Concentration of Non-Isotropic Random Tensors with Applications to Learning and Empirical Risk Minimization

Mathieu Even   MATHIEU.EVEN@INRIA.FR
INRIA, DI/ENS, PSL Research University, Paris, France

Laurent Massoulié   LAURENT.MASSOULIE@INRIA.FR
INRIA, DI/ENS, PSL Research University, MSR-Inria Joint Centre, Paris, France

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Dimension is an inherent bottleneck to some modern learning tasks, where optimization methods suffer from the size of the data. In this paper, we study non-isotropic distributions of data and develop tools that aim at reducing these dimensional costs by a dependency on an effective dimension rather than the ambient one. Based on non-asymptotic estimates of the metric entropy of ellipsoids -that prove to generalize to infinite dimensions- and on a chaining argument, our uniform concentration bounds involve an *effective dimension* instead of the global dimension, improving over existing results. We show the importance of taking advantage of non-isotropic properties in learning problems with the following applications: *i)* we improve state-of-the-art results in statistical preconditioning for communication-efficient distributed optimization, *ii)* we introduce a non-isotropic randomized smoothing for non-smooth optimization. Both applications cover a class of functions that encompasses empirical risk minization (ERM) for linear models.

Keywords: Effective Dimension, Large Deviation, Chaining Method, Metric Entropy, Ellipsoids, Random Tensors, Statistical Preconditioning, Smoothing Technique.

1. Introduction

The sum of *i.i.d.* symmetric random tensors of order 2 and rank 1 (*i.e.* symmetric random matrices of rank 1) is studied in probability and statistics both for theoretical and practical interests, the most classical application being covariance estimation. The empirical mean of such matrices follows the Wishart distribution (Wishart, 1928; Uhlig, 1994). Marčenko and Pastur (1967) proved the convergence in law of their spectrum when the number of observations and the dimension are of the same order. Machine Learning applications however require non-asymptotic properties, such as concentration bounds for a potentially large finite number of observations and finite dimension (Tropp, 2011, 2015; Donoho et al., 2017; Minsker, 2017), to control the eigenvalues of sums of independent matrices, namely:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} a_i a_i^\top - \mathbb{E}\left[aa^\top \right] \right\|_{\mathrm{op}} = \sup_{\|x\| \leq 1} \frac{1}{n} \sum_{i=1}^{n} x^\top \left(a_i a_i^\top - \mathbb{E}\left[aa^\top \right] \right) x \tag{1}$$

for $a, a_1, ..., a_n$ *i.i.d.* random variables in $\mathbb{R}^d$.

1.1. Theoretical Contributions

Our main contribution consists in new tools for the control of quantities generalizing (1). More precisely, for $r \geq 2$, $f_1, ..., f_r$ Lipschitz functions on $\mathbb{R}$, $a, a_1, ..., a_n$ *i.i.d.* random variables in $\mathbb{R}^d$, and $\mathcal{B}$ the d-dimensional unit ball, we derive in Section 2 concentration bounds on:

$$\sup_{x_1,...,x_r \in \mathcal{B}} \left\{ \frac{1}{n} \sum_{i \in [n]} \left(\prod_{k=1}^{r} f_k(a_i^\top x_k) - \mathbb{E}\left[\prod_{k=1}^{r} f_k(a^\top x_k) \right] \right) \right\}. \tag{2}$$

We thereby extend previous results in three directions. *i)* Matrices are tensors of order 2, which we generalize by treating symmetric random tensors of rank 1 and order $r \geq 2$ (Section 2.3). *ii)* We consider non linear functions f_i of scalar products $\langle a_i, x \rangle$, motivated by Empirical Risk Minimization. (2) can thus be seen as the uniform maximum deviation of a symmetric random tensor of order r and rank 1, with non-linearities $f_1, ..., f_r$. *iii)* Finally, by observing that data are usually distributed in a non-isotropic way (the *MNIST* dataset lies in a 712 dimensional space, yet its empirical covariance matrix is of effective dimension less than 3 for instance), we generalize classical isotropic assumptions on random variables a_i by introducing a non-isotropic counterpart:

Definition 1 (Σ-Subgaussian Random Vector) *A random variable a with values in $\mathbb{R}^d$ is Σ - subgaussian for $\Sigma \in \mathbb{R}^{d \times d}$ a positive-definite matrix if:*

$$\forall t > 0, \forall x \in \mathcal{B}, \mathbb{P}(|a^\top x| > t) \leq 2 \exp\left(-\frac{1}{2} \frac{t^2}{x^\top \Sigma x} \right). \tag{3}$$

A gaussian $\mathcal{N}(0, \Sigma)$ is for instance Σ-subgaussian. Note however that in the general case, Σ is not equal to the covariance matrix. The aim is then to derive concentration bounds on (2) (Section 2) that involve an *effective dimension* of Σ: a quantity smaller than the global dimension d, that reflects the non-isotropic repartition of the data:

Definition 2 (Effective Dimension $d_{\text{eff}}(r)$) *Let $\Sigma \in \mathbb{R}^{d \times d}$ a symmetric positive semi-definite matrix of size $d \times d$, where $d \in \mathbb{N}^*$. Let $\sigma_1^2 \geq \sigma_2^2 \geq ... \geq \sigma_d^2 \geq 0$ denote its ordered eigenvalues. For any $r \in \mathbb{N}^*$, let $d_{\text{eff}}(r)$ be defined as follows:*

$$d_{\text{eff}}(r) := \sum_{i=1}^{d} \left(\frac{\sigma_i}{\sigma_1} \right)^{\frac{2}{r}} = \frac{\text{Tr}(\Sigma^{1/r})}{\left\| \Sigma^{1/r} \right\|_{\text{op}}}. \tag{4}$$

This notion generalizes intrinsic dimension *in Tropp (2015) and* stable rank *in Vershynin (2011, 2014), both obtained for $r = 1$.*

Chaining Argument and Metric Entropy of Ellipsoids: Control of (2) involves a chaining argument (Boucheron et al. (2013), Chapter 13). In the simplest version of chaining, in order to bound a random variable of the form $\sup_{t \in \mathcal{T}} X_t$, one discretizes the set of indices $\mathcal{T}$ and approximates the value $\sup_{t \in \mathcal{T}} X_t$ by a supremum taken over successively refined discretizations. To exploit the non-isotropic properties of Σ-subgaussian random variables, we apply chaining based on a covering of the unit ball $\mathcal{B}$ with ellipsoids. Our approach yields similarities with that of Zhong (2017), who uses chaining with ellipsoids for a different purpose (control of eigenvectors). In section 2.3, in the setting where $f_1 = ... = f_r = Id$, control of (2) reduces to controlling the operator norm of

empirical tensors. This can be done using our bounds on the ε-entropy of ellipsoids, without the use of chaining.

In Section 3, we present results on the number of balls of fixed radius ε needed to cover an ellipsoid in dimension d. The logarithm of this quantity is often called the ε-*entropy* of an ellipsoid. Dumer et al. (2004) studied the limit $d \rightarrow \infty$, while we provide non-asymptotic estimates. Furthermore, in Appendix A.3, we extend these results to ellipsoids in infinite dimension, obtaining bounds on metric entropy in terms of power-law norm decay. We believe these technical results (both in finite and infinite dimension) to be of strong practical and theoretical interests: the bridge between covering numbers and suprema of random subgaussian processes is rather thin due to Dudley's inequality (Dudley, 1967). Bounding metric entropy of ellipsoids is thus a step towards uniform bounds on more general random variables than the one we consider in (2).

1.2. Applications in Learning Problems and ERM

We show the relevance of our concentration bounds through the following applications.

Operator Norm Of Tensors Setting $f_1 = ... = f_r = Id$ yields the operator norm of the empirical tensor $\frac{1}{n}\sum_i a_i^{\otimes r} - \mathbb{E}a^{\otimes r}$ in (2). In Section 2.3 we derive precise large deviation bounds on such tensors involving the effective dimension $d_{\text{eff}}(r)$, improving on previous works (Bubeck et al., 2020; Paouris et al., 2017) which depended on the global dimension. Optimal concentration inequalities on rank 1 symmetric tensors (*i.e.* of the form $a^{\otimes r}$) are not known. We refer the interested reader to Vershynin (2020) for the study of rank 1 tensors of the form $a_1 \otimes ... \otimes a_r$ where $a_1, ..., a_r$ are *i.i.d.* random variables, a different problem than ours. In Appendix F, we apply these bounds to the study of the Lipschitz constant of two-layered neural networks with polynomial activation, elaborating on the results in Bubeck et al. (2020).

