A Notation

For reader's convenience we collect the main notation we introduced in the paper.

Notation: We denote with the "hat", e.g. $\widehat{w}$, random quantities depending on the data. Given a linear operator A we denote by $A^\top$ its adjoint (transpose for matrices). For any $n \in \mathbb{N}$, we denote by $\langle , \rangle_n , \|\|_n$ the inner product and norm in $\mathbb{R}^n$. Given two quantities a, b (depending on some parameters), the notation $a \lesssim b$, or $a = O(b)$ means that there exists constant such that $a \leqslant Cb$.

Table 3: Definition of the main quantities used in the paper

	Definition
$L(w)$	$\int_{\mathcal{H}\times\mathcal{Y}} \ell(y, \langle w, x\rangle) dP(x, y)$
$L_\lambda(w)$	$L(w) + \lambda\|w\|^2$
$\widehat{L}(w)$	$n^{-1}\sum_{i=1}^n \ell(y_i, \langle w, x_i\rangle)$
$\widehat{L}_\lambda(w)$	$\widehat{L}(w) + \lambda\|w\|^2$
w_*	$\arg\min_{w\in\mathcal{H}} L(w)$
w_λ	$\arg\min_{w\in\mathcal{H}} L_\lambda(w)$
$\widehat{w}_\lambda$	$\arg\min_{w\in\mathcal{H}} \widehat{L}_\lambda(w)$
$\beta_{\lambda,\mathcal{B}}$	$\arg\min_{\beta\in\mathcal{B}} L_\lambda(\beta)$
$\widehat{\beta}_{\lambda,\mathcal{B}}$	$\arg\min_{\beta\in\mathcal{B}} \widehat{L}_\lambda(\beta)$
$f_*(x)$	$\arg\min_{a\in\mathbb{R}} \int_{\mathcal{Y}} \ell(y, a) dP(y\|x)$
$\mathcal{B}_m$	$\mathcal{B}_m = \text{span}\{\widetilde{x}_1, \dots, \widetilde{x}_m\}$
$\mathcal{P}_\mathcal{B}$	proj operator onto $\mathcal{B}$
$\mathcal{P}_m$	proj operator onto $\mathcal{B}_m$

B Proof of Theorem 1

This section is devoted to the proof of Theorem 1. In the following we restrict to linear functions, *i.e* $f(x) = \langle w, x\rangle$ for some $w \in \mathcal{H}$ and, with slight abuse of notation we set

$$\ell(w, z) = \ell(y, \langle w, x\rangle), \qquad z = (x, y) \in \mathcal{H}\times\mathcal{Y},\ w \in \mathcal{H}.$$

With this notation $L(w) = \int_{\mathcal{H}\times\mathcal{Y}} \ell(w, z) dP(z)$. The Lipschitz assumption implies that $\ell(\cdot, (X, Y))$ is almost surely Lipschitz in its argument, with Lipschitz constant $G\kappa$.

Specifically, we will show the following:

Theorem 6. *Under Assumptions 1, 2, for $\lambda > 0$ and $\delta \in (0, 1)$ let*

$$C_{\lambda,\delta} = 4\big\{1 + \sqrt{\log(1 + \log_2(3 + \ell_0\kappa^2/\lambda)) + \log(2/\delta)}\big\} = O(\sqrt{\log\log(3 + \ell_0\kappa^2/\lambda) + \log(1/\delta)}).$$

If Assumption 3 holds, then with probability $1 - \delta$,

$$L(\widehat{w}_\lambda) < \inf_{\mathcal{H}} L + \lambda\|w_*\|^2 + \frac{C_{\lambda,\delta}^2 G^2\kappa^2}{4\lambda n} + \frac{GC_{\lambda,\delta}}{\sqrt{n}} + (\ell_0 + G\kappa\|w_*\|)\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{27}$$

More generally, with probability $1 - \delta$, letting $\mathcal{A}(\lambda) := \inf_{w\in\mathcal{H}} L(w) + \lambda\|w\|^2 - \inf_{w\in\mathcal{H}} L(w)$,

$$\begin{aligned} L(\widehat{w}_\lambda) - \inf_{\mathcal{H}} L &< 2\mathcal{A}(\lambda) + \frac{C_{\lambda,\delta}^2 G^2\kappa^2 + 8G^2\kappa^2\log(2/\delta)}{4\lambda n} + \frac{GC_{\lambda,\delta}}{\sqrt{n}} + \ell_0\sqrt{\frac{2\log(2/\delta)}{n}} \\ &\leqslant 2\Big(\inf_{\|w\|\leqslant R} L(w) - \inf_{\mathcal{H}} L\Big) + 2\lambda R^2 + \frac{C_{\lambda,\delta}^2 G^2\kappa^2 + 8G^2\kappa^2\log(2/\delta)}{4\lambda n} + \frac{GC_{\lambda,\delta} + \ell_0\sqrt{2\log(2/\delta)}}{\sqrt{n}} \end{aligned} \tag{28}$$

for every $R > 0$.

The proof starts with the following bound on the generalization gap $L(w) - \widehat{L}(w)$ uniformly over balls. While this result is well-known and follows from standard arguments (see, e.g., Bartlett and Mendelson (2002); Koltchinskii (2011)), we include a short proof for completeness.

Lemma 1. *Under Assumptions 1 and 2 and, for every $R > 0$, one has with probability at least $1 - \delta$,*

$$\sup_{\|w\| \leqslant R} \big[L(w) - \widehat{L}(w)\big] < \frac{GR\kappa}{\sqrt{n}} \Big(2 + \sqrt{2\log(1/\delta)}\Big). \tag{29}$$

Proof of Lemma 1. The proof starts by a standard *symmetrization* step [Giné and Zinn (1984); Koltchinskii (2011)]. Let us call $D := (z_1, \dots, z_n)$ i.i.d. from P, as well as an independent $D' := (z'_1, \dots, z'_n)$ i.i.d. from P and $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. with $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$. We denote $\widehat{L}'(w) := n^{-1} \sum_{i=1}^n \ell(w, z'_i)$ the error on the sample D'. Then,

$$\begin{aligned}
\mathbb{E}_{D\sim P^n} \sup_{\|w\| \leqslant R} \Big[L(w) - \widehat{L}(w)\Big] &= \mathbb{E}_D \sup_{\|w\| \leqslant R} \Big[\mathbb{E}_{D'} \widehat{L}'(w) - \widehat{L}(w)\Big] \\
&\leqslant \mathbb{E}_{D,D'} \sup_{\|w\| \leqslant R} \Big[\widehat{L}'(w) - \widehat{L}(w)\Big] \\
&= \mathbb{E}_{D,D',\varepsilon} \sup_{\|w\| \leqslant R} \Big[\frac{1}{n} \sum_{i=1}^n \varepsilon_i \big(\ell(w, z_i) - \ell(w, z'_i)\big)\Big] \\
&= 2\mathbb{E}_{D,\varepsilon} \Big[\sup_{\|w\| \leqslant R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell(w, z_i)\Big]
\end{aligned}$$

where we used that $\mathbb{E}_{D'} \widehat{L}'(\cdot) = L(\cdot)$, and that $(\ell(f, z_i) - \ell(f, z'_i))_{1 \leqslant i \leqslant n}$ and $(\varepsilon_i(\ell(f, z_i) - \ell(f, z'_i)))_{1 \leqslant i \leqslant n}$ have the same distribution, as well as $(\varepsilon_i \ell(f, z_i))_i$ and $(-\varepsilon_i \ell(f, z'_i))_i$. The last term corresponds to the *Rademacher complexity* of the class of functions $\{\ell(w, \cdot) : \|w\| \leqslant R\}$ [Bartlett and Mendelson (2002); Koltchinskii (2011)]. Now, using that $\ell(w, z_i) = \ell(y_i, \langle w, x_i \rangle)$ for $z_i = (x_i, y_i)$, where $\ell(y_i, \cdot)$ is G-Lipschitz by Assumption 2, Ledoux-Talagrand's contraction inequality for Rademacher averages [Meir and Zhang (2003)] gives

$$\begin{aligned}
\mathbb{E}_{D,\varepsilon} \Big[\sup_{\|w\| \leqslant R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell(w, z_i)\Big] &\leqslant G\mathbb{E}_{D,\varepsilon} \Big[\sup_{\|w\| \leqslant R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle w, x_i \rangle\Big] \\
&= G\mathbb{E}_{D,\varepsilon} \Big[\sup_{\|w\| \leqslant R} \Big\langle w, \frac{1}{n} \sum_{i=1}^n \varepsilon_i x_i \Big\rangle\Big] \\
&\leqslant GR\mathbb{E}_{D,\varepsilon} \Big[\Big\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i x_i \Big\|^2\Big]^{1/2} \\
&= \frac{GR\mathbb{E}[\|x\|^2]^{1/2}}{\sqrt{n}} \\
&\leqslant \frac{GR\kappa}{\sqrt{n}}
\end{aligned}$$

where we used that $\mathbb{E}[\varepsilon_i \varepsilon_j \langle x_i, x_j \rangle] = 0$ for $i \neq j$ by independence, and that $\|x\| \leqslant \kappa$ almost surely (Assumption 1). Hence,

$$\mathbb{E}_{D\sim P^n} \sup_{\|w\| \leqslant R} \Big[L(w) - \widehat{L}(w)\Big] \leqslant \frac{2GR\kappa}{\sqrt{n}}. \tag{30}$$

To write the analogous bound in high probability we apply McDiarmid's inequality [Boucheron et al. (2013)]. We know that given $D := \{z_1, \dots, z_i, \dots, z_n\}$, $D^i = \{z_1, \dots, z'_i, \dots, z_n\}$ and defining $\phi(D) := \sup_{\|w\| \leqslant R} [L(w) - \widehat{L}(w)]$

we have

$$\begin{aligned}\left|\phi(D)-\phi(D^i)\right| &\leqslant \sup_{\|w\|\leqslant R}\left|\frac{1}{n}\ell(w,z_i)-\frac{1}{n}\ell(w,z_i')\right| \\ &\leqslant \frac{G}{n}\sup_{\|w\|\leqslant R}\left|\langle w, x_i - x_i'\rangle\right| \\ &\leqslant \frac{2GR\kappa}{n} \end{aligned} \tag{31}$$

using the Assumption 1 of boundedness of the input. Hence, by McDiarmid inequality:

$$\mathbb{P}\Big[\phi(D)-\mathbb{E}_D[\phi(D)]\geqslant t\Big]\leqslant \exp\Big(\frac{-t^2 n}{2G^2R^2\kappa^2}\Big); \tag{32}$$

taking $\delta = \exp\left(\frac{-t^2n}{2G^2R^2\kappa^2}\right)$ so that $t = GR\kappa\sqrt{\frac{2\log(1/\delta)}{n}}$, we obtain the desired bound (29). □

Lemma 1 suffices to control the excess risk of the constrained risk minimizer $\widehat{w} := \arg\min_{\|w\|\leqslant R} L(w)$ for $R = \|w_*\|$. On the other hand, this result cannot be readily applied to $\widehat{w}_\lambda$, since its norm $\|\widehat{w}_\lambda\|$ is itself random. Observe that, by definition and by Assumption 2,

$$\lambda\|\widehat{w}_\lambda\|^2 \leqslant \widehat{L}_\lambda(\widehat{w}_\lambda) \leqslant \widehat{L}_\lambda(0) = \widehat{L}(0) \leqslant \sup_{y\in\mathcal{Y}}\ell(y,0) = \ell_0,$$

so that $\|\widehat{w}_\lambda\| \leqslant \sqrt{\ell_0/\lambda}$. One could in principle apply this bound on $\widehat{w}_\lambda$, but this would yield a suboptimal dependence on λ and thus a suboptimal rate.

