-1/2}\theta_0,\\
\sigma &= \sqrt{\mathbb{E}\big[\big(\langle\theta_0, \overline{X}\rangle - Y\big)^2\big]},\\
\chi &= \sup_{u\in\mathbb{R}^d} \frac{\mathbb{E}\big(\langle u, \overline{X}\rangle^4\big)^{1/2}}{\mathbb{E}\big(\langle u, \overline{X}\rangle^2\big)},\\
\kappa &= \frac{\mathbb{E}\big(\|\overline{X}\|^4\big)^{1/2}}{\mathbb{E}\big(\|\overline{X}\|^2\big)} = \frac{\mathbb{E}\big(\|\overline{X}\|^4\big)^{1/2}}{D},\\
\kappa' &= \frac{\mathbb{E}\big[\big(\langle\theta_0, \overline{X}\rangle - Y\big)^4\big]^{1/2}}{\sigma^2},\\
\text{and}\quad T &= \|\overline{\Theta}\| = \max_{\theta,\theta'\in\overline{\Theta}} \|\theta - \theta'\|.
\end{aligned}$$

For $\alpha > 0$, we introduce

$$\begin{aligned}
J_i(\theta) &= \langle\theta, \overline{X}_i\rangle - Y_i, & J(\theta) &= \langle\theta, \overline{X}\rangle - Y\\
\overline{L}_i(\theta) &= \alpha\big(\langle\theta, \overline{X}_i\rangle - Y_i\big)^2, & \overline{L}(\theta) &= \alpha\big(\langle\theta, \overline{X}\rangle - Y\big)^2\\
W_i(\theta) &= \overline{L}_i(\theta) - \overline{L}_i(\theta_0), & W(\theta) &= \overline{L}(\theta) - \overline{L}(\theta_0),
\end{aligned}$$

and

$$r'(\theta,\theta') = \lambda\Big(\|Q_\lambda^{-1/2}\theta\|^2 - \|Q_\lambda^{-1/2}\theta'\|^2\Big) + \frac{1}{n\alpha}\sum_{i=1}^n \psi\big(\overline{L}(\theta) - \overline{L}(\theta')\big).$$

Let $\bar{\theta} = Q_\lambda^{1/2}\hat{\theta} \in \overline{\Theta}$. We have

$$-r'(\theta_0,\bar{\theta}) = r'(\bar{\theta},\theta_0) \quad \le \max_{\theta_1\in\overline{\Theta}} r'(\bar{\theta},\theta_1) \quad \le \gamma + \max_{\theta_1\in\overline{\Theta}} r'(\theta_0,\theta_1), \tag{2.1}$$

where the quantity $\gamma = \max_{\theta_1 \in \overline{\Theta}} r'(\bar{\theta}, \theta_1) - \inf_{\theta \in \overline{\Theta}} \max_{\theta_1 \in \overline{\Theta}} r'(\theta, \theta_1)$ can be made arbitrary small by a proper choice of the estimator. Using an upper bound $r'(\theta_0, \theta_1)$ that holds uniformly in θ_1, we will control both left and right hand sides of (2.1).

To achieve this, we will upper bound

$$r'(\theta_0, \theta_1) = \lambda\Big(\|Q_\lambda^{-1/2}\theta_0\|^2 - \|Q_\lambda^{-1/2}\theta_1\|^2\Big) + \frac{1}{n\alpha}\sum_{i=1}^{n} \psi\big[-W_i(\theta_1)\big] \tag{2.2}$$

by the expectation of a distribution depending on θ_1 of a *quantity that does not depend on* θ_1, and then use the PAC-Bayesian argument to control this expectation uniformly in θ_1. The distribution depending on θ_1 should therefore be taken such that for any $\theta_1 \in \overline{\Theta}$, its Kullback-Leibler divergence with respect to some fixed distribution is small (at least when θ_1 is close to θ_0).

Let us start with the following result.