Concentration of Hessians and Statistical Preconditioning For ℓ a twice differentiable function on $\mathbb{R}$ and Hessian-Lipschitz, let $f(x) = \frac{1}{n}\sum_{i=1}^{n}\ell(a_i^\top x)$. Then, $\nabla^2 f(x) = \frac{1}{n}\sum_{i=1}^{n}\ell''(a_i^\top x)a_i a_i^\top$, and setting $r = 3, f_1 = \ell'', f_2 = f_3 = Id$ in (2) yields $\sup_{x\in\mathcal{B}}\left\|\nabla^2 f(x) - \mathbb{E}[\nabla^2 f(x)]\right\|_{\text{op}}$. Controlling such quantities is relevant in optimization when studying functions that have an empirical risk structure. Methods such as statistical preconditioning (Shamir et al., 2014) take advantage of the *i.i.d.* structure of the observations, as we illustrate in Section 4. Our results improve on the state of the state-of-the-art (Hendrikx et al., 2020), establishing guarantees based on $d_{\text{eff}}(r)$ rather than d.

Randomized Smoothing Minimizing a non-smooth convex function f is a difficult problem, as acceleration methods cannot be used. Duchi et al. (2012); Scaman et al. (2018) propose to use the gradients of f^γ a smoothed version of f, where $f^\gamma(x) = \mathbb{E}_{X\sim\mathcal{N}(0,I_d)}[f(x + \gamma X)]$. This method suffers from a dimensional cost, a factor $d^{1/4}$ in the convergence time, that cannot in general be removed (Bubeck et al., 2019; Nemirovsky and Yudin, 1985). In Section 5, considering an empirical risk structure for f and a non-isotropic smoothing distribution for X, we take advantage of the non-isotropic repartition of data to obtain an effective dimension $d_{\text{eff}}(r)$ instead of the whole dimension d.

Organization of the paper We first present our 3 main uniform concentration bounds in Section 2: control of (2) and of the same quantity but un-centered (both using chaining), and a more precise control of (2) in the case where $f_1 = ... = f_r = Id$ (control of empirical mean of symmetric random tensors of rank 1 ann order p). In section 3, we provide bounds on the metric entropy of ellipsoids

in terms of effective dimension. We also investigate the case of infinite dimension with the notion of spectral dimension. The last two sections present two applications of the results presented in Section 2. In Section 4, we apply Theorem 1 to control uniform deviation of Hessians, in order to prove that statistical preconditioning methods naturally adapt to the underlying effective dimension. In Section 5, we introduce a non-isotropic smoothing method for empirical risk minimization.

2. Main Theoretical Results

2.1. Concentration Bound With Centering

Theorem 1 (Concentration With Centering) *Let $r \geq 2$ and $d, n \geq 1$ integers. Let $\Sigma \in \mathbb{R}^{d \times d}$ a positive-definite matrix and $a, a_1, ..., a_n$ i.i.d. $\Sigma-$subgaussian random variables. Let $d_{\text{eff}}(s), s \in \mathbb{N}^*$ be defined as in (4). Let $f_1, ..., f_r$ be 1-Lipshitz continuous functions on $\mathbb{R}$ such that $f_i(0) = 0$ for $i \in [n]$. For all $k = 1, ..., r$, let $B_k > 0$ such that:*

$$\forall x \in \mathcal{B}, \forall i \in [n], |f_k(a_i^\top x)| \leq B_k \text{ almost surely.} \tag{5}$$

Let $B = B_1...B_k$. Define the following random variable:

$$Z := \sup_{x_1,...,x_r \in \mathcal{B}} \left\{ \frac{1}{n} \sum_{i \in [n]} \left(\prod_{k=1}^r f_k(a_i^\top x_k) - \mathbb{E}\left[\prod_{k=1}^r f_k(a^\top x_k) \right] \right) \right\}. \tag{6}$$

Then, for any $\lambda > 0$ and for some universal constant C_r, the following large-deviation bound holds:

$$\mathbb{P}\left(Z \geq C_r \sigma_1^r \left(\frac{1}{n} \frac{\lambda + d_{\text{eff}}(r)\ln(d)}{\left(\sigma_1^{-r}B\right)^{2/r-1}} + \frac{\sqrt{\lambda} + \sqrt{d_{\text{eff}}(1)\ln(d)}}{\sqrt{n}} \right) \right) \leq e^{-\lambda}. \tag{7}$$

2.2. Concentration Bound Without Centering

Theorem 2 (Concentration Without Centering) *Let $r \geq 2$ and $d, n \geq 1$ integers. Let $\Sigma \in \mathbb{R}^{d \times d}$ a positive-definite matrix and $a, a_1, ..., a_n$ i.i.d. $\Sigma-$subgaussian random variables (3). Let $d_{\text{eff}}(s), s \in \mathbb{N}^*$ be defined as in (4). Let $f_1, ..., f_r$ be 1-Lipshitz continuous functions on $\mathbb{R}$ such that $f_i(0) = 0$ for $i \in [n]$. For all $k = 1, ..., r$, let $B_k > 0$ such that:*

$$\forall x \in \mathcal{B}, \forall i \in [n], |f_k(a_i^\top x)| \leq B_k \text{ almost surely.}$$

Let $B = B_1...B_k$. Define the following random variable:

$$Y := \sup_{x_1,...,x_r \in \mathcal{B}} \frac{1}{n} \sum_{i \in [n]} \prod_{k=1}^r f_k(a_i^\top x_k). \tag{8}$$

Then, for any $\lambda > 0$ and for some universal constant C_r, the following large-deviation bound holds:

$$\mathbb{P}\left(Y \geq \sigma_1^r C_r \left(1 + \frac{d_{\text{eff}}(r)\ln(d) + \lambda}{n}(\sigma_1^{-r}B)^{1-2/r} \right) \right) \leq e^{-\lambda}. \tag{9}$$

Remark 1 *Assumptions* (5) *in Theorem* 1 *can be replaced by high-probability bounds on the random variables* $a_1, ..., a_n$ *in the following way. If we denote* $R = \sup_{i=1,...,n} \|a_i\|$, *we always have* $B_k \leq R$ *and* $B \leq R^r$ *using Lipshitz continuity of functions* f_k. *Furthermore, a Chernoff bound gives with probability* $1 - \delta$:

$$R^2 \leq 4\sigma_1^2 (2d_{\text{eff}}(1) + \ln(1/\delta) + \ln(n)),$$

yielding, with probability $1 - \delta$, *where* Z *is defined in* (6):

$$Z \leq C_r \sigma_1^r \left(\frac{\ln(\delta^{-1}) + d_{\text{eff}}(r) \ln(d)}{n} \left(d_{\text{eff}}(1) + \ln(\delta^{-1}) + \ln(n) \right)^{\frac{r}{2} - 1} + \frac{\sqrt{\ln(\delta^{-1})} + \sqrt{d_{\text{eff}}(1) \ln(d)}}{\sqrt{n}} \right).$$

The same reasoning applies to Theorem 2 *in the case without centering.*

Remark 2 *Theorems* 1 *and* 2 *assume that the functions* f_k *are 1-Lipshitz and that the supremum is taken over* $\mathcal{B}$ *the centered unit ball. By considering* L_k-*Lipshitz functions and a ball* $\mathcal{B}(x_0, \rho)$, *one obtains the same bound, up to a factor* $\rho L_1 ... L_r$.