The next step in the proof is to make the bound of Lemma 1 valid for all norms R, so that it can be applied to the random quantity $R = \|\widehat{w}_\lambda\|$. This is done in Lemma 2 below though a union bound.

Lemma 2. *Under Assumptions 1 and 2, with probability $1-\delta$, one has:*

$$\forall w\in\mathcal{H},\qquad L(w)-\widehat{L}(w)\leqslant \frac{4G(1+\kappa\|w\|)}{\sqrt{n}}\Big(1+\sqrt{\log(2+\log_2(1+\kappa\|w\|))+\log(1/\delta)}\Big).$$

Proof of Lemma 2. Fix $\delta\in(0,1)$. For $p\geqslant 1$, let $R_p := \kappa^{-1}2^p$ and $\delta_p = \delta/(p(p+1))$. By Lemma 1, one has for every $p\geqslant 1$,

$$\mathbb{P}\left(\sup_{\|w\|\leqslant R_p}\Big[L(w)-\widehat{L}(w)\Big]\geqslant \frac{G\kappa R_p}{\sqrt{n}}\Big(2+\sqrt{2\log\frac{1}{\delta_p}}\Big)\right)\leqslant\delta_p.$$

Taking a union bound over $p\geqslant 1$ and using that $\sum_{p\geqslant1}\delta_p=\delta$ and $\delta_p\geqslant\delta^2/(p+1)^2$, we get:

$$\mathbb{P}\left(\exists p\geqslant 1,\quad \sup_{\|w\|\leqslant R_p}\Big[L(w)-\widehat{L}(w)\Big]\geqslant \frac{G\kappa R_p}{\sqrt{n}}\Big(2+2\sqrt{\log\frac{p+1}{\delta}}\Big)\right)\leqslant\delta.$$

Now, for $w\in\mathcal{H}$, let $p=\lceil\log_2(1+\kappa\|w\|)\rceil$; then, $1+\kappa\|w\|\leqslant\kappa R_p = 2^p\leqslant 2(1+\kappa\|w\|)$, so $\|w\|\leqslant R_p$. Hence, with probability $1-\delta$,

$$\forall w\in\mathcal{H},\qquad L(w)-\widehat{L}(w)\leqslant\frac{4G(1+\kappa\|w\|)}{\sqrt{n}}\Big(1+\sqrt{\log\big(2+\log_2(1+\kappa\|w\|)\big)+\log(1/\delta)}\Big).$$

This is precisely the desired bound. □

Since the bound of Lemma 2 holds simultaneously for all $w\in\mathcal{H}$, one can apply it to $\widehat{w}_\lambda$; using the inequality $\kappa\|\widehat{w}_\lambda\|\leqslant\kappa\sqrt{\ell_0/\lambda}\leqslant(1+\ell_0\kappa^2/\lambda)/2$ to bound the log log term, this gives with probability $1-\delta$,

$$L(\widehat{w}_\lambda)-\widehat{L}(\widehat{w}_\lambda)\leqslant\frac{4G(1+\kappa\|\widehat{w}_\lambda\|)}{\sqrt{n}}\Big(1+\sqrt{\log\big(1+\log_2(3+\ell_0\kappa^2/\lambda)\big)+\log(1/\delta)}\Big). \tag{33}$$

Now, let $C = C_{\lambda,\delta} = 4\{1 + \sqrt{\log(1 + \log_2(3 + \ell_0\kappa^2/\lambda)) + \log(1/\delta)}\}$; (33) writes $L(\widehat{w}_\lambda) - \widehat{L}(\widehat{w}_\lambda) \leqslant CG(1 + \kappa\|\widehat{w}_\lambda\|)/\sqrt{n}$. Using that $ab \leqslant \lambda a^2 + b^2/(4\lambda)$ for $a, b \geqslant 0$, one can then write

$$\begin{aligned} L(\widehat{w}_\lambda) &\leqslant \widehat{L}(\widehat{w}_\lambda) + \frac{CG\kappa\|\widehat{w}_\lambda\|}{\sqrt{n}} + \frac{CG}{\sqrt{n}} \\ &\leqslant \widehat{L}(\widehat{w}_\lambda) + \lambda\|\widehat{w}_\lambda\|^2 + \frac{C^2G^2\kappa^2}{4\lambda n} + \frac{CG}{\sqrt{n}} \qquad (34) \\ &\leqslant \widehat{L}(w_\lambda) + \lambda\|w_\lambda\|^2 + \frac{C^2G^2\kappa^2}{4\lambda n} + \frac{CG}{\sqrt{n}} \qquad (35) \end{aligned}$$

where (35) holds by definition of $\widehat{w}_\lambda$. Now, since $|\ell(w_\lambda, Z)| \leqslant |\ell(Y, 0)| + |\ell(Y, \langle w_\lambda, X\rangle) - \ell(Y, 0)| \leqslant \ell_0 + G\kappa\|w_\lambda\|$ almost surely, Hoeffding's inequality [Boucheron et al. (2013)] implies that, with probability $1 - \delta$,

$$\widehat{L}(w_\lambda) < L(w_\lambda) + (\ell_0 + G\kappa\|w_\lambda\|)\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Combining this inequality with (35) with a union bound, with probability $1 - 2\delta$:

$$L(\widehat{w}_\lambda) < L(w_\lambda) + \lambda\|w_\lambda\|^2 + \frac{C^2G^2\kappa^2}{4\lambda n} + \frac{GC}{\sqrt{n}} + (\ell_0 + G\kappa\|w_\lambda\|)\sqrt{\frac{2\log(1/\delta)}{n}}. \qquad (36)$$

First case: w_* exists. First, assume that $w_* = \arg\min_{w\in\mathcal{H}} L(w)$ exists. Then, by definition of w_λ, $L(w_\lambda) + \lambda\|w_\lambda\|^2 \leqslant L(w_*) + \lambda\|w_*\|^2$. In addition, $\|w_\lambda\| \leqslant \|w_*\|$, since otherwise $\|w_*\| < \|w_\lambda$ and $L(w_*) \leqslant L(w_\lambda)$ would imply $L(w_*) + \lambda\|w_*\|^2 < L(w_\lambda) + \lambda\|w_\lambda\|^2$, contradicting the above inequality. Since $L(w_*) = \inf_{\mathcal{H}} L$, it follows that, with probability $1 - 2\delta$,

$$\begin{aligned} L(\widehat{w}_\lambda) &< \inf_{\mathcal{H}} L + \lambda\|w_*\|^2 + \frac{C^2G^2\kappa^2}{4\lambda n} + \frac{GC}{\sqrt{n}} + (\ell_0 + G\kappa\|w_*\|)\sqrt{\frac{2\log(1/\delta)}{n}} \\ &\leqslant \inf_{\mathcal{H}} L + \lambda\|w_*\|^2 + \frac{8G^2\kappa^2\{1 + \log(1 + \log_2(3 + \ell_0\kappa^2/\lambda)) + \log(1/\delta)\}}{\lambda n} + \\ &+ \frac{4G\{1 + \sqrt{\log(1 + \log_2(3 + \ell_0\kappa^2/\lambda)) + \log(1/\delta)}\}}{\sqrt{n}} + (\ell_0 + G\kappa\|w_*\|)\sqrt{\frac{2\log(1/\delta)}{n}} \\ &= \inf_{\mathcal{H}} L + O\Big(\lambda\|w_*\|^2 + \frac{G^2\kappa^2\{\log\log(3 + \ell_0\kappa^2/\lambda) + \log(1/\delta)\}}{\lambda n} + \frac{(G + \ell_0)\sqrt{\log(1/\delta)}}{\sqrt{n}}\Big), \qquad (37) \end{aligned}$$

where the $O(\dots)$ hide universal constants. The bound (37) precisely corresponds to the desired bound (27) after replacing δ by $\delta/2$. In particular, tuning $\lambda \asymp (G\kappa/\|w_*\|)\sqrt{\log(1/\delta)/n}$ yields

$$L(\widehat{w}_\lambda) - \inf_{\mathcal{H}} L \lesssim \frac{\{\ell_0 + G(1 + \kappa\|w_*\|)\}\{\log\log(\kappa\|w_*\|n/G) + \sqrt{\log(1/\delta)}\}}{\sqrt{n}}.$$

Omitting the $\log\log n$ term, this bound essentially scales as $\widetilde{O}(G\kappa\|w_*\|\sqrt{\log(1/\delta)/n})$.

General case. Let us now drop the assumption that $w_* = \arg\min_{w\in\mathcal{H}} L(w)$ exists, and let (see (24)) for $\lambda > 0$:

$$\begin{aligned} \mathcal{A}(\lambda) &= L(w_\lambda) + \lambda\|w_\lambda\|^2 - \inf_{\mathcal{H}} L \\ &= \inf_{w\in\mathcal{H}} [L(w) + \lambda\|w\|^2] - \inf_{\mathcal{H}} L\,. \end{aligned}$$

Note that, again using that $ab \leqslant \lambda a^2 + b^2/(4\lambda)$,

$$\begin{aligned} G\kappa\|w_\lambda\|\sqrt{\frac{2\log(1/\delta)}{n}} &\leqslant \lambda\|w_\lambda\|^2 + \frac{2G^2\kappa^2\log(1/\delta)}{\lambda n} \\ &\leqslant \mathcal{A}(\lambda) + \frac{2G^2\kappa^2\log(1/\delta)}{\lambda n} \end{aligned}$$

so that (36) implies, with probability $1-2\delta$,

$$L(\widehat{w}_\lambda) - \inf_{\mathcal{H}} L < 2\mathcal{A}(\lambda) + \frac{C^2G^2\kappa^2}{4\lambda n} + \frac{GC}{\sqrt{n}} + \ell_0\sqrt{\frac{2\log(1/\delta)}{n}} + \frac{2G^2\kappa^2\log(1/\delta)}{\lambda n}.$$

Finally, note that for all $w \in \mathcal{H}$ with $\|w\| \leqslant R$, $\mathcal{A}(\lambda) \leqslant L(w) + \lambda\|w\|^2 - \inf_{\mathcal{H}} L \leqslant L(w) - \inf_{\mathcal{H}} L + \lambda R^2$, hence $\mathcal{A}(\lambda) \leqslant \inf_{\|w\|\leqslant R} L(w) - \inf_{\mathcal{H}} L + \lambda R^2$ and

$$L(\widehat{w}_\lambda) - \inf_{\mathcal{H}} L < 2\Big(\inf_{\|w\|\leqslant R} L(w) - \inf_{\mathcal{H}} L\Big) + 2\lambda R^2 + \frac{C^2G^2\kappa^2 + 8G^2\kappa^2\log(1/\delta)}{4\lambda n} + \frac{GC + \ell_0\sqrt{2\log(1/\delta)}}{\sqrt{n}}.$$

Letting $\lambda \asymp 1/(R\sqrt{n})$, this gives $L(\widehat{w}_\lambda) - \inf_{\mathcal{H}} L \leqslant 2(\inf_{\|w\|\leqslant R} L(w) - \inf_{\mathcal{H}} L) + O(R/\sqrt{n})$ with high probability.