LEMMA 2.1. *Let $f, g : \mathbb{R} \to \mathbb{R}$ be two Lebesgue measurable functions such that $f(x) \le g(x)$, $x \in \mathbb{R}$. Let us assume that there exists $h \in \mathbb{R}$ such that $x \mapsto g(x) + hx^2/2$ is convex. Then for any probability distribution μ on the real line,*

$$f\left(\int x\mu(dx)\right) \le \int g(x)\mu(dx) + \min\Big\{\sup f - \inf f, \frac{h}{2}\,\mathbb{V}\mathrm{ar}(\mu)\Big\}.$$

PROOF. Let us put $x_0 = \int x\mu(dx)$ The function

$$x \mapsto g(x) + \frac{h}{2}(x - x_0)^2$$

is convex. Thus, by Jensen's inequality

$$f(x_0) \le g(x_0) \le \int \mu(dx)\left[g(x) + \frac{h}{2}(x - x_0)^2\right] = \int g(x)\mu(dx) + \frac{h}{2}\,\mathbb{V}\mathrm{ar}(\mu).$$

On the other hand

$$f(x_0) \le \sup f \le \sup f + \int \big[g(x) - \inf f\big]\mu(dx)$$
$$= \int g(x)\mu(dx) + \sup f - \inf f.$$

The lemma is a combination of these two inequalities. □

The above lemma will be used with $f = g = \psi$, where ψ is the increasing influence function

$$\psi(x) = \begin{cases} -\log(2), & x \le -1, \\ \log(1 + x + x^2/2), & -1 \le x \le 0, \\ -\log(1 - x + x^2/2), & 0 \le x \le 1, \\ \log(2), & x \ge 1. \end{cases}$$

Since we have for any $x \in \mathbb{R}$

$$-\log\left(1 - x + \frac{x^2}{2}\right) = \log\left(\frac{1 + x + \frac{x^2}{2}}{1 + \frac{x^4}{4}}\right) < \log\left(1 + x + \frac{x^2}{2}\right),$$

the function ψ satisfies for any $x \in \mathbb{R}^*$

$$-\log\left(1 - x + \frac{x^2}{2}\right) < \psi(x) < \log\left(1 + x + \frac{x^2}{2}\right).$$

Moreover

$$\psi'(x) = \frac{1 - x}{1 - x + \frac{x^2}{2}}, \quad \psi''(x) = \frac{x(x-2)}{2\left(1 - x + \frac{x^2}{2}\right)^2} \ge -2, \quad 0 \le x \le 1,$$

showing (by symmetry) that the function $x \mapsto \psi(x) + 2x^2$ is convex on the real line.

For any $\theta' \in \mathbb{R}^d$ and $\beta > 0$, we consider the Gaussian distribution with mean θ' and covariance $\beta^{-1} I$:

$$\rho_{\theta'}(d\theta) = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta}{2}\|\theta - \theta'\|^2\right) d\theta.$$

From Lemmas 1.2 and 2.1 (with μ the distribution of $-W_i(\theta) + \frac{\alpha\|\overline{X}_i\|^2}{\beta}$ when θ is drawn from ρ_{θ_1} and for a fixed pair (X_i, Y_i)), we can see that

$$\begin{aligned} \psi\big[-W_i(\theta_1)\big] &= \psi\left\{\int \rho_{\theta_1}(d\theta)\left[-W_i(\theta) + \frac{\alpha\|\overline{X}_i\|^2}{\beta}\right]\right\} \\ &\le \int \rho_{\theta_1}(d\theta)\psi\left[-W_i(\theta) + \frac{\alpha\|\overline{X}_i\|^2}{\beta}\right] \\ &\qquad + \min\Big\{\log(4), \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[\overline{L}_i(\theta)\big]\Big\}. \end{aligned}$$