Remark 3 *In Appendix* B.4, *we study the tightness of these results. We prove that, for* $f_1 = ... = f_r = I_d$:

$$\begin{cases} \mathbb{E}[Y_{\text{Non-Centered}}] & \geq & C_1 \sigma_1^r \left(1 + \frac{d_{\text{eff}}(1)^{r/2}}{n} \right) \\ \mathbb{E}[Y_{\text{Centered}}^2] & \geq & C_2 \sigma_1^{2r} \frac{C(\Sigma)}{n}. \end{cases}$$

Dependency in terms of n *is thus optimal in both Theorems* 1 $(O(1/\sqrt{n}))$ *and* 2 $(\sigma_1^r (1 + O(1/n)))$. *However, we believe that both the factor* $\ln(d)$ *and having* $d_{\text{eff}}(r)$ *instead of* $d_{\text{eff}}(1)$ *are artifacts of the proof, coming from our non-asymptotic estimates of the metric entropy of ellipsoids (next Section).*

2.3. Concentration of Non-Isotropic Random Tensors

In this section, we provide a concentration bound on the empirical mean of symmetric random tensors of rank 1, involving an effective dimension. In the appendix, we exploit this result to derive some results on the robustness of two-layered neural networks with polynomial activations (Appendix F). Methods such as in Paouris et al. (2017); Bubeck et al. (2020), which do not rely ellipsoids, cannot yield results as sharp as ours, as detailed in Appendix C.

Definition 3 (Tensor) *A tensor of order* $p \in \mathbb{N}^*$ *is an array* $T = (T_{i_1,...,i_p})_{i_1,...,i_p \in [d]} \in \mathbb{R}^{dp}$. T *is said to be of rank* 1 *if it can be written as:*

$$T = u_1 \otimes \cdots \otimes u_p$$

for some $u_1, ..., u_p \in \mathbb{R}^p$.
Scalar product between two tensors of same order p *is defined as:*

$$\langle T, S \rangle = \sum_{i_1,...,i_p} T_{i_1,...,i_p} S_{i_1,...,i_p}, \text{ giving the norm: } \|T\|^2 = \sum_{i_1,...,i_p} T_{i_1,...,i_p}^2.$$

We define the operator norm of a tensor as:

$$\|T\|_{\text{op}} = \sup_{\|x_1 \otimes ... \otimes x_p\| \leq 1} \langle T, x_1 \otimes ... \otimes x_p \rangle.$$

Definition 4 (Symmetric Random Tensor of Rank 1) *A symmetric random tensor of rank 1 and order p is a random tensor of the form:*

$$T = X^{\otimes p}, \tag{10}$$

where $X \in \mathbb{R}^d$ is a random variable. We say that T is Σ-subgaussian if X is a Σ-subgaussian random variable.

We wish to bound the operator norm of tensors of the form $T = \frac{1}{n} \sum_{i=1}^{n} T_i$, where $T_1, ..., T_n$ are *i.i.d.* subgaussian random tensors of rank 1 and order p, using a dependency in an effective dimension rather than the global one. We have:

$$\|T - \mathbb{E}T\|_{\text{op}} = \frac{1}{n} \sup_{x_1,...,x_p \in \mathcal{S}} \sum_{i=1}^{n} \left\{ \prod_{k=1}^{p} \langle a_i, x_k \rangle - \mathbb{E}\left[\prod_{k=1}^{p} \langle a_i, x_k \rangle \right] \right\}.$$

This quantity can be upper-bounded using chaining as in Theorem 1. However, using a simpler argument (Appendix C) inspired by Bubeck et al. (2020) and our bounds on the metric entropy of ellipsoids, we have the following.

Theorem 3 (Non-Isotropic Concentration Bound on Random Tensors) *Let $T_1, ..., T_n$ be* i.i.d. *random tensors of order p, rank 1, symmetric and Σ-subgaussian. Let $T = \frac{1}{n} \sum_{i=1}^{n} T_i$. With probability $1 - \delta$ for any $\delta > 0$ and universal constant $C_p > 0$, we have:*

$$\|T - \mathbb{E}T\|_{\text{op}} \le C_p \sigma_1^p \sqrt{\frac{d_{\text{eff}}(1) + \ln(d) + \ln(\delta^{-1})}{n}}.$$

Equivalently, for any $\lambda > 0$:

$$\mathbb{P}\left(\|T - \mathbb{E}T\|_{\text{op}} \ge C_p \sigma_1^p \sqrt{\frac{d_{\text{eff}}(1) + \ln(d) + \lambda}{n}} \right) \le e^{-\lambda}.$$

We next present the upper-bounds on the number of balls needed to cover an ellipsoid we use to prove the concentration bounds with chaining (Theorems 1 and 2) or without (Theorem 3). We believe these technical results to be of independent interest due to the strong link between metric entropy and uniform concentration bounds.

3. Results on Covering of Balls with Ellipsoids and Metric Entropy

3.1. Metric Entropy of an Ellipsoid

Definition 5 (Ellipsoid and ε-Entropy) *Given a vector $b = (b_1, \ldots, b_d)$ with $b_1 \ge \cdots \ge b_d > 0$, the ellipsoid E_b is defined as*

$$E_b = \left\{ x \in \mathbb{R}^d : \sum_{i \in [d]} \frac{x_i^2}{b_i^2} \le 1 \right\}.$$

The ε-entropy $\mathcal{H}_\varepsilon(E_b)$ of ellipsoid E_b is the logarithm of the size of a minimal ε-covering (or ε-net in information theory terminology) of E_b. More formally:

$$\mathcal{H}_\varepsilon(E_b) = \ln\left(\min\left\{|A| : A \subset \mathbb{R}^d, E_b \subset \bigcup_{x \in A} \mathcal{B}(x, \varepsilon)\right\}\right), \tag{11}$$

where $\mathcal{B}(x, \varepsilon)$ is the Euclidean ball of radius ε. The unit entropy is the ε-entropy for $\varepsilon = 1$.

Given an ellipsoid E_b, define the following quantities:

$$K_b = \sum_{i=1}^{m_b} \ln(b_i) \text{ and } m_b = \sum_{i \in [d]} \mathbf{1}_{b_i > 1}. \tag{12}$$

Provided that:

$$\ln(b_1) = o\left(\frac{K_b^2}{m_b \ln(d)}\right), \tag{13}$$

[Dumer et al. (2004)](#) (Theorem 2 in their article) prove the following asymptotic equivalent of $\mathcal{H}_1(E_b)$ when $d \to \infty$:

$$\mathcal{H}_1(E_b) \sim K_b. \tag{14}$$

However, we need non-asymptotic bounds on $\mathcal{H}_1(E_b)$. Using techniques introduced in [Dumer et al. (2004)](#), we thus establish Theorem 4, whose proof appears in Appendix A, together with an extension to ellipsoids in infinite dimension.

Theorem 4 (Unit Entropy of an Ellipsoid in Fixed Dimension) *One has, for some universal constant $c > 0$, the following bound on the unit entropy of ellipsoid E_b:*

$$\mathcal{H}_1(E_b) \leq K_b + c\left[\ln(d) + \sqrt{\ln(b_1)m_b \ln(d)}\right].$$

This theorem gives the following corollary, bounding the number of ellipsoids required to cover the unit ball, directly linked with the number of balls required to cover an ellipsoid thanks to a linear transformation.

3.2. Coverings of the Unit Ball With Ellipsoids

Corollary 1 *Let $\varepsilon > 0$. Let random vector $a \in \mathbb{R}^d$ satisfy subgaussian tail assumption (3) for matrix Σ, with spectrum $\sigma_1^2 \geq \cdots \geq \sigma_d^2 > 0$. Then there exists a collection $\mathcal{N}_\varepsilon$ of vectors in $\mathcal{S}_1$ the unit sphere of $\mathbb{R}^d$ such that, for all $x \in \mathcal{S}_1$, there exists $y = \Pi_\varepsilon x \in \mathcal{N}_\varepsilon$ such that*

$$\|x - y\|_\Sigma^2 := (x - y)^\top \Sigma (x - y) \leq \varepsilon^2 \sigma_1^2, \tag{15}$$

and the covering $\mathcal{N}_\varepsilon$ verifies

$$\ln(|\mathcal{N}_\varepsilon|) \leq \mathcal{H}_\varepsilon := \sum_{i=1}^{m_\varepsilon} \ln\left(\frac{\sigma_i}{\varepsilon \sigma_1}\right) + c\left[\ln(d) + \sqrt{\ln(\varepsilon^{-1})\ln(d)m_\varepsilon}\right], \tag{16}$$

where

$$m_\varepsilon = \sum_{i=1}^{d} \mathbf{1}_{\sigma_i > \varepsilon \sigma_1} \tag{17}$$

and c is some universal constant. Furthermore, we have:

$$\mathcal{H}_\varepsilon \leq \ln(\varepsilon^{-1}) + \frac{\min\left(d-1, \varepsilon^{-\frac{2}{r}\frac{d_{\text{eff}}(r)-1}{e}}\right)}{2/r} \ln\left(\max(e, \varepsilon^{-\frac{2}{r}\frac{d_{\text{eff}}(r)-1}{d-1}})\right) + c\left[\ln(d) + \sqrt{\ln(\varepsilon^{-1})\ln(d)m_\varepsilon}\right],$$

and

$$m_\varepsilon \leq 1 + (d_{\text{eff}}(r)-1)\varepsilon^{-\frac{2}{r}}.$$

This last bound on $\mathcal{H}_\varepsilon$ is a core technical lemma behind Theorems 1 and 2 . It is to be noted that $\mathcal{H}_\varepsilon$ is not linear in an effective dimension. Indeed, for $\varepsilon \leq C_r \left(\frac{d-1}{d_{\text{eff}}(r)-1}\right)^{r/2}$, our expression is linear in d. This difficulty is the non-asymptotic equivalent of Dumer et al. (2004)'s assumption in (13).