C Proof of Theorem 2

The proof of Theorem 2 is given by decomposing the excess risk as in (44) where $\mathcal{P}_m$ is replaced by $\mathcal{P}_\mathcal{B}$, (47) bounds term A, (48) bounds term B and (49) and the Definition 14 bound term C.

D T-approximate leverage scores and proof of Proposition 1

Since in practice the leverage scores $l_i(\alpha)$ defined by (10) are onerous to compute, approximations $(\hat{l}_i(\alpha))_{i=1}^n$ have been considered [Drineas et al. (2012); Cohen et al. (2015); Alaoui and Mahoney (2015)]. In particular, in the following we are interested in suitable approximations defined as follows.

Definition 2. (T-approximate leverage scores) Let $(l_i(\alpha))_{i=1}^n$ be the leverage scores associated to the training set for a given α. Let $\delta > 0$, $t_0 > 0$ and $T \geqslant 1$. We say that $(\hat{l}_i(\alpha))_{i=1}^n$ are T-approximate leverage scores with confidence δ, when with probability at least $1-\delta$,

$$\frac{1}{T} l_i(\alpha) \leqslant \hat{l}_i(\alpha) \leqslant T l_i(\alpha), \qquad \forall i \in \{1,\dots,n\}, \quad \alpha \geqslant t_0 \tag{38}$$

So, given T-approximate leverage score for $\alpha \geqslant t_0$, $\{\tilde{x}_1,\dots,\tilde{x}_m\}$ are sampled from the training set independently with replacement, and with probability to be selected given by $Q_\alpha(i) = \hat{l}_i(\alpha)/\sum_j \hat{l}_j(\alpha)$.

First part of Proposition 1 is the content of the following two results from Rudi et al. (2015).

Lemma 3 (Uniform sampling, Lemma 6 in Rudi et al. (2015))**.** *Under Assumption 1, let J be a partition of $\{1,\dots,n\}$ chosen uniformly at random from the partitions of cardinality m. Let $\alpha > 0$, for any $\delta > 0$, such that $m \geqslant 67\log\frac{4\kappa^2}{\alpha\delta} \vee 5d_{\alpha,\infty}\log\frac{4\kappa^2}{\alpha\delta}$, the following holds with probability at least $1-\delta$*

$$\left\|(I - \mathcal{P}_{\mathcal{B}_m})\Sigma^{1/2}\right\|^2 \leqslant 3\alpha \tag{39}$$

Lemma 4 (ALS sampling, Lemma 7 in Rudi et al. (2015))**.** *Let $(\hat{l}_i(t))_{i=1}^n$ be the collection of approximate leverage scores. Let $\alpha > 0$ and the sampling probability Q_α be defined as $Q_\alpha(i) = \hat{l}_i(\alpha)/\sum_{j\in N}\hat{l}_j(\alpha)$ for any $i \in N$ with $N = \{1,\dots,n\}$. Let $\mathcal{I} = (i_1,\dots,i_m)$ be a collection of indices independently sampled with replacement from N according to the probability distribution P_α. Let $\mathcal{B}_m = span\{x_j | j \in J\}$ where J be the subcollection of $\mathcal{I}$ with all the duplicates removed. Under Assumption 1, for any $\delta > 0$ the following holds with probability at least $1-\delta$*

$$\left\|(I - \mathcal{P}_{\mathcal{B}_m})\Sigma_\alpha^{1/2}\right\|^2 \leqslant 3\alpha \tag{40}$$

where the following conditions are satisfied:

1. *there exists a* $T \geqslant 1$ *and a* $t_0 > 0$ *such that* $(l_i(t))_{i=1}^n$ *are* T*-approximate leverage scores for any* $t \geqslant t_0$,
2. $n \geqslant 1655\kappa^2 + 223\kappa^2 \log \frac{4\kappa^2}{\delta}$,
3. $t_0 \vee \frac{19\kappa^2}{n} \log \frac{4n}{\delta} \leqslant \alpha \leqslant \|\Sigma\|$,
4. $m \geqslant 334 \log \frac{16n}{\delta} \vee 78 T^2 d_{\alpha,2} \log \frac{16n}{\delta}$.

If the spectrum of Σ satisfies the decay property (15), the second part of Proposition 1 is a consequence of Lemma 4.

E Proof of Theorem 3

Theorem 3 is a compact version of the following result.

Theorem 7. *Fix* $\alpha, \lambda, \delta > 0$. *Under Assumption 1, 2 and 3, with probability at least* $1 - \delta$*:*

$$L(\widehat{\beta}_\lambda) - L(w_*) \leqslant \frac{C_{\lambda,\delta}^2 G^2 \kappa^2}{4\lambda n} + \frac{C_{\lambda,\delta} G}{\sqrt{n}} + G\kappa \|w_*\| \sqrt{\frac{2\log(3/\delta)}{n}} + 2G\sqrt{\alpha}\|w_*\| + \lambda \|w_*\|_{\mathcal{H}}^2 \tag{41}$$
$$C_{\lambda,\delta} = 4\Big\{1 + \sqrt{\log(1 + \log_2(3 + \ell_0 \kappa^2/\lambda)) + \log(1/\delta)}\Big\}$$

provided that $n \geqslant 1655\kappa^2 + 223\kappa^2 \log \frac{4\kappa^2}{\delta}$ *and*

1. *for uniform sampling*
$$m \geqslant 67 \log \frac{4\kappa^2}{\alpha\delta} \vee 5 d_{\alpha,\infty} \log \frac{4\kappa^2}{\alpha\delta} \tag{42}$$
2. *for ALS sampling and* T*-approximate leverage scores with subsampling probabilities* Q_α, $t_0 > \frac{19\kappa^2}{n} \log \frac{4n}{\delta}$ *and*
$$m \geqslant 334 \log \frac{16n}{\delta} \vee 78 T^2 d_{\alpha,2} \log \frac{16n}{\delta} \tag{43}$$
where $\alpha \geqslant \frac{19\kappa^2}{n} \log \frac{4n}{\delta}$

Proof. We recall the notation.

$$\{\tilde{x}_1, \dots, \tilde{x}_m\} \subseteq \{x_1, \dots, x_n\}$$
$$\mathcal{B}_m = \operatorname{span}\{\tilde{x}_1, \dots, \tilde{x}_m\}$$
$$\widehat{\beta}_\lambda = \operatorname*{arg\,min}_{w \in \mathcal{B}_m} \widehat{L}(w)$$
$$w_* = \operatorname*{arg\,min}_{w \in \mathcal{H}} \widehat{L}_\lambda(w).$$

and $\mathcal{P}_m = \mathcal{P}_{\mathcal{B}_m}$ is the orthogonal projector operator onto $\mathcal{B}_m$.
In order to bound the excess risk of $\widehat{\beta}_\lambda$, we decompose the error as follows:

$$L(\widehat{\beta}_\lambda) - L(w_*) = \underbrace{L(\widehat{\beta}_\lambda) - \widehat{L}(\widehat{\beta}_\lambda) - \lambda\|\widehat{\beta}_\lambda\|_{\mathcal{H}}^2}_{\mathbf{A}} + \underbrace{\widehat{L}(\widehat{\beta}_\lambda) + \lambda\|\widehat{\beta}_\lambda\|_{\mathcal{H}}^2 - \widehat{L}(\mathcal{P}_m w_*) - \lambda\|\mathcal{P}_m w_*\|_{\mathcal{H}}^2}_{\leqslant 0} +$$
$$+ \underbrace{\widehat{L}(\mathcal{P}_m w_*) - L(\mathcal{P}_m w_*)}_{\mathbf{B}} + \underbrace{L(\mathcal{P}_m w_*) - L(w_*)}_{\mathbf{C}} + \lambda\|\mathcal{P}_m w_*\|_{\mathcal{H}}^2 \tag{44}$$

Bound for term A
To bound term **A** we apply Lemma 2 for $\widehat{\beta}_\lambda$ and we get with probability a least $1 - \delta$

$$\forall \lambda \geqslant \frac{\ell_0 \kappa^2}{n^{2K}}, \qquad L(\widehat{\beta}_\lambda) \leqslant \widehat{L}(\widehat{\beta}_\lambda) + \frac{C_{\lambda,\delta} G (1 + \kappa\|\widehat{\beta}_\lambda\|)}{\sqrt{n}} \tag{45}$$

where $C_{\lambda,\delta} = 4\big\{1 + \sqrt{\log(1 + \log_2(3 + \ell_0\kappa^2/\lambda)) + \log(1/\delta)}\big\}$. Now since $xy \leqslant \lambda x^2 + y^2/(4\lambda)$, we can write

$$\frac{C_{\lambda,\delta} G\kappa \|\widehat{\beta}_\lambda\|}{\sqrt{n}} \leqslant \lambda \|\widehat{\beta}_\lambda\|^2 + \frac{C^2_{\lambda,\delta} G^2\kappa^2}{4\lambda n} \tag{46}$$

hence,

$$L(\widehat{\beta}_\lambda) \leqslant \widehat{L}(\widehat{\beta}_\lambda) + \lambda \|\widehat{\beta}_\lambda\|^2 + \frac{C^2_{\lambda,\delta} G^2\kappa^2}{4\lambda n} + \frac{C_{\lambda,\delta} G}{\sqrt{n}} \tag{47}$$

Bound for term B
As regards term **B**, since $|\ell(\mathcal{P}_m w_*, Z) - \ell(0, Z)| \leqslant G\kappa \|\mathcal{P}_m w_*\| \leqslant G\kappa \|w_*\|$, using Hoeffding's inequality, we have with probability at least $1 - \delta$

$$\mathbf{B} \leqslant \left|\widehat{L}(\mathcal{P}_m w_*) - L(\mathcal{P}_m w_*)\right| \leqslant G\kappa \|w_*\| \sqrt{\frac{2\log(1/\delta)}{n}} \tag{48}$$

Bound for term C
Finally, term **C** can be rewritten as

$$\begin{aligned} \mathbf{C} &= L(\mathcal{P}_m w_*) - L(w_*) \\ &\leqslant G \|\Sigma^{1/2}(I - \mathcal{P}_m) w_*\|_{\mathcal{H}} \\ &\leqslant G \|\Sigma^{1/2}(I - \mathcal{P}_m)\| \|w_*\|_{\mathcal{H}} \end{aligned} \tag{49}$$

We bound equation (49) using Lemma 3 for uniform sampling and Lemma 4 for ALS selection.
Putting the pieces together and noticing that $\lambda \|\mathcal{P}_m w_*\|^2_{\mathcal{H}} \leqslant \lambda \|w_*\|^2_{\mathcal{H}}$ we finally get the result in Theorem 7. □

The following corollary shows that there is choice of the parameters $\lambda = \lambda_n, \alpha = \alpha_n$ such that the excess risk of the β_{λ_n} converges to zero with the optimal rate (up to a logarithmic factor) $O(\log(n/\delta)/\sqrt{n})$.