Let us compute

$$\frac{1}{\alpha^2}\,\mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\Big(\overline{L}_i(\theta)\Big) = \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[J_i^2(\theta) - J_i^2(\theta_1)\big]$$

$$
\begin{aligned}
&= \int \rho_{\theta_1}(d\theta)\big[J_i^2(\theta) - J_i^2(\theta_1)\big]^2 - \frac{\|\overline{X}_i\|^4}{\beta^2} \\
&= \int \rho_{\theta_1}(d\theta)\Big[\langle \theta-\theta_1, \overline{X}_i\rangle^2 + 2\langle \theta-\theta_1, \overline{X}_i\rangle J_i(\theta_1)\Big]^2 - \frac{\|\overline{X}_i\|^4}{\beta^2} \\
&= \frac{2\|\overline{X}_i\|^4}{\beta^2} + \frac{4\overline{L}_i(\theta_1)\|\overline{X}_i\|^2}{\alpha\beta}. \qquad (2.3)
\end{aligned}
$$

Let $\xi \in (0,1)$, and let us remark that

$$
\overline{L}_i(\theta_1) \le \frac{\overline{L}_i(\theta)}{\xi} + \frac{\alpha\langle \theta-\theta_1, \overline{X}_i\rangle^2}{1-\xi}.
$$

We get

$$
\begin{aligned}
&\min\Big\{\log(4), \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[\overline{L}_i(\theta)\big]\Big\} \\
&\qquad = \min\Big\{\log(4), \frac{4\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta_1)}{\beta} + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2}\Big\} \\
&\le \int \rho_{\theta_1}(d\theta)\min\Big\{\log(4), \\
&\qquad\qquad \frac{4\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta)}{\beta\xi} + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2} + \frac{4\alpha^2\|\overline{X}_i\|^2\langle \theta-\theta_1, \overline{X}_i\rangle^2}{\beta(1-\xi)}\Big\} \\
&\qquad \le \int \rho_{\theta_1}(d\theta)\min\Big\{\log(4), \frac{4\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta)}{\beta\xi} + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2}\Big\} \\
&\qquad\qquad + \min\Big\{\log(4), \frac{4\alpha^2\|\overline{X}_i\|^4}{\beta^2(1-\xi)}\Big\}.
\end{aligned}
$$

Let us now put $a = \dfrac{3}{\log(4)} < 2.17$, $b = a + a^2\log(4) < 8.7$ and let us remark that

$$
\begin{aligned}
&\min\big\{\log(4), x\big\} + \min\big\{\log(4), y\big\} \\
&\qquad \le \log\big[1 + a\min\{\log(4), x\}\big] + \log(1+ay) \\
&\qquad\qquad \le \log\big(1 + ax + by\big), \qquad x, y \in \mathbb{R}_+.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\min\Big\{\log(4), \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[\overline{L}_i(\theta)\big]\Big\} \\
&\qquad \le \int \rho_{\theta_1}(d\theta)\log\Big[1 + \frac{4a\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta)}{\beta\xi} + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2}\Big(a + \frac{2b}{1-\xi}\Big)\Big].
\end{aligned}
$$

We can then remark that

$$\psi(x)+\log(1+y)=\log\big[\exp[\psi(x)]+y\exp[\psi(x)]\big]$$
$$\leq\log\big[\exp[\psi(x)]+2y\big]\leq\log\bigg(1+x+\frac{x^2}{2}+2y\bigg),\qquad x\in\mathbb{R},y\in\mathbb{R}_+.$$

Thus, putting $c_0=a+\dfrac{2b}{1-\xi}$, we get

$$\psi\big[-W_i(\theta_1)\big]\leq\int\rho_{\theta_1}(d\theta)\log[A_i(\theta)],\tag{2.4}$$

with

$$A_i(\theta)=1-W_i(\theta)+\frac{\alpha\|\overline{X}_i\|^2}{\beta}+\frac{1}{2}\bigg(-W_i(\theta)+\frac{\alpha\|\overline{X}_i\|^2}{\beta}\bigg)^2$$
$$+\frac{8a\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta)}{\beta\xi}+\frac{4c_0\alpha^2\|\overline{X}_i\|^4}{\beta^2}.$$