3.3. Ellipsoids in Infinite Dimension

We here define ellipsoids in infinite dimension and upper-bound asymptotically their ε-entropy in terms of *spectral dimension*. Although not used in the applications described in the present article, uniform concentration of infinite-dimensional random vectors that satisfy an infinite-dimensional subgaussian property require results such as the one we provide below.

Let $\mathcal{V}$ be a separable real Hilbert space (*e.g.* $\mathbb{R}^\mathbb{N}$, $\ell^2([0,1])$).

Definition 6 (Ellipsoids in Hilbert Spaces) *Let A a self-adjoint and semi-definite positive operator on $\mathcal{V}$ i.e. such that $\forall(x,y) \in \mathcal{V}^2$, we have $\langle A(x), y\rangle = \langle x, A(y)\rangle \geq 0$. We define the ellipsoid $E_A \subset \mathcal{V}$ by:*

$$E_A = \left\{x \in \mathcal{V} : \left\|A^\dagger(x)\right\|^2 \leq 1\right\},$$

where $A^\dagger$ is the pseudo-inverse of A.

This notion generalizes Definition 5: taking $\mathcal{V} = \mathbb{R}^d$ and $A = \text{Diag}(b_1, ..., b_d)$, we have $E_A = E_b$. We next define the spectral dimension of an ellipsoid. We recall that if A is a self-adjoint and semi-definite positive operator on $\mathcal{V}$, there exists a Hilbert basis of eigenvectors of A, and the eigenvalues of A are non-negative.

Definition 7 (Spectral Dimension and Effective Dimension) *Let E_A an ellipsoid in $\mathcal{V}$, where A is a self-adjoint and semi-definite positive operator. Assume that the eigenvalues of A can be ordered as a decreasing sequence $(b_i)_{i\in\mathbb{N}^*}$. E_A is of spectral dimension $d \in \mathbb{R}^{+,*}$ if $\sum_{i\in\mathbb{N}^*} b_i^2 < \infty$ and when $n \to \infty$:*

$$\sum_{i \geq n+1} b_i^2 = \mathcal{O}(n^{-\frac{2}{d}}).$$

The effective dimension of ellipsoid E_A is then $\sum_{i\in\mathbb{N}^*} b_i^2$.

The right notion of dimension for the control of metric entropy in infinite dimension is the spectral dimension, as shown in the next proposition: the ε-entropy of an ellipsoid scales as the spectral dimension in infinite dimension.

Proposition 1 *Let E_A be an ellipsoid in $\mathcal{V}$, of spectral dimension $d > 0$. We have, when $\varepsilon \to 0$:*

$$\mathcal{H}_\varepsilon(E_A) \leq d\ln\left(\varepsilon^{-1}\right)^2\left(1 + o(1)\right),$$

where $\mathcal{H}_\varepsilon(E_A)$ is the number (possibly infinite) of balls of radius ε required to cover E_A.

4. Statistical Preconditioning: Bounding Relative Condition Numbers

In this section, we present an application of Theorem 1 to optimization. Essentially, we show that statistical preconditioning-based optimization automatically benefits from low effective dimension in the data, thus proving a conjecture made in Hendrikx et al. (2020).

4.1. Large Deviation of Hessians

Let f be a convex function defined on $\mathbb{R}^d$. We assume that the following holds, which is true for logistic or ridge regressions (Appendix D.4).

Assumption 1 (Empirical Risk Structure) *Let $\ell : \mathbb{R} \to \mathbb{R}$ convex, twice differentiable such that ℓ'' is $\|\ell''\|_{\mathrm{Lip}}$-Lipschitz. Let $n \in \mathbb{N}^*$, some convex functions $\ell_j : \mathbb{R} \to \mathbb{R}, j \in [n]$ such that $\forall j \in [n], \ell_j'' = \ell''$ and i.i.d. Σ-subgaussian random variables $(a_j)_{j \in [n]}$. We assume that:*

$$\forall x \in \mathbb{R}^d, f(x) = \frac{1}{n} \sum_{j=1}^{n} \ell_j(a_j^\top x). \tag{18}$$

Proposition 2 *Denote H_x the Hessian of f at some point $x \in \mathbb{R}^d$ and $\bar{H}_x$ its mean. We have:*

$$H_x = \frac{1}{n} \sum_{i=1}^{n} \ell''(a_i^\top x) a_i a_i^\top, \text{ and } \bar{H}_x = \mathbb{E}_a \left[\ell''(a_1^\top x) a_1 a_1^\top \right].$$

Let:

$$Z = \sup_{\|x\| \leq 1} \left\| H_x - \bar{H}_x \right\|_{\mathrm{op}}. \tag{19}$$

With probability $1 - \delta$, we have, with C a universal constant:

$$Z \leq C \sigma_1^3 \|\ell''\|_{\mathrm{Lip}} \left(\frac{(d_{\mathrm{eff}}(3)\ln(d) + \ln(1/\delta))\sqrt{d_{\mathrm{eff}}(1) + \ln(n/\delta)}}{n} + \frac{\sqrt{\ln(1/\delta)} + \sqrt{d_{\mathrm{eff}}(1)\ln(d)}}{\sqrt{n}} \right).$$

Previous works (Hendrikx et al., 2020) obtained:

$$C' \sigma_1^3 \|\ell''\|_{\mathrm{Lip}} \frac{(d + \ln(1/\delta))\sqrt{d_{\mathrm{eff}}(1) + \ln(n/\delta)}}{\sqrt{n}} \left[\frac{1}{\sqrt{d}} + \frac{1}{\sqrt{n}} \right]. \tag{20}$$

In order for this bound to be of order 1, n was required to be of order the whole dimension d, while we only need n to be of order $d_{\mathrm{eff}}(3)$.

4.2. Statistical Preconditioning

Consider the following optimization problem:

$$\min_{x \in \mathbb{R}^d} \Phi(x) := F(x) + \psi(x), \tag{21}$$

where $F(x) = \frac{1}{n} \sum_{j=1}^{n} f_j(x)$ has a finite sum structure and ψ is a convex regularization function. Standard assumptions are the following:

$$\forall x, \sigma_F I_d \leq \nabla^2 F(x) \leq L_F I_d. \tag{22}$$

9

We focus on a basic setting of distributed optimization. At each iteration $t = 0, 1, ...,$ the server broadcasts the parameter x_t to all workers $j \in \{1, ..., n\}$. Each machine j then computes in parallel $\nabla f_j(x_t)$ and sends it back to the server, who finally aggregates the gradients to form $\nabla F(x_t) = \frac{1}{n} \sum_j \nabla f_j(x_t)$ and use it to update x_t in the following way, using a standard proximal gradient descent, for some parameter $\eta_t \leq 1/L_F$:

$$x_{t+1} \in \underset{x \in \mathbb{R}^d}{\arg\min} \left\{ \langle \nabla F(x_t), x \rangle + \psi(x) + \frac{1}{2\eta_t} \|x - x_t\|^2 \right\}. \tag{23}$$

Setting $\eta_t = 1/L_F$ yields linear convergence:

$$\Phi(x_t) - \Phi(x^*) \leq L_F (1 - \kappa_F^{-1})^t \|x_0 - x^*\|^2. \tag{24}$$

In general, using an accelerated version of (23), one obtains a communication complexity (*i.e.* number of steps required to reach a precision $\varepsilon > 0$) of $O(\kappa_F^{1/2} \ln(1/\varepsilon))$ (where $\kappa_F = \frac{L_F}{\sigma_F}$) that cannot be improved in general. Statistical preconditioning is then a technique to improve each iteration's efficiency, based on the following insight: considering *i.i.d.* datasets leads to statistically similar local gradients ∇f_j. The essential tool for preconditioning is the Bregman divergence.