Corollary 1. *Fix $\delta > 0$. Under the assumption of Theorem (7), let*

$$\lambda \asymp \frac{1}{\|w_*\|} n^{-1/2} \qquad \alpha \asymp \frac{\log(n/\delta)}{n}$$

with probability at least $1 - \delta$:

$$L(\widehat{\beta}_\lambda) - L(w_*) \lesssim \frac{\|w_*\| \sqrt{\log(n/\delta)}}{\sqrt{n}} \tag{50}$$

Despite of the fact that the rate is optimal (up to the logarithmic term), the required number of subsampled points is $m \gtrsim n \log n$, so that the procedure is not effective. However, the following proposition shows that under a fast decay for the spectrum of the covariance operator Σ, the ALS method becomes computationally efficient. We denote by $(\sigma_i(\Sigma))_I$ the sequence of strictly positive eigenvalues of Σ where the eigenvalues are counted with respect to their multiplicity and ordered in a non-increasing way.

Proposition 2. *Fix $\delta > 0$. Under the assumptions of Theorem (7) and using ALS sampling*

1. *for polynomial decay, i.e. $\sigma_i(\Sigma) \leqslant \gamma i^{-\frac{1}{p}}$, $\gamma \in \mathbb{R}^+$, $p \in (0, 1)$, for $\delta > 0$, with probability at least $1 - \delta$:*

$$L(\widehat{\beta}_\lambda) - L(w_*) \leqslant \frac{C^2_{\lambda,\delta} G^2\kappa^2}{4\lambda n} + \frac{C_{\lambda,\delta} G}{\sqrt{n}} + G\kappa \|w_*\| \sqrt{\frac{2\log(3/\delta)}{n}} + 2G\|w_*\|\sqrt{\alpha} + \lambda \|w_*\|^2_{\mathcal{H}} \tag{51}$$

where $O(\sqrt{\frac{\log(n/\delta)}{n}})$ rate can be achieved optimizing the choice of the parameters, i.e. $\lambda \asymp \frac{1}{\|w_*\|} n^{-1/2}$, $\alpha \asymp \frac{\log(n/\delta)}{n}$, $m \gtrsim n^p (\log n)^{1-p}$.

2. *for exponential decay, i.e.* $\sigma_i(\Sigma) \leqslant \gamma e^{-\beta i}$, $\gamma, \beta \in \mathbb{R}^+$, *for* $\delta > 0$, *with probability at least* $1-\delta$:

$$L(\widehat{\beta}_\lambda) - L(w_*) \leqslant \frac{C_{\lambda,\delta}^2 G^2 \kappa^2}{4\lambda n} + \frac{C_{\lambda,\delta} G}{\sqrt{n}} + G\kappa\|w_*\|\sqrt{\frac{2\log(3/\delta)}{n}} + 2G\|w_*\|\sqrt{\alpha} + \lambda\|w_*\|_{\mathcal{H}}^2 \tag{52}$$

where $O(\sqrt{\frac{\log(n/\delta)}{n}})$ *rate can be achieved optimizing the choice of the parameter, i.e.* $\lambda \asymp \frac{1}{\|w_*\|} n^{-1/2}$, $\alpha \asymp \frac{\log(n/\delta)}{n}$, $m \gtrsim \log^2 n$.

Proof. The claim is a consequence of Appendix H where the link with m is obtained using Leverage Score sampling so that in Lemma 4 using proposition 4 we have that

$$m \gtrsim d_{\alpha,2} \log n, \qquad d_{\alpha,2} \lesssim \alpha^{-p}, \qquad m \asymp n^p (\log n)^{1-p} \tag{53}$$

while using Proposition 5 we have that

$$m \gtrsim d_{\alpha,2} \log n, \qquad d_{\alpha,2} \lesssim \log(1/\alpha), \qquad m \asymp \log^2 n \tag{54}$$

□

From proposition above we have the following asymptotic rate.

Corollary 2. *Fix* $\delta > 0$. *Under the assumptions of Theorem (7) and using ALS sampling, with probability at least* $1-\delta$

1. *assuming polynomial decay of the spectrum of* Σ *and choosing* $\lambda \asymp \frac{1}{\|w_*\|} n^{-1/2}$, $m \gtrsim n^p (\log n)^{1-p}$ *then:*

$$L(\widehat{\beta}_\lambda) - L(w_*) \lesssim \frac{\|w_*\|\sqrt{\log(n/\delta)}}{\sqrt{n}} \tag{55}$$

2. *assuming exponential decay of the spectrum of* Σ *and choosing* $\lambda \asymp \frac{1}{\|w_*\|} n^{-1/2}$, $m \gtrsim \log^2 n$ *then:*

$$L(\widehat{\beta}_\lambda) - L(w_*) \lesssim \frac{\|w_*\|\sqrt{\log(n/\delta)}}{\sqrt{n}} \tag{56}$$

F Proof of Theorem 4

Before proving Theorem 4 we introduce a modification of the above Lemma 4 in the case of sub-gaussian random variables

Lemma 5. *(ALS sampling for sub-gaussian variables). Let* $(\hat{l}_i(t))_{i=1}^n$ *be the collection of approximate leverage scores. Let* $\alpha > 0$ *and the sampling probability* Q_α *be defined as* $Q_\alpha(i) = \hat{l}_i(\alpha) / \sum_{j \in N} \hat{l}_j(\alpha)$ *for any* $i \in N$ *with* $N = \{1, ..., n\}$. *Let* $\mathcal{I} = (i_1, ..., i_m)$ *be a collection of indices independently sampled with replacement from* N *according to the probability distribution* P_α. *Let* $\mathcal{B}_m = span\{x_j | j \in J\}$ *where* J *be the subcollection of* $\mathcal{I}$ *with all the duplicates removed. Under Assumption 4, for any* $\delta > 0$ *the following holds with probability at least* $1 - 5\delta$

$$\left\| (I - \mathcal{P}_{\mathcal{B}_m}) \Sigma_\alpha^{1/2} \right\|^2 \lesssim \alpha \tag{57}$$

when the following conditions are satisfied:

1. *there exists a* $T \geqslant 1$ *and a* $t_0 > 0$ *such that* $(l_i(t))_{i=1}^n$ *are* T*-approximate leverage scores for any* $t \geqslant t_0$,

2.
$$n \gtrsim d_{\alpha,2}(\Sigma) \vee \log(1/\delta) \tag{58}$$

3.
$$m \gtrsim d_{\alpha,2}(\Sigma) \log(\frac{2n}{\delta}) \tag{59}$$

Proof. The proof follows the structure of the one in Lemma 4 (see Rudi et al. (2015)). Exploiting sub-gaussianity anyway the various terms are bounded differently. In particular, to bound β_1 we refer to Theorem 9 in Koltchinskii and Lounici (2014), obtaining with probability at least $1-\delta$

$$\beta_1(\alpha) \lesssim \max\left\{\sqrt{\frac{d_{\alpha,2}(\Sigma)}{n}}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}. \tag{60}$$

As regards β_3 term we apply Proposition 3 below to get with probability greater than $1-3\delta$

$$\beta_3(\alpha) \leqslant \frac{2\log\frac{2n}{\delta}}{3m} + \sqrt{\frac{32T^2 d_{\alpha,2}(\Sigma)\log\frac{2n}{\delta}}{m}}$$

for $n \geqslant 2C^2\log(1/\delta)$.
Finally, taking a union bound we have with probability at least $1-5\delta$

$$\begin{aligned}\beta(\alpha) \lesssim & \max\left\{\sqrt{\frac{d_{\alpha,2}(\Sigma)}{n}}, \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right\} + \\ & + \left(1+\max\left\{\sqrt{\frac{d_{\alpha,2}(\Sigma)}{n}}, \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right\}\right)\left(\frac{2\log\frac{2n}{\delta}}{3m} + \sqrt{\frac{32T^2 d_{\alpha,2}(\Sigma)\log\frac{2n}{\delta}}{m}}\right) \lesssim 1\end{aligned}$$

when $n \gtrsim d_{\alpha,2}(\Sigma) \vee \log(1/\delta)$ and $m \gtrsim d_{\alpha,2}(\Sigma)\log\frac{2n}{\delta}$. See Rudi et al. (2015) to conclude the proof. □

Corollary 3. *Given the assumptions in Theorem 5 if we further assume a polynomial decay of the spectrum of Σ with rate $1/p \in (0,\infty)$, for any $\delta > 0$ the following holds with probability $1-\delta$*

$$\left\|(I - \mathcal{P}_{\mathcal{B}_m})\,\Sigma_\alpha^{1/2}\right\|^2 \lesssim \alpha$$

when the following conditions are satisfied:

1. *there exists a $T \geqslant 1$ and a $t_0 > 0$ such that $(l_i(t))_{i=1}^n$ are T-approximate leverage scores for any $t \geqslant t_0$,*
2. $$n \gtrsim \log(5/\delta) \tag{61}$$
3. $$\alpha \gtrsim n^{-1/p} \tag{62}$$
4. $$m \gtrsim \alpha^{-p}\log(\frac{10n}{\delta}) \tag{63}$$

Proof. Using Proposition 4 $d_{\alpha,2}(\Sigma) \lesssim \alpha^{-p}$, the result simply follows from the substitution in Lemma 5. □