Similarly, we define $A(\theta)$ by replacing (X_i,Y_i) by (X,Y). Since we have

$$\mathbb{E}\bigg[\exp\bigg(\sum_{i=1}^n\log[A_i(\theta)]-n\log[\mathbb{E}A(\theta)]\bigg)\bigg]=1,$$

from the usual PAC-Bayesian argument, we have with probability at least $1-\varepsilon$, for any $\theta_1\in\mathbb{R}^d$,

$$\int\rho_{\theta_1}(d\theta)\bigg(\sum_{i=1}^n\log[A_i(\theta)]\bigg)-n\int\rho_{\theta_1}(d\theta)\log[A(\theta)]\leq K(\rho_{\theta_1},\rho_{\theta_0})+\log(\varepsilon^{-1})$$
$$\leq\frac{\beta\|\theta_1-\theta_0\|^2}{2}+\log(\varepsilon^{-1}).$$

From (2.2) and (2.4), with probability at least $1-\varepsilon$, for any $\theta_1\in\mathbb{R}^d$, we get

$$r'(\theta_0,\theta_1)\leq\frac{1}{\alpha}\log\bigg\{1+\mathbb{E}\bigg[\int\rho_{\theta_1}(d\theta)\bigg(-W(\theta)+\frac{\alpha\|\overline{X}\|^2}{\beta}$$
$$+\frac{1}{2}\bigg(-W(\theta)+\frac{\alpha\|\overline{X}\|^2}{\beta}\bigg)^2+\frac{8a\alpha\|\overline{X}\|^2\overline{L}(\theta)}{\beta\xi}+\frac{4c_0\alpha^2\|\overline{X}\|^4}{\beta^2}\bigg)\bigg]\bigg\}$$
$$+\frac{\beta\|\theta_1-\theta_0\|^2}{2n\alpha}+\frac{\log(\varepsilon^{-1})}{n\alpha}+\lambda\Big(\|Q_\lambda^{-1/2}\theta_0\|^2-\|Q_\lambda^{-1/2}\theta_1\|^2\Big).$$

Moreover from (2.3) and $\frac{\alpha\|\overline{X}\|^2}{\beta} = -\overline{L}(\theta_1) + \int \rho_{\theta_1}(d\theta)\overline{L}(\theta)$, we deduce that

$$\int \rho_{\theta_1}(d\theta)\Big(-W(\theta) + \frac{\alpha\|\overline{X}\|^2}{\beta}\Big)^2 = \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[\overline{L}(\theta)\big] + W(\theta_1)^2$$
$$= W(\theta_1)^2 + \frac{4\alpha\overline{L}(\theta_1)\|\overline{X}\|^2}{\beta} + \frac{2\alpha^2\|\overline{X}\|^4}{\beta^2}.$$

PROPOSITION 2.2. *With probability at least $1-\varepsilon$, for any $\theta_1 \in \mathbb{R}^d$,*

$$r'(\theta_0,\theta_1) \le \frac{1}{\alpha}\log\Big\{1 + \mathbb{E}\Big[-W(\theta_1) + \frac{W(\theta_1)^2}{2} + \frac{\big(2+8a/\xi\big)\alpha\|\overline{X}\|^2\overline{L}(\theta_1)}{\beta}$$
$$+ \frac{\big(1+8a/\xi+4c_0\big)\alpha^2\|\overline{X}\|^4}{\beta^2}\Big]\Big\} + \frac{\beta\|\theta_1-\theta_0\|^2}{2n\alpha} + \frac{\log(\varepsilon^{-1})}{n\alpha}$$
$$+ \lambda\Big(\|Q_\lambda^{-1/2}\theta_0\|^2 - \|Q_\lambda^{-1/2}\theta_1\|^2\Big)$$
$$\le \mathbb{E}\Big[J(\theta_0)^2 - J(\theta_1)^2 + \frac{1}{2\alpha}W(\theta_1)^2 + \frac{(2+8a/\xi)\|\overline{X}\|^2\overline{L}(\theta_1)}{\beta}$$
$$+ \frac{(1+8a/\xi+4c_0)\alpha\|\overline{X}\|^4}{\beta^2}\Big] + \frac{\beta\|\theta_1-\theta_0\|^2}{2n\alpha} + \frac{\log(\varepsilon^{-1})}{n\alpha}$$
$$+ \lambda\Big(\|Q_\lambda^{-1/2}\theta_0\|^2 - \|Q_\lambda^{-1/2}\theta_1\|^2\Big).$$