Definition 8 (Bregman divergence and Relative Smoothness) *For a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, we define D_ϕ its Bregman divergence by:*

$$\forall x, y \in \mathbb{R}^d, D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle. \tag{25}$$

For convex functions $\phi, F : \mathbb{R}^d \to \mathbb{R}$, we say that F is relatively $L_{F/\phi}$-smooth and $\sigma_{F/\phi}$-strongly-convex if, for all $x, y \in \mathbb{R}^d$:

$$\sigma_{F/\phi} D_\phi(x, y) \leq D_F(x, y) \leq L_{F/\phi} D_\phi(x, y), \tag{26}$$

or equivalently:

$$\sigma_{F/\phi} \nabla^2 \phi(x) \leq \nabla^2 F(x) \leq L_{F/\phi} \nabla^2 \phi(x), \tag{27}$$

We consequently define $\kappa_{F/\phi} = \frac{L_{F/\phi}}{\sigma_{F/\phi}}$ the relative condition number of F with respect to ϕ.

Taking $\phi = \frac{1}{2} \|.\|^2$ gives $D_\phi = \frac{1}{2} \|.\|^2$ and thus yields classical smoothness and strong-convexity definitions. The idea of preconditioning is then to replace $\frac{1}{2\eta_t} \|x - x_t\|^2$ in (23) by $D_\phi(x, y)$ for a *convenient* function ϕ which the server has access to, leading to:

$$x_{t+1} \in \underset{x \in \mathbb{R}^d}{\arg\min} \left\{ \langle \nabla F(x_t), x \rangle + \psi(x) + \frac{1}{\eta_t} D_\phi(x, x_t) \right\}. \tag{28}$$

With $\eta_t = 1/L_{F/\phi}$, the sequence generated by (28) satisfies:

$$\Phi(x_t) - \Phi(x^*) \leq L_{F/\phi} (1 - \kappa_{F/\phi}^{-1})^t. \tag{29}$$

Hence, the effectiveness of preconditioning hinges on how smaller $\kappa_{F/\phi}$ is compared to κ_F. Next subsection presents how our large deviation bound of Hessians (Proposition 2) comes into place. The better ϕ approximates F, the smaller $\kappa_{F/\phi}$ and the more efficient each iteration of (28) is.

4.3. Main Results in Statistical Preconditioning

We furthermore assume that $F(x) = f(x) + \frac{\lambda}{2}\|x\|^2$ where f verifies Assumption 1 and $\lambda > 0$. Assume that the server has access to an *i.i.d.* sample $\tilde{a}_1, ..., \tilde{a}_N$ of the same law as the a_j's and to functions $\tilde{\ell}_1, ..., \tilde{\ell}_N$ such that $\tilde{\ell}_i'' = \ell''$. Define $\tilde{f}(x) = \frac{\lambda}{2}\|x\|^2 + \frac{1}{N}\sum_{i=1}^{N}\tilde{\ell}_i(a_i^\top x)$. The preconditioner ϕ is chosen as, for some $\mu > 0$:

$$\phi(x) = \frac{\lambda}{2}\|x\|^2 + \frac{1}{N}\sum_{i=1}^{N}\tilde{\ell}_i(\tilde{a}_i^\top x) + \frac{\mu}{2}\|x\|^2, \tag{30}$$

Parameter $\mu > 0$ is chosen such that, with high probability:

$$\forall x \in \mathrm{Dom}_\psi, \left\|\nabla^2\tilde{f}(x) - \nabla^2 F(x)\right\|_{\mathrm{op}} \leq \mu. \tag{31}$$

For such a $\mu > 0$, we have: $L_{F/\phi} \leq 1$, $\sigma_{F/\phi} \geq (1 + 2\mu/\lambda)^{-1}$ and $\kappa_{F/\phi} \leq 1 + \frac{2\mu}{\lambda}$. Recall that for $t = 0, 1, 2, ...$, we have $\|x_t - x^*\|^2 \leq C(1 - \kappa_{F/\phi})^t$.

Proposition 3 (Statistical Preconditioning: Non-Isotropic Results) *Assume that for all $x \in \mathrm{Dom}_\psi$, $\|x\| \leq R$. Under Assumption 1, with probability $1 - \delta$, we have:*

$$\sup_{\|x\|\leq R}\left\|\nabla^2\tilde{f}(x) - \nabla^2 F(x)\right\| \leq CR\sigma_1^3\|\ell''\|_{\mathrm{Lip}}\left(\frac{(d_{\mathrm{eff}}(3)\ln(d) + \ln(1/\delta))\sqrt{d_{\mathrm{eff}}(1) + \ln(n/\delta)}}{n}\right.$$
$$\left. + \frac{\sqrt{\ln(1/\delta)} + \sqrt{d_{\mathrm{eff}}(1)\ln(d)}}{\sqrt{n}}\right).$$

If μ is taken as this upper bound, then we control the rate of convergence in (29) with:

$$\kappa_{F/\phi} = 1 + \tilde{O}\left\{\frac{R\sigma_1^3\|\ell''\|_{\mathrm{Lip}}}{\lambda}\max\left(\frac{\sqrt{d_{\mathrm{eff}}(1)}}{\sqrt{n}}, \frac{\sqrt{d_{\mathrm{eff}}(1)}d_{\mathrm{eff}}(3)}{n}\right)\right\}, \tag{32}$$

where $\tilde{O}$ hides logarithmic factors in d, n and δ^{-1}.

Contrast this with known results:

Remark 4 (Statistical Preconditioning: Isotropic Results) *Still under Assumption 1, Hendrikx et al. (2020) obtained:*

$$\kappa_{F/\phi} = 1 + \tilde{O}\left\{\frac{R\sigma_1^3\|\ell''\|_{\mathrm{Lip}}}{\lambda}\max\left(\sqrt{\frac{d}{n}}, \frac{d^{3/2}}{n}\right)\right\}. \tag{33}$$

The only parameter required is an upper-bound on $d_{\mathrm{eff}}(3)$ and $d_{\mathrm{eff}}(1)$ in order to tune μ. Simply knowing that data are distributed according to a highly non-isotropic subgaussian law can thus improve the efficiency of statistical preconditioning, by decreasing drastically estimates of $\kappa_{F/\phi}$ and the number of samples required in the preconditioning function.

5. Non-Isotropic Randomized Smoothing

5.1. General Considerations on the Randomized Smoothing Technique

Consider an objective function $f : \mathbb{R}^d \to \mathbb{R}$ and a known convex regularizer ψ. f is assumed to be convex and L-Lispchitz for some $L > 0$. We assume that $\mathrm{Dom}_\psi \subset \mathcal{B}(0, R)$. The following minimization problem:

$$\min_{x \in \mathrm{Dom}_\psi} \Phi(x) := f(x) + \psi(x) \tag{34}$$

is potentially hard as f is not necessarily smooth. Moreover, f is assumed to be of the form:

$$\forall x \in \mathbb{R}^d, f(x) = \mathbb{E}_a[F(x, a)], \tag{35}$$

for some random variable a and F a convex function, Lipschitz in its first variable. The second difficulty is thus that f may not be directly computable, and a stochastic framework is required.

Principle of the randomized smoothing technique and description of the algorithm: in order to both use acceleration techniques and stochasticity of the gradients, the objective function f is approximated by a smoothed version f^γ, where $\gamma > 0$ is a parameter of the algorithm:

$$\forall x \in \mathbb{R}^d, f^\gamma(x) = \mathbb{E}_Z[f(x + \gamma Z)] \tag{36}$$

where Z is a random variable, following a *smoothing ditribution* μ. Scaman et al. (2018) consider isotropic gaussians ($\mu = \mathcal{N}(0, I_d)$), while Duchi et al. (2012) consider more general smoothing distributions (encompassing uniform distributions on the euclidean ball or on the ℓ^1-ball). The algorithm then:

1. Draws $Z_{1,t}, ..., Z_{m,t}$ *i.i.d.* random variables according to the smoothing distribution μ, for m a fixed integer.