Proposition 3. *Let $X, X_1, \dots, X_n$ be iid C-sub-gaussian random variables in $\mathcal{H}$. Let $d_{\alpha,2}(\widehat{\Sigma}) = \mathrm{Tr}(\widehat{\Sigma}_\alpha^{-1}\widehat{\Sigma})$ the empirical effective dimension and $d_{\alpha,2}(\Sigma) = \mathrm{Tr}(\Sigma_\alpha^{-1}\Sigma)$ the correspondent population quantity. For any $\delta > 0$ and $n \geqslant 2C^2\log(1/\delta)$, then the following hold with probability $1-\delta$*

$$d_{\alpha,2}(\widehat{\Sigma}) \leqslant 16 d_{\alpha,2}(\Sigma) \tag{64}$$

Proof. Let V_α be the space spanned by eigenvectors of Σ with corresponding eigenvalues $\alpha_j \geqslant \alpha$, and call D_α its dimension. Notice that $D_\alpha \leqslant 2d_{\alpha,2}(\Sigma)$ since $d_{\alpha,2}(\Sigma) = \mathrm{Tr}(\Sigma_\alpha^{-1}\Sigma) = \sum \frac{\alpha_i}{\alpha_i+\alpha}$, where in the sum we have D_α terms greater or equal than $1/2$.
Let $X = X_1 + X_2$, where X_1 is the orthogonal projection of X on the space V_α, we have

$$\widehat{\Sigma} = \widehat{\Sigma}_1 + \widehat{\Sigma}_2 + n^{-1}\sum_{i=1}^n (X_{1,i}X_{2,i}^\top + X_{2,i}X_{1,i}^\top) \preccurlyeq 2(\widehat{\Sigma}_1 + \widehat{\Sigma}_2) \tag{65}$$

Now, since the function $g : t \mapsto \frac{t}{t+\alpha}$ is sub-additive (meaning that $g(t+t') \leqslant g(t) + g(t')$), denoting $d_{\alpha,2}(\Sigma) = \operatorname{Tr} g(\Sigma) = \operatorname{Tr}(\Sigma_\alpha^{-1}\Sigma)$,

$$d_{\alpha,2}(\widehat{\Sigma}) \leqslant 2(d_{\alpha,2}(\widehat{\Sigma}_1) + d_{\alpha,2}(\widehat{\Sigma}_2)) \tag{66}$$

and, since $(\widehat{\Sigma}_1 + \alpha)^{-1}\widehat{\Sigma}_1 \preccurlyeq I_{V_\alpha}$,

$$\operatorname{Tr}(\widehat{\Sigma}_\alpha^{-1}\widehat{\Sigma}) \leqslant 2D_\alpha + \frac{2\operatorname{Tr}(\widehat{\Sigma}_2)}{\alpha} = 4d_{\alpha,2}(\Sigma) + \frac{2\operatorname{Tr}(\widehat{\Sigma}_2)}{\alpha} \tag{67}$$

Now,

$$\operatorname{Tr}(\widehat{\Sigma}_2) = \frac{1}{n}\sum_{i=1}^n \|X_{2,i}\|^2$$

It thus suffices establish concentration for averages of the random variable $\|X_2\|^2$.
Since X is sub-gaussian then $\|X_2\|^2$ is sub-exponential. In fact, since X is C-sub-gaussian then

$$\|\langle v, X\rangle\|_{\psi_2} \leqslant C\|\langle v, X\rangle\|_{L_2} \qquad \forall v \in \mathcal{H} \tag{68}$$

and given that $\langle v, \mathcal{P}X\rangle = \langle \mathcal{P}v, X\rangle$ with $\mathcal{P}$ an orthogonal projection, then also X_2 is C-sub-gaussian. Now take e_i the orthonormal basis of V composed by the eigenvectors of $\Sigma_2 = \mathbb{E}[X_2 X_2^T]$, then

$$\left\|\|X_2\|^2\right\|_{\psi_1} = \left\|\sum_i \langle X_2, e_i\rangle^2\right\|_{\psi_1} \leqslant \sum_i \left\|\langle X_2, e_i\rangle^2\right\|_{\psi_1} \tag{69}$$

$$= \sum_i \|\langle X_2, e_i\rangle\|_{\psi_2}^2 \leqslant C^2 \|\langle X_2, e_i\rangle\|_{L_2}^2 \tag{70}$$

$$= C^2 \sum_i \alpha_i = C^2 \operatorname{Tr}[\Sigma_2] = C^2 \mathbb{E}\left[\|X_2\|^2\right] \tag{71}$$

so $\|X_2\|^2$ is $C^2\mathbb{E}\left[\|X_2\|^2\right]$-sub-exponential. Note that $\mathbb{E}\|X_2\|^2 = \mathbb{E}[\operatorname{Tr}(X_2 X_2^\top)] = \operatorname{Tr}(\Sigma_2) \leqslant 2\alpha d_{\alpha,2}(\Sigma)$, in fact

$$d_{\alpha,2}(\Sigma) = \sum_{i=1}^\infty \frac{\alpha_i}{\alpha_i + \alpha} \geqslant \sum_{i:\alpha_i<\alpha} \frac{\alpha_i}{\alpha_i + \alpha} \geqslant \sum_{i:\alpha_i<\alpha} \frac{\alpha_i}{2\alpha} = \frac{\operatorname{Tr}(\Sigma_2)}{2\alpha} \tag{72}$$

Hence, we can apply then Bernstein inequality for sub-exponential scalar variables (see Theorem 2.10 in Boucheron et al. (2013)), with parameters ν and c given by

$$n\mathbb{E}\left[\|X_2\|^4\right] \leqslant \underbrace{4nC^2\alpha^2 d_{\alpha,2}^2(\Sigma)}_{\nu} \tag{73}$$

$$c = C\alpha d_{\alpha,2}(\Sigma) \tag{74}$$

where we used the bound on the moments of a sub-exponential variable (see Vershynin (2010)).
With high probability (67) becomes

$$d_{\alpha,2}(\widehat{\Sigma}) \leqslant 8d_{\alpha,2}(\Sigma) + \frac{4Cd_{\alpha,2}(\Sigma)\sqrt{2\log(1/\delta)}}{\sqrt{n}} + \frac{2Cd_{\alpha,2}(\Sigma)\log(1/\delta)}{n} \leqslant 16 d_{\alpha,2}(\Sigma) \tag{75}$$

for $n \geqslant 2C^2\log(1/\delta)$ □

In the following we will exploit the adaptation of Theorem 7.23 in Steinwart and Christmann (2008) for X sub-gaussian, before presenting it we introduce some of the required quantities as defined in Steinwart and Christmann (2008):

$$r^* := \inf_{f\in\mathcal{H}} \Upsilon(f) + L(f^{cl}) - L(f_*)$$

$$\mathcal{H}_r := \left\{f \in \mathcal{H} : \Upsilon(f) + L(f^{cl}) - L(f_*) \leqslant r\right\} \qquad r > r^*$$

$$\mathcal{F}_r := \left\{\ell \circ f^{cl} - \ell \circ f_* : f \in \mathcal{H}_r\right\} \qquad r > r^*$$

$$h_{f_0}(X) := \ell(Y, f_0(X)) - \ell(Y, f_*(X)).$$

Theorem 8 (Adaptation Theorem 7.20 in Steinwart and Christmann (2008) to sub-gaussian framework)**.** *Let $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a continuous loss that can be clipped at $M > 0$ and that satisfies (25) for a constant $B > 0$. Moreover, let $\mathcal{H} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete, separable metric dominating the pointwise convergence, and let $\Upsilon : \mathcal{H} \to [0, \infty)$ be a continuous function. Given a distribution P on $\mathcal{H} \times Y$ that satisfies the variance bound (26). Assume that for fixed $n \geqslant 1$ there exists a $\varphi_n : [0, \infty) \to [0, \infty)$ such that $\varphi_n(4r) \leqslant 2\varphi_n(r)$ and the expectation with respect to the empirical distribution of the empirical Rademacher averages of $\mathcal{F}_r$ can be upper bounded by*

$$\mathbb{E}_{\widehat{P}} \widehat{\mathrm{Rad}}\left(\mathcal{F}_r, n\right) \leqslant \varphi_n(r)$$

for all $r > r^*$. Finally, fix an $f_0 \in \mathcal{H}$ such that $h_{f_0}(X) - h_{f_0^{cl}}(X)$ is a real c-sub-gaussian random variable and $\mathbb{E}(h_{f_0}(X) - h_{f_0^{cl}}(X))^2 \leqslant D\mathbb{E}\left(h_{f_0}(X) - h_{f_0^{cl}}(X)\right)$ for some $D > 0$. Then, for all fixed $\epsilon \geqslant 0, \tau > 0$, and $r > 0$ satisfying

$$r > \max\left\{30\varphi_n(r), \left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}}, \frac{D\tau}{2n} + \frac{c\tau}{n} + \frac{4B\tau}{3n}, r^*\right\}$$

every measurable ϵ-CR-ERM (ϵ-approximate clipped regularized empirical risk minimization) $\widehat{f}$ satisfies

$$\Upsilon(\widehat{f}) + L(\widehat{f}^{cl}) - L(f_*) \leqslant 6\left(\Upsilon\left(f_0\right) + L\left(f_0\right) - L(f_*)\right) + 3r + 3\epsilon$$

with probability not less than $1 - 3e^{-\tau}$.

Proof. The proof mimics the one in Steinwart and Christmann (2008). Clearly Bernstein inequality for bounded variables must be replaced with its sub-gaussian version. Let $\eta := h_{f_0}(X) - h_{f_0^{cl}}(X)$, which is a c-sub-gaussian scalar variable by hypothesis, and define $\xi = \eta - \mathbb{E}[\eta]$. We can apply then Bernstein inequality for sub-gaussian i.i.d variables $\xi, \xi_1, ..., \xi_n$ (see Theorem 2.10 in Boucheron et al. (2013)):

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \leqslant \frac{\sqrt{2\nu\tau}}{n} + \frac{c\tau}{n}\right) \geqslant 1 - e^{-\tau} \tag{76}$$

with $\sum_{i=1}^n \mathbb{E}[\xi_i^2] \leqslant \sum_{i=1}^n \mathbb{E}[\eta_i^2] \leqslant Dn\mathbb{E}[\eta] = \nu$ for hypothesis, so that with probability at least $1 - e^{-\tau}$

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \leqslant \sqrt{\frac{2D\mathbb{E}[\eta]\tau}{n}} + \frac{c\tau}{n} \leqslant \mathbb{E}[\eta] + \frac{D\tau}{2n} + \frac{c\tau}{n} \tag{77}$$

which replaces eq. (7.41) in Steinwart and Christmann (2008). Following the proof in Steinwart and Christmann (2008) while taking into account the above modification leads to the assertion. □

Theorem 9 (Adaptation Theorem 7.23 in Steinwart and Christmann (2008) to sub-gaussian framework)**.** *Let $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, be a locally Lipschitz continuous loss that can be clipped at $M > 0$ and satisfies the supremum bound (25) for a $B > 0$. Moreover, let P be a distribution on $\mathcal{H} \times Y$ such that the variance bound (26) is satisfied for constants $\vartheta \in [0, 1], V \geqslant B^{2-\vartheta}$, and all $f \in \mathcal{H}$. Assume that for fixed $n \geqslant 1$ there exist constants $p \in (0, 1)$ and $a \geqslant B$ such that*

$$\mathbb{E}_{\widehat{\mathrm{P}}} e_i\left(\mathrm{id} : \mathcal{H} \to L_2(\widehat{P}_{\mathcal{H}})\right) \leqslant a i^{-\frac{1}{2p}}, \quad i \geqslant 1 \tag{78}$$

Finally, fix an $f_0 \in \mathcal{H}$ such that $h_{f_0}(X) - h_{f_0^{cl}}(X)$ is a real c-sub-gaussian random variable and $\mathbb{E}(h_{f_0}(X) - h_{f_0^{cl}}(X))^2 \leqslant D\mathbb{E}\left(h_{f_0}(X) - h_{f_0^{cl}}(X)\right)$. Then, for all fixed $\tau > 0$, $\lambda > 0$, and $\widehat{f}_\lambda$ ϵ-approximate clipped regularized empirical risk minimization (ϵ-CR-ERM):

$$\begin{aligned}\lambda\|\widehat{f}_\lambda\|_H^2 + L(\widehat{f}_\lambda^{cl}) - L(f_*) \leqslant 9\left(\lambda\|f_0\|_H^2 + L(f_0) - L(f_*)\right) + K\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \\ + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{(3D + 6c + 8B)\tau}{2n}\end{aligned} \tag{79}$$

with probability not less than $1 - 3e^{-\tau}$, where $K \geqslant 1$ is a constant only depending on p, M, B, ϑ, and V.