Using the triangular inequality and Cauchy-Schwarz's inequality, we get

$$\frac{1}{\alpha^2}\mathbb{E}\big[W(\theta_1)^2\big] = \mathbb{E}\Big\{\big[\langle\theta_1-\theta_0,\overline{X}\rangle^2 + 2\langle\theta_1-\theta_0,\overline{X}\rangle J(\theta_0)\big]^2\Big\}$$
$$\le \Big\{\mathbb{E}\big[\langle\theta_1-\theta_0,\overline{X}\rangle^4\big]^{1/2} + 2\mathbb{E}\big[\langle\theta_1-\theta_0,\overline{X}\rangle^4\big]^{1/4}\mathbb{E}\big[J(\theta_0)^4\big]^{1/4}\Big\}^2 \tag{2.5}$$
$$\le \Bigg\{\chi\|\theta_1-\theta_0\|^2\mathbb{E}\Big[\Big\langle\frac{\theta_1-\theta_0}{\|\theta_1-\theta_0\|},\overline{X}\Big\rangle^2\Big]$$
$$+ 2\|\theta_1-\theta_0\|\sigma\sqrt{\kappa'\chi}\sqrt{\mathbb{E}\Big[\Big\langle\frac{\theta_1-\theta_0}{\|\theta_1-\theta_0\|},\overline{X}\Big\rangle^2\Big]}\Bigg\}^2$$
$$\le \frac{\chi q_{\max}}{q_{\max}+\lambda}\|\theta_1-\theta_0\|^2\Big\{\|\theta_1-\theta_0\|\sqrt{\frac{\chi q_{\max}}{q_{\max}+\lambda}} + 2\sigma\sqrt{\kappa'}\Big\}^2,$$

and

$$\frac{1}{\alpha}\mathbb{E}\big[\|\overline{X}\|^2\overline{L}(\theta_1)\big] = \mathbb{E}\Big\{\big[\|\overline{X}\|\langle\theta_1-\theta_0,\overline{X}\rangle + \|\overline{X}\|J(\theta_0)\big]^2\Big\}$$

$$\leq \mathbb{E}\big[\|\overline{X}\|^4\big]^{1/2}\Big\{\mathbb{E}\big[\langle\theta_1-\theta_0,\overline{X}\rangle^4\big]^{1/4}+\mathbb{E}\big[J(\theta_0)^4\big]^{1/4}\Big\}^2 \tag{2.6}$$
$$\leq \kappa D\left\{\|\theta_1-\theta_0\|\sqrt{\frac{\chi q_{\max}}{q_{\max}+\lambda}}+2\sigma\sqrt{\kappa'}\right\}^2.$$

Let us put

$$\begin{aligned}
\widetilde{R}(\theta) &= \overline{R}(\theta)+\lambda\|Q_\lambda^{-1/2}\theta\|^2,\\
c_1 &= 4(2+8a/\xi),\\
c_2 &= 4(1+8a/\xi+4c_0),\\
\delta &= \frac{c_1\kappa\kappa' D\sigma^2}{n}+\frac{2\chi\left(\dfrac{\log(\varepsilon^{-1})}{n}+\dfrac{c_2\kappa^2D^2}{n^2}\right)\left(2\sqrt{\kappa'}\sigma+\|\overline{\Theta}\|\sqrt{\chi}\right)^2}{1-\dfrac{4c_1\kappa\chi D}{n}}.
\end{aligned}$$

We have proved the following result.