2. Queries the oracle at the m points $y_t + u_t Z_{i,t}, i = 1, ..., m$, yielding stochastic gradients $g_{i,t} \in \partial F(y_t + u_t Z_{i,t}, a_{i,t})$, where y_t is the query point.

3. Computes the average $g_t = \frac{1}{m} \sum_{i=1}^m g_{i,t}$.

4. Uses this estimated gradient to perform an accelerated stochastic and proximal gradient step.

For brevity, precise formulations of the algorithm and in particular of that last point are deferred to Appendix E.

5.2. Isotropic Randomized Smoothing

We restrict ourselves to gaussian smoothing distributions μ. In the isotropic case $\mu = \mathcal{N}(0, I_d)$ considered by Duchi et al. (2012); Scaman et al. (2018), the following crucial property holds, leading to a trade-off between precision and the smoothness parameter of f^γ.

Proposition 4 (Properties of Isotropic Gaussian Smoothing) *Let $\gamma > 0$ and assume that $\mu = \mathcal{N}(0, I_d)$. Recall that $f^\gamma(x) = \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)}[f(x + \gamma Z)]$ and f is L-Lipschitz. We have:*

$$\forall x \in \mathbb{R}^d, f(x) \leq f^\gamma(x) \leq f(x) + \gamma L \sqrt{d}, \tag{37}$$

and f^γ is $\frac{L}{\gamma}$-smooth. In order to reach an $\varepsilon > 0$ precision, one can take $\gamma = \frac{\varepsilon}{L\sqrt{d}}$, for which f^γ is then $\frac{L^2 \sqrt{d}}{\varepsilon}$-smooth.

Proposition 5 (Convergence Guarantees with Isotropic Smoothing) *Take $\mu = \mathcal{N}(0, I_d)$ for the smoothing distribution. For a smoothing parameter $\gamma = Rd^{-1/4}$ and varying stepsizes in the accelerated gradient descent (Appendix E), we have:*

$$\mathbb{E}[f(x_T) + \psi(x_T) - f(x^*) - \psi(x^*)] \leq \frac{10LRd^{1/4}}{T} + \frac{5LR}{\sqrt{Tm}}, \tag{38}$$

and this $d^{1/4}$ factor cannot be improved: there exist objective functions f and dimension-free constants such that we have an effective $d^{1/4}$ dependency in the global dimension.

For m big enough, the dominant term is $O(LRd^{1/4}/T)$, a dimensional dependency that cannot be alleviated (Nemirovsky and Yudin, 1985; Duchi et al., 2012; Bubeck et al., 2019).

5.3. Non-Isotropic Randomized Smoothing

In order to improve over Proposition 5, as it is optimal on the class of Lipschitz functions of the form (35), more assumptions are required in order to take advantage of an enventual underlying small effective dimension. We restrict ourselves to empirical measures of subgaussian random variables for ν in (35) and to an empirical risk assumption for linear models such as in Assumption 1. We will hence assume that:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(a_i^\top x), \tag{39}$$

for convex functions ℓ_i, and Σ-subgaussian random variables a_i. We furthermore assume that each ℓ_i is L_ℓ-Lispchitz. Our interest in empirical measures lies in the fact that in practice one does not have access to an infinite number of samples. Our assumptions encompass non-smooth losses, such as $\ell_i(x) = \max(0, a_i^\top x - b_i)$. As in Proposition 3, one can hope to replace the $d^{1/4}$ factor in Proposition 5 by an effective dimension dependent factor. A non-isotropic analog of Proposition 4 for a smoothing distribution μ of the form $\mathcal{N}(0, \Sigma')$ is required. It is quite intuitive to conjecture that adapting the smoothing distribution to the distribution of the data should indeed improve the efficiency of the algorithm. An analysis in the appendix shows that an optimal Σ' is $\sqrt{\Sigma}$, hence the following proposition.

Proposition 6 (Properties of Non-Isotropic Gaussian Smoothing) *Set $\mu = \mathcal{N}(0, \sqrt{\Sigma})$, $\gamma > 0$. We have, with probability $1 - \delta$:*

$$\begin{aligned}
\forall x \in \mathbb{R}^d, f^\gamma(x) &\leq & f(x) + \gamma L_\ell \sqrt{\sigma_1^3(d_{\text{eff}}(1) + \ln(n\delta^{-1}))d_{\text{eff}}(2)} \\
\|\nabla f^\gamma\|_{\text{Lip}} &\leq & \frac{L_\ell \sigma_1^{1/2} d_{\text{eff}}(2)^{1/2}}{\gamma d_{\text{eff}}(1)} \left(1 + C\sqrt{\frac{d_{\text{eff}}(1)\ln(d) + \ln(\delta^{-1})}{N}}\right).
\end{aligned} \tag{40}$$

In order to reach an $\varepsilon > 0$ precision, one can take $\gamma = \frac{\varepsilon}{L_\ell \sqrt{\sigma_1^3 d_{\text{eff}}(1) d_{\text{eff}}(2)}}$, for which f^γ is then $\frac{\sigma_1^2 L_\ell^2 d_{\text{eff}}(2)}{\varepsilon \sqrt{d_{\text{eff}}(1)}}(1 + C\sqrt{\frac{d_{\text{eff}}(1)\ln(d) + \ln(\delta^{-1})}{N}})$-smooth with probability $1 - \delta$.

Proposition 7 (Convergence Guarantees with Non-Isotropic Smoothing) *Taking $\mu = \mathcal{N}(0, \sqrt{\Sigma})$, for time-varying stepsizes defined in Appendix E, we have with probability $1 - \delta$ conditionally on*

the random variables $a_i, \tilde{a}_j$:

$$\mathbb{E}[f(x_T) + \psi(x_T) - f(x^*) - \psi(x^*)] \leq \tilde{O}\left(\frac{LR\sigma_1\sqrt{d_{\text{eff}}(2)/\sqrt{d_{\text{eff}}(1)}}}{T}\right) + \frac{5L_0 R}{\sqrt{T}m}, \qquad (41)$$

where $\tilde{O}$ hides logarithmic factors in d and δ^{-1}.

(40) corresponds to (38), with d replaced by $\frac{d_{\text{eff}}(2)^2}{d_{\text{eff}}(1)}$. Taking advantage of the underlying geometric repartition of the data thus yields better convergence guarantees, if we assume a more restrictive structure on the objective function. The knowledge of Σ is here required to apply the previous considerations, whereas in the previous section only $\text{Tr}(\Sigma)$ is needed. One may wonder to what extent our assumptions on f could be generalized in order to obtain similar results.

6. Conclusion

Achieving *effective dimension*-dependent bounds thus yields several applications, and we believe many others than the ones we studied exist. Broadening the set of applications could be achieved by: considering more general random variables, other models of effective dimension such as spectral dimension (Durhuus, 2009) or doubling dimension (Karbasi et al., 2012), and infinite dimension d but finite effective dimension such as in Appendix A.3 in order to take into account functional spaces for instance. Also, efficient methods for testing Σ-subgaussianity do not seem to exist, which should be an interesting problem to tackle.

References

Stéphane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration inequalities : a non asymptotic theory of independence*. Oxford University Press, 2013. URL https://hal.inria.fr/hal-00942704.

Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Complexity of highly parallel non-smooth convex optimization. *arXiv preprint arXiv:1906.10655*, 2019.

Sébastien Bubeck, Yuanzhi Li, and Dheeraj Nagaraj. A law of robustness for two-layers neural networks. *arXiv preprint arXiv:2009.14444*, 2020.

David L. Donoho, Matan Gavish, and Iain M. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851*, 2017.

Radu-Alexandru Dragomir, Adrien Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of bregman first-order methods. *arXiv preprint arXiv:1911.08510*, 2019.

Radu-Alexandru Dragomir, Mathieu Even, and Hadrien Hendrikx. Fast stochastic bregman gradient methods: Sharp analysis and variance reduction. *arXiv preprint arXiv:2104.09813*, 2021.