Proof. The proof mimics the one in Steinwart and Christmann (2008), but here we exploit Theorem 8, i.e. the adaptation of Theorem 7.20 in Steinwart and Christmann (2008) in the sub-gaussian framework. □

We can now proceed with the proof of Theorem 4 that is the content of Theorem 10 and Corollary 4.

Theorem 10. *Fix $\lambda > 0$, $\alpha \gtrsim n^{-1/p}$ and $0 < \delta < 1$. Under Assumptions 2, 4, 5, 6, and a polynomial decay of the spectrum of Σ with rate $1/p \in (1, \infty)$, as in (15), including also the additional hypothesis $\mathbb{E}(\ell(Y, \langle \mathcal{P}_m w_*, X\rangle) - \ell(Y, \langle \mathcal{P}_m w_*, X\rangle^{cl})^2 \leqslant D\mathbb{E}(\ell(Y, \langle \mathcal{P}_m w_*, X\rangle) - \ell(Y, \langle \mathcal{P}_m w_*, X\rangle^{cl}))$, with $D > 0$, then, with probability at least $1 - 2\delta$*

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \leqslant 9\lambda\|w_*\|^2 + 9C_0G\sqrt{\alpha}\|w_*\| + K\frac{a^{2p}}{\lambda^p n} + 216V\frac{\log(3/\delta)}{n} + \\ + \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\|w_*\|\log(3/\delta)}{n} \tag{80}$$

provided that n satisfies (58) and m satisfies (42) (uniform sampling) or (59) (ALS sampling), and where ℓ can be clipped at $M > 0$, $B > 0$ and $V > 0$ come from the supremum bound(19) and variance bound (20) respectively, $a \geqslant B$ and $K \geqslant 1$ is a constant only depending on p, M, B and V.

Proof. The proof mimics the proof of Theorem 11 where in (79), following the same reasoning as in (95), we choose

$$f_0 = \langle \mathcal{P}_{\mathcal{B}_m} w_*, \cdot\rangle \qquad \tilde{C} := 2B + 2CG\sqrt{2\operatorname{Tr}(\Sigma)}\,\|w_*\|\,.$$

Hence (79) with $\theta = 1$ reads

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) &\leqslant 9(\lambda\|\mathcal{P}_{\mathcal{B}_m} w_*\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)) + K\frac{a^{2p}}{\lambda^p n} + 216V\frac{\log(3/\delta)}{n} + \\ &+ \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\|w_*\|\log(3/\delta)}{n} \\ &\leqslant 9\lambda\|w_*\|^2 + 9(L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)) + K\frac{a^{2p}}{\lambda^p n} + 216V\frac{\log(3/\delta)}{n} + \\ &+ \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\|w_*\|\log(3/\delta)}{n}\end{aligned} \tag{81}$$

We can deal with the term $L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)$ as in (49) (but where we use Lemma 5 instead of Lemma 4), so that for $\alpha \gtrsim n^{-1/p}$ with probability greater than $1 - \delta$

$$L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*) \leqslant C_0 G\sqrt{\alpha}\,\|w_*\|$$

for some $C_0 > 0$. Hence, with probability at least $1 - 2\delta$

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \leqslant 9\lambda\|w_*\|^2 + 9C_0G\sqrt{\alpha}\|w_*\| + K\frac{a^{2p}}{\lambda^p n} + 216V\frac{\log(3/\delta)}{n} + \\ + \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\|w_*\|\log(3/\delta)}{n} \tag{82}$$

which proves the claim. □

The following corollary provides the optimal rates, whose proof is the same as for Corollary 5

Corollary 4. *Fix $\delta > 0$. Under the Theorem 10 set*

$$\lambda \asymp n^{-\frac{1}{1+p}} \tag{83}$$

$$\alpha \asymp n^{-\frac{2}{1+p}} \tag{84}$$

$$m \gtrsim n^{\frac{2p}{1+p}}\log n \tag{85}$$

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

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2_{\mathcal{H}} + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim \|w_*\|\Big(\frac{1}{n}\Big)^{\frac{1}{1+p}} \tag{86}$$

Notice that $\alpha \asymp n^{-\frac{2}{1+p}}$ is compatible with condition $\alpha \gtrsim d_{\alpha,2}(\Sigma) \asymp n^{-1/p}$ in Lemma 5.

G Proof of Theorem 5

Theorem 5 is the content of Theorem 11 and Corollary 5

Theorem 11. *Fix $\lambda > 0$, $\alpha \gtrsim n^{-1/p}$ and $0 < \delta < 1$. Under Assumptions 2, 4, 7, and a polynomial decay of the spectrum of Σ with rate $1/p \in (1, \infty)$, as in (15), including also the additional hypothesis $\mathbb{E}(\ell(Y, \langle \mathcal{P}_m w_\lambda, X\rangle) - \ell(Y, \langle \mathcal{P}_m w_\lambda, X\rangle^{cl})^2 \leqslant D\mathbb{E}(\ell(Y, \langle \mathcal{P}_m w_\lambda, X\rangle) - \ell(Y, \langle \mathcal{P}_m w_\lambda, X\rangle^{cl}))$, with $D > 0$, then with probability at least $1 - 2\delta$*

$$\lambda\|\widehat{\beta}_{\lambda,m}\|_{\mathcal{H}}^2 + L(\widehat{\beta}_{\lambda,m}^{cl}) - L(f_*) \leqslant 9\mathcal{A}(\lambda) + 9C_0G\sqrt{\frac{\alpha\mathcal{A}(\lambda)}{\lambda}} + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + 3\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \\ + \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \tag{87}$$

provided that n satisfies (58) and m satisfies (42) (uniform sampling) or (59) (ALS sampling), and where ℓ can be clipped at $M > 0$, $B > 0$ and $\theta \in [0, 1]$ come from the supremum bound(25) and variance bound (26) respectively, $a \geqslant B$ and $K \geqslant 1$ is a constant only depending on p, M, B, θ and V.

Proof. We adapt the proof of Theorem 7.23 in Steinwart and Christmann (2008) to $\widehat{\beta}_{\lambda,m}$. Set

$$r_{\mathcal{H}}^* = \inf_{w\in\mathcal{H}} \lambda\|w\|^2 + L(w^{cl}) - L(f_*) \tag{88}$$

$$r_{\mathcal{B}_m}^* = \inf_{w\in\mathcal{B}_m} \lambda\|w\|^2 + L(w^{cl}) - L(f_*) \tag{89}$$

$$\mathcal{H}_r = \{w \in \mathcal{H} : \lambda\|w\|^2 + L(w^{cl}) - L(f_*) \leqslant r\} \qquad r > r_{\mathcal{H}}^* \tag{90}$$

$$(\mathcal{B}_m)_r = \{w \in \mathcal{B}_m : \lambda\|w\|^2 + L(w^{cl}) - L(f_*) \leqslant r\} \qquad r > r_{\mathcal{B}_m}^* \tag{91}$$

(see Eq. (7.32)-(7.33) in Steinwart and Christmann (2008)). Let's notice that $r_{\mathcal{B}_m}^* \geqslant r_{\mathcal{H}}^*$, which means that $(\mathcal{B}_m)_r \subseteq \mathcal{H}_r$. As a consequence, using also Theorem 15 in Steinwart et al. (2009) stating that the decay condition (15) of the spectrum of the covariance operator Σ is equivalent to the polynomial decay of the (dyadic) entropy numbers e_j (see Lemma 6), we have that, analogously to the proof of Theorem 7.23 in Steinwart and Christmann (2008) (see Lemma 7.17 and eq. (A.36) in Steinwart and Christmann (2008) for details):

$$\mathbb{E}_{\widehat{P}}[e_j(\mathrm{id} : (\mathcal{B}_m)_r \to L_2(\widehat{P}_{\mathcal{H}}))] \leqslant \mathbb{E}_{\widehat{P}}[e_j(\mathrm{id} : \mathcal{H}_r \to L_2(\widehat{P}_{\mathcal{H}}))] \leqslant 2\left(\frac{r}{\lambda}\right)^{1/2} a j^{-\frac{1}{2p}}$$

for some $a \geqslant B$, where the first inequality is a consequence of $(\mathcal{B}_m)_r \subseteq \mathcal{H}_r$ and $\widehat{P}_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical (marginal) measure.