Proposition 2.3. *With probability at least $1-\varepsilon$, for any $\theta_1\in\mathbb{R}^d$,*

$$\begin{aligned}
r'(\theta_0,\theta_1) \leq \widetilde{R}(\theta_0)-\widetilde{R}(\theta_1)&+\frac{\alpha}{2}\chi\|\theta_1-\theta_0\|^2\big[2\sqrt{\kappa'}\sigma+\|\theta_1-\theta_0\|\sqrt{\chi}\big]^2\\
&+\frac{c_1\alpha}{4\beta}\kappa D\big[\sqrt{\kappa'}\sigma+\|\theta_1-\theta_0\|\sqrt{\chi}\big]^2+\frac{c_2\alpha\kappa^2D^2}{4\beta^2}\\
&+\frac{\beta\|\theta_1-\theta_0\|^2}{2n\alpha}+\frac{\log(\varepsilon^{-1})}{n\alpha}.
\end{aligned}$$

Let us assume from now on that $\theta_1\in\overline{\Theta}$, our convex bounded parameter set. In this case, as seen in (1.14), we have $\|\theta_0-\theta_1\|^2\leq\widetilde{R}(\theta_1)-\widetilde{R}(\theta_0)$. We can also use the fact that

$$\big[\sqrt{\kappa'}\sigma+\|\theta_1-\theta_0\|\sqrt{\chi}\big]^2\leq 2\kappa'\sigma^2+2\chi\|\theta_1-\theta_0\|^2.$$

We deduce from these remarks that with probability at least $1-\varepsilon$,

$$\begin{aligned}
r'(\theta_0,\theta_1) \leq \left\{-1+\frac{\alpha\chi}{2}\big[2\sqrt{\kappa'}\sigma+\|\overline{\Theta}\|\sqrt{\chi}\big]^2+\frac{\beta}{2n\alpha}+\frac{c_1\alpha\kappa D\chi}{2\beta}\right\}\big[\widetilde{R}(\theta_1)-\widetilde{R}(\theta_0)\big]&\\
+\frac{c_1\alpha\kappa D\kappa'\sigma^2}{2\beta}+\frac{c_2\alpha\kappa^2D^2}{4\beta^2}+\frac{\log(\varepsilon^{-1})}{n\alpha}.&
\end{aligned}$$

Let us assume that $n>4c_1\kappa\chi D$ and let us choose

$$\beta=\frac{n\alpha}{2},$$

$$\alpha = \frac{1}{2\chi\left[2\sqrt{\kappa'}\sigma + \|\overline{\Theta}\|\sqrt{\chi}\right]^2}\left(1 - \frac{4c_1\kappa\chi D}{n}\right),$$

to get

$$r'(\theta_0, \theta_1) \leq -\frac{\widetilde{R}(\theta_1) - \widetilde{R}(\theta_0)}{2} + \delta.$$

Plugging this into (2.1), we get

$$\frac{\widetilde{R}(\bar{\theta}) - \widetilde{R}(\theta_0)}{2} - \delta \leq r'(\bar{\theta}, \theta_0) \leq \max_{\theta_1 \in \overline{\Theta}} \left(\frac{\widetilde{R}(\theta_0) - \widetilde{R}(\theta_1)}{2}\right) + \gamma + \delta = \gamma + \delta,$$

hence

$$\widetilde{R}(\bar{\theta}) - \widetilde{R}(\theta_0) \leq 2\gamma + 4\delta.$$

Computing the numerical values of the constants when $\xi = 0.8$ gives $c_1 < 95$ and $c_2 < 1511$.

References.

[1] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. Pages 1–269.

[2] M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

[3] P. Dupuis and R.S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley-Interscience, 1997.