John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized smoothing for stochastic optimization. *arXiv preprint arXiv:1103.4296*, 2012.

R.M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967. ISSN 0022-1236. doi: https://doi.org/10.1016/0022-1236(67)90017-1. URL http://www.sciencedirect.com/science/article/pii/0022123667900171.

I. Dumer, M. Pinsker, and V. V. Prelov. On coverings of ellipsoids in euclidean spaces. *IEEE Transactions on Information Theory*, 50:2348–2356, 2004.

Bergfinnur Durhuus. Hausdorff and spectral dimension of infinite random graphs. *Acta Phys. Polon. B*, 40:3509–3532, 2009.

Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. 2020.

Amin Karbasi, Stratis Ioannidis, and laurent Massoulie. Comparison-based learning with rank nets. *arXiv preprint arXiv:1206.4674*, 2012.

V A Marčenko and L A Pastur. DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES. *Mathematics of the USSR-Sbornik*, 1(4):457–483, apr 1967. doi: 10.1070/sm1967v001n04abeh001994. URL https://doi.org/10.1070/sm1967v001n04abeh001994.

Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *arXiv preprint arXiv:1112.5448*, 2017.

A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin). *SIAM Review*, 27(2):264–265, 1985. doi: 10.1137/1027074. URL https://doi.org/10.1137/1027074.

Grigoris Paouris, Petros Valettas, and Joel Zinn. Random version of Dvoretzky's theorem in lpn. *Stochastic Processes and their Applications*, 127(10):3187–3227, 2017. doi: 10.1016/j.spa.2017.02.007. URL https://ideas.repec.org/a/eee/spapps/v127y2017i10p3187-3227.html.

C. A. Rogers. Covering a sphere with spheres. *Mathematika*, 10(2):157–164, 1963. doi: 10.1112/S0025579300004083.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for non-smooth distributed optimization in networks. *arXiv preprint arXiv:1806.00291*, 2018.

Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication efficient distributed optimization using an approximate newton-type method. *arXiv preprint arXiv:1312.7853*, 2014.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, Aug 2011. ISSN 1615-3383. doi: 10.1007/s10208-011-9099-z. URL http://dx.doi.org/10.1007/s10208-011-9099-z.

Joel A. Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Harald Uhlig. On singular wishart and singular multivariate beta distributions. *Ann. Statist.*, 22(1):395–405, 03 1994. doi: 10.1214/aos/1176325375. URL https://doi.org/10.1214/aos/1176325375.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2011.

Roman Vershynin. Estimation in high dimensions: a geometric perspective. *arXiv preprint arXiv:1405.5103*, 2014.

Roman Vershynin. Concentration inequalities for random tensors. *arXiv preprint arXiv:1905.00802*, 2020.

John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2):32–52, 1928. ISSN 00063444. URL http://www.jstor.org/stable/2331939.

Yiqiao Zhong. Eigenvector under random perturbation: A nonasymptotic rayleigh-schrödinger theory. *arXiv preprint arXiv:1702.00139*, 2017.

Summary of the Article and of the Appendix

Appendix A. Covering Ellipsoids with Balls

A.1. Proof of Theorem 4

Dumer et al. (2004) prove the asymptotic version of our result. We use their method in order to prove Theorem 4 in what follows.

Proof The proof involves three steps. In the first one, we cover ellipsoid E_b by direct products of balls of lesser dimensions. Then, in Step 2, we derive a general upper bound. Finally, tuning our parameters from the bound obtained in Step 2 leads to the desired result in Step 3. Steps 1 and 2 of the proof are directly inspired by Dumer et al. (2004), while step 3 and the corollaries that follow are the non-asymptotic improvements over their results.

Step 1. Let $t \in \mathbb{N}, t \leq d, 0 = n_0 < n_1 < ... < n_t = d$ and $I_i = \{n_{i-1}+1, ..., n_i\}$ for $i = 1, ..., t$, in order to divide $[d]$ into t subsets. Let $s_i = n_i - n_{i-1}$ for $i = 1, ..., t$. For some parameter $h \in (0, 1)$ let the set of numbers:

$$H = \{ih, i = 1, ..., \lfloor h^{-1} \rfloor + 1\}.$$

For any $w \in [0, 1]$, let $\bar{w}$ be the closest point in H exceeding w. Consider the following subset of H^t:

$$U = \{(u_1, ..., u_t) \in H^t | \sum_{i=1}^{t} u_i \leq 1 + th\}$$

Let $u \in U$ be fixed. For $i = 1, ..., t$, consider the ball of dimension s_i:

$$B_i^u = \left\{ x \in \mathbb{R}^{I_i} | \sum_{j \in I_i} x_j^2 \leq \rho_i^2 \right\}, \text{ where } \rho_i^2 = u_i b_{n_{i-1}+1}^2.$$

Let D_u be the direct product of all t balls:

$$D_u = \prod_{i=1}^{t} B_i^u = \left\{ x \in \mathbb{R}^d | \sum_{j \in I_i} x_j^2 / b_{n_{i-1}+1}^2 \leq u_i, i = 1, ..., t \right\}.$$

We have:

$$E_b \subset \bigcup_{u \in U} D_u.$$

Indeed, for $x \in E_b$, let $w_i = \sum_{j \in I_i} x_j^2 / b_j^2$ and take $u_i = \bar{w}_i$. First, $x_{I_i} \subset B_i$:

$$\sum_{j \in I_i} x_j^2 / b_{n_{i-1}+1}^2 \leq \sum_{j \in I_i} x_j^2 / b_j^2 = w_i \leq u_i.$$

Moreover, $u \in U$:

$$\sum_{i=1}^{t} u_i \leq \sum_{i=1}^{t} w_i + h \leq 1 + th.$$

Hence, for any $x \in E_b$, there exists $u \in U$ such that $x \in D_u$.

Step 2. Given D_u for some $u \in U$, denote $\mathcal{H}_1(D_u)$ its unit entropy. We have:

$$\mathcal{H}_1(D_u) = \inf_{e \in (R^{+*})^t : \sum_i e_i^2 \leq 1} \sum_{i=1}^{t} \mathcal{H}_{e_i}(B_i^u),$$

where $\mathcal{H}_{e_i}(B_i^u)$ is the e_i-entropy of B_i^u. As $E_b \subset \bigcup_{u \in U} D_u$, we have that:

$$\mathcal{H}_1(E_b) \leq \ln(|U|) + \sup_{u \in U} \left(\inf_{e \in (R^{+*})^t : \sum_i e_i^2 \leq 1} \sum_{i=1}^{t} \mathcal{H}_{e_i}(B_i^u) \right).$$

We have:

$$|U| = \binom{t + \lfloor h^{-1} \rfloor}{t} := \mathcal{N}(t, h).$$

In order to estimate quantities such as $\mathcal{H}_{e_i}(B_i^u)$, we will need results on the ε-entropies of balls that directly come from Rogers (1963):

Lemma 5 *For any dimension $d > 0$, any ball B_ρ of radius $\rho > 0$ has a unit entropy $\mathcal{H}_1(B_\rho)$ upper-bounded by:*

$$\mathcal{H}_1(B_\rho) \le n\ln(\rho) + c\ln(n+1), \tag{42}$$

for some universal constant $c > 0$.