Furthermore $\widehat{\beta}_{\lambda,m}$ is a *clipped regularized empirical risk minimizer* over $\mathcal{B}_m$ (see Definition 7.18 in Steinwart and Christmann (2008)) since

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + \widehat{L}(\widehat{\beta}_{\lambda,m}^{cl}) \leqslant \lambda\|\widehat{\beta}_{\lambda,m}\|^2 + \widehat{L}(\widehat{\beta}_{\lambda,m}) = \inf_{\beta\in\mathcal{B}_m} [\lambda\|w\|^2 + \widehat{L}(w)].$$

Then, applying Theorem 9 (sub-gaussian adaptation of Theorem 7.23 in Steinwart and Christmann (2008)) with probability at least $1 - \delta$

$$\lambda\|\widehat{\beta}_{\lambda,m}\|_{\mathcal{H}}^2 + L(\widehat{\beta}_{\lambda,m}^{cl}) - L(f_*) \leqslant 9\left(\lambda\|f_0\|_H^2 + L(f_0) - L(f_*)\right) + K\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \\ + 3\left(\frac{72V\log(3/\delta)}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{3D+6c+8B\log(3/\delta)}{2n}. \tag{92}$$

We define $w_\lambda := \arg\min_{w\in\mathcal{H}} L(w) + \lambda\|w\|^2$. Now, since

$$\begin{aligned}\|\langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle\|_{\psi_2} &\leqslant \|\|\mathcal{P}_{\mathcal{B}_m} w_\lambda\| \cdot \|X\|\|_{\psi_2} \leqslant \|\|w_\lambda\| \cdot \|X\|\|_{\psi_2} \\ &= \|w_\lambda\| \cdot \|\|X\|\|_{\psi_2} \leqslant \sqrt{2}C\,\|w_\lambda\| \cdot \|\|X\|\|_{L_2} \\ &= C\sqrt{2\operatorname{Tr}(\Sigma)}\,\|w_\lambda\|\end{aligned}$$

where we used the fact that $\|X\|$ is sub-gaussian since, given an orthonormal basis e_i,

$$\begin{aligned}\left\| \|X\| \right\|_{\psi_2}^2 \leqslant \left\| \|X\|^2 \right\|_{\psi_1} &= \left\| \sum_i \langle X, e_i\rangle^2 \right\|_{\psi_1} \leqslant \sum_i \left\| \langle X, e_i\rangle^2 \right\|_{\psi_1} \\ &\leqslant 2\sum_i \|\langle X, e_i\rangle\|_{\psi_2}^2 \leqslant 2C^2 \|\langle X, e_i\rangle\|_{L_2}^2 = 2C^2 \operatorname{Tr}[\Sigma]\end{aligned}$$

Then $\langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle$ is a real $C\sqrt{2\operatorname{Tr}(\Sigma)}\,\|w_\lambda\|$-sub-gaussian random variable. Moreover we have

$$\|\ell(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle)\|_{\psi_2} \leqslant B + CG\sqrt{2\operatorname{Tr}(\Sigma)}\,\|w_\lambda\|\,. \tag{93}$$

Recalling the definition of *clipping*, we have $\ell(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle^{cl}) \leqslant \ell\left(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle\right)$ which implies

$$\|\ell(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle^{cl})\|_{\psi_2} = \sup_{p\geqslant 2} \frac{\|\ell(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle^{cl})\|_{L_p}}{\sqrt{p}} \leqslant \sup_{p\geqslant 2} \frac{\|\ell(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle)\|_{L_p}}{\sqrt{p}} = \|\ell(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle)\|_{\psi_2} \tag{94}$$

for the monotonicity of the L_p-norm. Putting everything together we get

$$\|h_{\mathcal{P}_{\mathcal{B}_m} w_\lambda}(X) - h_{\mathcal{P}_{\mathcal{B}_m} w_\lambda^{cl}}(X)\|_{\psi_2} = \|\ell\left(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle\right) - \ell\left(y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle^{cl}\right)\|_{\psi_2} \leqslant 2B + 2CG\sqrt{2\operatorname{Tr}(\Sigma)}\,\|w_\lambda\| = \tilde{C}. \tag{95}$$

We can finally conclude that $h_{\mathcal{P}_{\mathcal{B}_m} w_\lambda}(X) - h_{\mathcal{P}_{\mathcal{B}_m} w_\lambda^{cl}}(X)$ is a $\tilde{C}$-sub-gaussian random variable. Assumption $\mathbb{E}(\ell(Y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle) - \ell(Y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle^{cl})^2 \leqslant D\mathbb{E}(\ell(Y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle) - \ell(Y, \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, X\rangle^{cl}))$ allows us to apply Theorem 9 for $f_0 := \langle \mathcal{P}_{\mathcal{B}_m} w_\lambda, \cdot\rangle$ with $c = \tilde{C}$. We rewrite (79) as:

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) &\leqslant 9(\lambda\|\mathcal{P}_{\mathcal{B}_m} w_\lambda\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(f_*)) + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + 3\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \\ &\quad + \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\,\|w_\lambda\|\log(3/\delta)}{n} \\ &= 9(\lambda\|\mathcal{P}_{\mathcal{B}_m} w_\lambda\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda) + L(w_\lambda) - L(f_*)) + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + \\ &\quad + 3\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \\ &\leqslant 9(L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda) + \lambda\|w_\lambda\|^2 + L(w_\lambda) - L(f_*)) + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + \\ &\quad + 3\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \\ &= 9\mathcal{A}(\lambda) + 9(L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda)) + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + 3\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \\ &\quad + \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}\end{aligned} \tag{96}$$

where we used the fact that $\|w_\lambda\| \leqslant \sqrt{\mathcal{A}(\lambda)/\lambda}$.

We can deal with the term $L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda)$ as in (49) (but where we use Lemma 5 instead of Lemma 4 to exploit sub-gaussianity), so that for $\alpha \gtrsim n^{-1/p}$ with probability greater than $1-\delta$

$$L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda) \leqslant C_0 G\sqrt{\alpha}\,\|w_\lambda\| \leqslant C_0 G\sqrt{\alpha}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}$$

for some $C_0 > 0$. We finally obtain with probability greater than $1-2\delta$

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|_{\mathcal{H}}^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) &\leqslant 9\mathcal{A}(\lambda) + 9C_0 G\sqrt{\frac{\alpha\mathcal{A}(\lambda)}{\lambda}} + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + 3\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \\ &\quad + \frac{(3D+8B)\log(3/\delta)}{2n} + \frac{6CG\sqrt{2\operatorname{Tr}(\Sigma)}\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}\end{aligned} \tag{97}$$

which proves the first claim. □

The following corollary provides the optimal rates.

Corollary 5. *Fix $\delta > 0$. Under the Theorem 11 and the source condition*

$$\mathcal{A}(\lambda) \leqslant A_0 \lambda^r$$

for some $r \in (0, 1]$, set

$$\lambda \asymp n^{-\min\{\frac{2}{r+1}, \frac{1}{r(2-p-\theta+\theta p)+p}\}} \tag{98}$$

$$\alpha \asymp n^{-\min\{2, \frac{r+1}{r(2-p-\theta+\theta p)+p}\}} \tag{99}$$

$$m \gtrsim n^{\min\{2p, \frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}} \tag{100}$$

for ALS sampling, with probability at least $1 - 2\delta$

$$\lambda \|\widehat{\beta}_{\lambda,m}\|_{\mathcal{H}}^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) \lesssim n^{-\min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\theta+\theta p)+p}\}} \tag{101}$$

Proof. Lemma 4 with Proposition 4 gives

$$m \gtrsim d_{\alpha,2} \log(n/\delta), \qquad d_{\alpha,2} \lesssim \alpha^{-p} \qquad \alpha \asymp \frac{\log^{1/p}(n/\delta)}{m^{1/p}} \tag{102}$$

Lemma A.1.7 in Steinwart and Christmann (2008) with $r = 2$, $1/\gamma = (2 - p - \theta + \theta p)$, $\alpha = p$, $\beta = r$ shows that the choice of λ, α and m given by (98)–(100) provides the optimal rate. □

Notice that $\alpha \asymp n^{-\min\{2, \frac{r+1}{r(2-p-\theta+\theta p)+p}\}}$ is compatible with condition $\alpha \gtrsim d_{\alpha,2}(\Sigma) \asymp n^{-1/p}$ in Lemma 5.

H Effective Dimension and Eigenvalues Decay

In this section, we derive tight bounds for $d_{\alpha,2}$ defined by (13) when assuming respectively polynomial and exponential decay of the eigenvalues $\sigma_j(\Sigma)$ of Σ.

Proposition 4 (Polynomial eigenvalues decay, Proposition 3 in Caponnetto and De Vito (2007))**.**
If for some $\gamma \in \mathbb{R}^+$ and $1 < \beta < +\infty$

$$\sigma_i \leqslant \gamma i^{-\beta}$$

then

$$d_{\alpha,2} \leqslant \gamma \frac{\beta}{\beta - 1} \alpha^{-1/\beta} \tag{103}$$

Proof. Since the function $\sigma/(\sigma + \alpha)$ is increasing in σ and using the spectral theorem $\Sigma = UDU^*$ combined with the fact that $\mathrm{Tr}(UDU^*) = \mathrm{Tr}(U(U^*D)) = \mathrm{Tr}D$

$$d_{\alpha,2} = \mathrm{Tr}(\Sigma(\Sigma + \alpha I)^{-1}) = \sum_{i=1}^{\infty} \frac{\sigma_i}{\sigma_i + \alpha} \leqslant \sum_{i=1}^{\infty} \frac{\gamma}{\gamma + i^\beta \alpha} \tag{104}$$

The function $\gamma/(\gamma + x^\beta \alpha)$ is positive and decreasing, so

$$\begin{aligned} d_{\alpha,2} &\leqslant \int_0^\infty \frac{\gamma}{\gamma + x^\beta \alpha} dx \\ &= \alpha^{-1/\beta} \int_0^\infty \frac{\gamma}{\gamma + \tau^\beta} d\tau \\ &\leqslant \gamma \frac{\beta}{\beta - 1} \alpha^{-1/\beta} \end{aligned} \tag{105}$$

since $\int_0^\infty (\gamma + \tau^\beta)^{-1} \leqslant \beta/(\beta - 1)$. □

Proposition 5 (Exponential eigenvalues decay)**.**
If for some $\gamma, \beta \in \mathbb{R}^+ \sigma_i \leqslant \gamma e^{-\beta i}$ *then*

$$d_{\alpha,2} \leqslant \frac{\log(1+\gamma/\alpha)}{\beta} \tag{106}$$

Proof.

$$d_{\alpha,2} = \sum_{i=1}^{\infty} \frac{\sigma_i}{\sigma_i + \alpha} = \sum_{i=1}^{\infty} \frac{1}{1+\alpha/\sigma_i} \leqslant \sum_{i=1}^{\infty} \frac{1}{1+\alpha' e^{\beta i}} \leqslant \int_0^{+\infty} \frac{1}{1+\alpha' e^{\beta x}} dx \tag{107}$$

where $\alpha' = \alpha/\gamma$. Using the change of variables $t = e^{\beta x}$ we get

$$\begin{aligned} (107) &= \frac{1}{\beta} \int_1^{+\infty} \frac{1}{1+\alpha' t} \frac{1}{t} dt = \frac{1}{\beta} \int_1^{+\infty} \Big[\frac{1}{t} - \frac{\alpha'}{1+\alpha' t}\Big] dt = \frac{1}{\beta} \Big[\log t - \log(1+\alpha' t)\Big]_1^{+\infty} \\ &= \frac{1}{\beta} \Big[\log\Big(\frac{t}{1+\alpha' t}\Big)\Big]_1^{+\infty} = \frac{1}{\beta} \Big[\log(1/\alpha') + \log(1+\alpha')\Big] \end{aligned} \tag{108}$$

So we finally obtain

$$d_{\alpha,2} \leqslant \frac{1}{\beta} \Big[\log(\gamma/\alpha) + \log(1+\alpha/\gamma)\Big] = \frac{\log(1+\gamma/\alpha)}{\beta} \tag{109}$$

□

The following result is the content of Theorem 15 in Steinwart et al. (2009). Given a bounded operator A between two Hilbert spaces $\mathcal{H}_1$ and H_2, denote by $e_j(A)$ the entropy numbers of A and by $\widehat{P}_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ the empirical (marginal) measure associated with the input data $x_i, \dots, x_n$. Regard the data matrix $\widehat{X}$ as the inclusion operator $\mathrm{id} : \mathcal{H} \to L_2(\widehat{P})$

$$(\mathrm{id}\, w)(x_i) = \langle w, x_i \rangle \qquad i = 1, \dots, n$$

Lemma 6. *Let* $p \in (0,1)$*. Then*

$$\mathbb{E}_{\widehat{P}}[e_j(\mathrm{id} : \mathcal{H} \to L_2(\widehat{P}))] \sim j^{-\frac{1}{2p}} \tag{110}$$

if and only if

$$\sigma_j(\Sigma) \sim j^{-\frac{1}{p}} \tag{111}$$

I Constrained problem

In this section we investigate the so called *constrained problem*. As (9) the hypothesis space is still the subspace $\mathcal{B}_m \subseteq \mathcal{H}$ spanned by $\{\tilde{x}_1, \cdots, \tilde{x}_m\}$ with $\{\tilde{x}_1, \cdots, \tilde{x}_m\}$ being the sampled input points and the empirical estimator is the minimizer of ERM on the ball of radius R belonging to the subspace $\mathcal{B}_m$. More precisely, for any $R > 0$ we set

$$\widehat{\beta}_{R,m} = \operatorname*{arg\,min}_{w \in \mathcal{B}_m, \|w\| \leqslant R} \widehat{L}(w) \tag{112}$$

For sake of simplicity we assume the best in model to exist. We start presenting the finite sample error bounds for uniform and approximate leverage scores subsampling of the m points.

Theorem 12. *Fix* $R > 0$*,* $\alpha > 0$*,* $0 < \delta < 1$*. Under Assumptions 1, 2, 3, with probability at least* $1 - \delta$

$$L(\widehat{\beta}_{R,m}) - L(f_{\mathcal{H}}) \leqslant \frac{2GR\kappa}{\sqrt{n}} \Big(2 + \sqrt{2\log(1/\delta)}\Big) + 2GR\sqrt{\alpha} \tag{113}$$

provided that $R \geqslant \|w_*\|$*,* $n \geqslant 1655\kappa^2 + 223\kappa^2 \log\frac{4\kappa^2}{\delta}$ *and* m *satisfies*

1. *for uniform sampling*

$$m \geqslant 67 \log\frac{4\kappa^2}{\alpha\delta} \vee 5 d_{\alpha,\infty} \log\frac{4\kappa^2}{\alpha\delta} \tag{114}$$

2. *for ALS sampling and T-approximate leverage scores with subsampling probabilities Q_α, $t_0 > \frac{19\kappa^2}{n}\log\frac{4n}{\delta}$,*

$$m \geqslant 334\log\frac{16n}{\delta} \vee 78T^2 d_{\alpha,2}\log\frac{16n}{\delta} \tag{115}$$

where $\alpha \geqslant \frac{19\kappa^2}{n}\log\frac{4n}{\delta}$.

Under the above condition, with the choice $\alpha \asymp 1/n$, the estimator achieves the optimal bound

$$\begin{aligned} L(\widehat{\beta}_{R,m}) - L(f_{\mathcal{H}}) &\leqslant \frac{2GR\kappa}{\sqrt{n}}\Big(2+\sqrt{2\log(1/\delta)}\Big) + 2GR\frac{1}{\sqrt{n}} \\ &= R\sqrt{\log(1/\delta)}\, O\Big(\frac{1}{\sqrt{n}}\Big) \end{aligned} \tag{116}$$

Proof. We decompose the excess risk of $\widehat{\beta}_{R,m}$ with respect to the target w_*

$$\begin{aligned} L(\widehat{\beta}_{R,m}) - L(w_*) &= L(\widehat{\beta}_{R,m}) - \widehat{L}(\widehat{\beta}_{R,m}) + \underbrace{\widehat{L}(\widehat{\beta}_{R,m}) - \widehat{L}(\mathcal{P}_m w_*)}_{\leqslant 0} + \\ &+ \quad \widehat{L}(\mathcal{P}_m w_*) - L(\mathcal{P}_m w_*) + L(\mathcal{P}_m w_*) - L(w_*) \\ &\leqslant 2\underbrace{\sup_{w\in\mathcal{B}_m, \|w\|\leqslant R}\Big(L(w)-\widehat{L}(w)\Big)}_{\mathbf{A}} + \underbrace{L(\mathcal{P}_m w_*) - L(w_*)}_{\mathbf{B}} \end{aligned} \tag{117}$$

where $\|\mathcal{P}_m w_*\| \leqslant R$ since $R \geqslant \|w_*\|$.

Bound for the term A:
Term **A** is bounded by Lemma 1, so that with probability at least $1-\delta$

$$\mathbf{A} \leqslant \frac{GR\kappa}{\sqrt{n}}\Big(2+\sqrt{2\log(1/\delta)}\Big). \tag{118}$$

Bound for term B:
Term **B** is bounded as Term C in the proof of Theorem 7, see (49)

$$\mathbf{B} \leqslant G\|\Sigma^{1/2}(I-\mathcal{P}_m)\|\|w_*\| \leqslant GR\|\Sigma^{1/2}(I-\mathcal{P}_m)\| \tag{119}$$

and we estimate $\|\Sigma^{1/2}(I-\mathcal{P}_m)\|$ using Lemma 3 for uniform sampling and Lemma 4 for ALS selection. □

Again, bound 116 provides a convergence rate, which is optimal from a statistical point of view, but that requires at least $m \sim n\log n$ subsampled points since, without further assumptions the effective dimension $d_{\alpha,2}$, as well as $d_{\alpha,\infty}$, can in general be bounded only by κ^2/α. Clearly, this makes the approach completely useless. As for the regularized estimator, to overcome this issue we are forced to assume fast decay of the eigenvalues of the covariance operator Σ, as in Bach (2017). Under this condition the following results – whose proof is identical to the proof of Proposition 2, shows that the optimal rate can be achieved with an efficient computational cost at least for ALS.

Corollary 6. *Under the condition of Theorem 12,*

1. *if Σ has a polynomial decay, i.e. for some $\gamma \in \mathbb{R}^+$, $p \in (0,1)$,*

$$\sigma_j(\Sigma) \leqslant \gamma j^{-\frac{1}{p}},$$

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

$$L(\widehat{\beta}_{R,m}) - L(w_*) \lesssim R\sqrt{\frac{\log(1/\delta)}{n}} + R\sqrt{\frac{\log^{1/p} n}{m^{1/p}}} = R\sqrt{\log(1/\delta)}\, O\Big(\frac{\log^{1/p} n}{\sqrt{n}}\Big) \tag{120}$$

with $m \gtrsim n^p \log n$ subsampled points according to ALS method.

2. *if Σ has an exponential decay, i.e. for some $\gamma, \beta \in \mathbb{R}^+$,*

$$\sigma_j(\Sigma) \leqslant \gamma e^{-\beta j}$$

with probability at least $1-\delta$:

$$L(\widehat{\beta}_{R,m}) - L(w_*) \lesssim R\sqrt{\frac{\log(1/\delta)}{n}} + Re^{-\frac{m}{2\log n}} = R\sqrt{\log(1/\delta)}\, O\Big(\frac{1}{\sqrt{n}}\Big) \tag{121}$$

with $m \gtrsim \log^2 n$ subsampled points according to ALS method.

J Experiments: datasets and tuning

Here we report further information on the used data sets and the set up used for parameter tuning.
For Nyström SVM with Pegaos we tuned the kernel parameter σ and λ regularizer with a simple grid search ($\sigma \in [0.1, 20]$, $\lambda \in [10^{-8}, 10^{-1}]$, initially with a coarse grid and then more refined around the best candidates). An analogous procedure has been used for K-SVM with its parameters C and γ. The details of the considered data sets and the chosen parameters for our algorithm in Table 2 and 4 are the following:
SUSY (Table 2 and 4, $n = 5 \times 10^6$, $d = 18$): we used a Gaussian kernel with $\sigma = 4$, $\lambda = 3 \times 10^{-6}$ and $m_{ALS} = 2500$, $m_{uniform} = 2500$.
Mnist binary (Table 2 and 4, $n = 7 \times 10^4$, $d = 784$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 3 \times 10^{-6}$ and $m_{ALS} = 15000$, $m_{uniform} = 20000$.
Usps (Table 2 and 4, $n = 9298$, $d = 256$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 5 \times 10^{-6}$ and $m_{ALS} = 2500$, $m_{uniform} = 4000$.
Webspam (Table 2 and 4, $n = 3.5 \times 10^5$, $d = 254$): we used a Gaussian kernel with $\sigma = 0.25$, $\lambda = 8 \times 10^{-7}$ and $m_{ALS} = 11500$, $m_{uniform} = 20000$.
a9a (Table 2 and 4, $n = 48842$, $d = 123$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 1 \times 10^{-5}$ and $m_{ALS} = 800$, $m_{uniform} = 1500$.
CIFAR (Table 2 and 4, $n = 6 \times 10^4$, $d = 400$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 2 \times 10^{-6}$ and $m_{ALS} = 20500$, $m_{uniform} = 20000$.

Table 4: Comparison between ALS and uniform sampling. To achieve similar accuracy, uniform sampling usually requires larger m than ALS sampling. Therefore, even if it does not need leverage scores computations, Nyström-Pegasos with uniform sampling can be more expensive both in terms of memory and time (in seconds).

	Nyström-Pegasos (ALS)			Nyström-Pegasos (Uniform)		
Datasets	c-err	t train	t pred	c-err	t train	t pred
SUSY	$20.0\% \pm 0.2\%$	608 ± 2	134 ± 4	$20.1\% \pm 0.2\%$	592 ± 2	129 ± 1
Mnist bin	$2.2\% \pm 0.1\%$	1342 ± 5	491 ± 32	$2.3\% \pm 0.1\%$	1814 ± 8	954 ± 21
Usps	$3.0\% \pm 0.1\%$	19.8 ± 0.1	7.3 ± 0.3	$3.0\% \pm 0.2\%$	66.1 ± 0.1	48 ± 8
Webspam	$1.3\% \pm 0.1\%$	2440 ± 5	376 ± 18	$1.3\% \pm 0.1\%$	4198 ± 40	1455 ± 180
a9a	$15.1\% \pm 0.2\%$	29.3 ± 0.2	1.5 ± 0.1	$15.1\% \pm 0.2\%$	30.9 ± 0.2	3.2 ± 0.1
CIFAR	$19.2\% \pm 0.1\%$	2408 ± 14	820 ± 47	$19.0\% \pm 0.1\%$	2168 ± 19	709 ± 13