∎

Using this, we obtain:

$$\mathcal{H}_1(E_b) \le \ln(\mathcal{N}(t,h)) + \sup_{u \in U}\left(\inf_{e \in (R^{+*})^t : \sum_i e_i^2 \le 1} \sum_{i=1}^{t} s_i \ln(\rho_i/e_i) + c\ln(s_i + 1)\right).$$

As $\rho_i^2 = u_i b_{n_{i-1}+1}^2$, for fixed $u \in U$ we have:

$$\inf_{e \in (R^{+*})^t : \sum_i e_i^2 \le 1} \sum_{i=1}^{t} s_i \ln(\rho_i/e_i) = \inf_{e \in (R^{+*})^t : \sum_i \varepsilon_i \le 1} \frac{1}{2}\sum_{i=1}^{t} s_i \ln(u_i b_{n_{i-1}+1}^2/\varepsilon_i)$$

$$\le \frac{1}{2}\sum_{i=1}^{t} s_i \ln(u_i) + \sum_{i=1}^{t} s_i \ln^+(b_i)$$

$$\le \ln(\gamma)\sum_{i=1}^{t} s_i + \sum_{i=1}^{t} s_i \ln^+(b_i),$$

where we note $\gamma = \sqrt{1 + th}$. Now consider $\hat{b} \in \mathbb{R}^d$ the vector with coefficients:

$$\hat{b}_j = b_{n_{i-1}+1}, j \in I_i, i = 1, ..., t,$$

such that:

$$\sum_{i=1}^{t} s_i \ln^+(b_i) = K_{\hat{b}}.$$

Furthermore, comparing K_b and $K_{\hat{b}}$:

$$\sum_{i=1}^{t} s_i \ln^+(b_i) \le \sum_{j=1}^{d} \ln^+(b_j) + \sum_{i=1}^{t-1}\sum_{j \in I_i} \ln(b_{n_{i-1}+1}/b_i)$$

$$\le K_b + \sum_{i=1}^{t-1}(s_i - 1)\ln(b_{n_{i-1}+1}/b_{n_i}).$$

The sum above ends at $t-1$ by definition of m and of the interval I_t. We hence obtain the following general upper-bound on $\mathcal{H}(E_b)$, concluding Step 2:

$$\mathcal{H}(E_b) \le K_b + \ln(\mathcal{N}(t,s)) + \sum_{i=1}^{t}(s_i - 1)\ln(b_{n_{i-1}+1}/b_{n_i}) + n\ln(\gamma) + c\sum_{i=1}^{t}\ln(s_i + 1). \tag{43}$$

Step 3. To provide the desired result, we tune $h, n_1, ..., n_t, t, s_1, ..., s_t$ in the following way. Let $h = 1/d$. For simplicity, we denote $m = m_b = \sum_{i:b_i>1} 1$. We choose $s_t = d - m$ and for $i = 2, ..., t - 1$, we set $s_i = s$ for some $s \in \mathbb{N}^*$ to determine. We have $s_1 \leq s$, and $t \leq 1 + \lceil m/s \rceil$. Let us bound the terms appearing in (43) from left to right.

$$\ln(\mathcal{N}(t, s)) \leq \ln((e(1 + t^{-1}h^{-1}))^t) \leq t(1 + \ln(1 + d)).$$

Then, since $s_1 \leq s_2 = ... = s_{t-1} = s$, and by definition of m:

$$\sum_{i=1}^{t-1} (s_i - 1) \ln(b_{n_{i-1}+1}/b_{n_i}) = (s - 1) \ln(b_1/b_m) \leq (s - 1) \ln(b_1).$$

We chose $h = 1/n$ such that, using $\gamma = \sqrt{1 + th}$:

$$n \ln(\gamma) = \frac{n}{2} \ln(1 + t/n) \leq t/2.$$

Finally:

$$\sum_{i=1}^{t} \ln(s_i + 1) = (t - 1) \ln(s + 1) + \ln(d - m + 1).$$

Combining these inequalities leads to:

$$\mathcal{H}_1(E_b) \leq K_b + C \left(t \ln(d) + (s - 1) \ln(b_1) \right).$$

As $t \leq 1 + \lceil m/s \rceil$:

$$\mathcal{H}_1(E_b) \leq K_b + C \left((1 + \lceil m/s \rceil) \ln(d) + (s - 1) \ln(b_1) \right).$$

We now need to tune s. We take s of the form:

$$s = \left\lceil \frac{m \ln(d)}{\eta} \right\rceil,$$

for some $\eta > 0$, leading to:

$$\mathcal{H}_1(E_b) \leq K_b + C \left((1 + \lceil \eta/\ln(d) \rceil) \ln(d) + (\left\lceil \frac{m \ln(d)}{\eta} \right\rceil - 1) \ln(b_1) \right).$$

Using $\lceil x \rceil \leq 1 + 2x$ for any $x \geq 0$:

$$\mathcal{H}_1(E_b) \leq K_b + C \left(2\eta + 2\frac{m \ln(d)}{\eta} \ln(b_1) + 2 \ln(d) \right).$$

Optimizing and taking $\eta = \sqrt{m \ln(d) \ln(b_1)}$ gives:

$$\mathcal{H}_1(E_b) \leq K_b + C \left(4\sqrt{m \ln(d) \ln(b_1)} + 2 \ln(d) \right),$$

concluding our proof.

A.2. Proof of Corollary 1

We start by proving Corollay 1. Let $\varepsilon > 0$. Consider the ellipsoid E_b, where $b_i = \varepsilon^{-1}\sigma_i/\sigma_1$, $i \in [d]$, and a covering $C(E_b)$ of E_b by unit Euclidean balls. Theorem 4 gives us an upper bound on the minimal size of such coverings. Let $S = \mathrm{Diag}(\varepsilon^{-1}\sigma_i/\sigma_1)_{i\in[d]}$. We then define $\mathcal{N}_\varepsilon = S^{-1}C(E_b)$. By definition of E_b, $x \in \mathcal{S}_1$ if and only if $Sx \in E_b$, so that $\mathcal{N}_\varepsilon$ consists of vectors in $\mathcal{S}_1$. Moreover, for each $y \in \mathcal{S}_1$, by definition of $C(E_b)$, there exists $x \in \mathcal{N}_\varepsilon$ such that $\|Sy - Sx\|^2 \leq 1$. This is precisely the condition required.

We now derive an upper bound on $\mathcal{H}_\varepsilon$ in terms of $d_{\mathrm{eff}}(r)$. Let m_ε be given, so that:

$$\varepsilon^{-1}\sigma_{m_\varepsilon}/\sigma_1 > 1 \geq \varepsilon^{-1}\sigma_{m_\varepsilon+1}/\sigma_1.$$

Given $d_{\mathrm{eff}}(r)$, the value of σ_{m_ε} is maximized by taking the values of $\sigma_2^2/\sigma_1^2, \ldots, \sigma_{m_\varepsilon}^2/\sigma_1^2$ all equal to $(d_{\mathrm{eff}}(r) - 1)/(m_\varepsilon - 1)$. We thus find that necessarily,

$$\varepsilon^{-\frac{2}{r}}(d_{\mathrm{eff}}(r) - 1) \geq m_\varepsilon - 1.$$

We also have the trivial bound $m_\varepsilon - 1 \leq d - 1$. Next, we note that, for fixed m_ε, the value of $\sum_{i=1}^{m_\varepsilon}\ln(\varepsilon^{-1}\sigma_i/\sigma_1)$ is maximized again, using concavity of $\ln$, by taking $\sigma_2^{2/r}/\sigma_1^{2/r}, \ldots, \sigma_{m_\varepsilon}^{2/r}/\sigma_1^{2/r}$ all equal to $(d_{\mathrm{eff}}(r) - 1)/(m_\varepsilon - 1)$. This then evaluates to:

$$\sum_{i=1}^{m_\varepsilon}\ln(\varepsilon^{-1}\sigma_i/\sigma_1) = \ln(\varepsilon^{-1}) + \frac{m_\varepsilon - 1}{2/r}\ln(\varepsilon^{-\frac{2}{r}}(d_{\mathrm{eff}}(r) - 1)/(m_\varepsilon - 1))$$

We then use the fact that $x \to x\ln(A/x)$ is increasing over $[0, A/e]$. Here, x plays the role of $m_\varepsilon - 1$, and A the role of $\varepsilon^{-\frac{2}{r}}(d_{\mathrm{eff}}(r) - 1)$. We end up with:

$$\sum_{i=1}^{m_\varepsilon}\ln(\varepsilon^{-1}\sigma_i/\sigma_1) \leq \ln(\varepsilon^{-1}) + \frac{\min(d - 1, \varepsilon^{-\frac{2}{r}}(d_{\mathrm{eff}}(r) - 1)/e)}{2/r}\ln(\max(e, \varepsilon^{-\frac{2}{r}}(d_{\mathrm{eff}}(r)-1)/(d-1))).$$

A.3. Ellipsoids in Infinite Dimension

We here present results on the unit entropy of ellipsoids in infinite dimension, which we believe to be of interest. More precisely, consider the space $V = \ell^2(\mathbb{R}) = \{x \in \mathbb{R}^{\mathbb{N}^*} | \sum_{i\geq1} x_i^2 < \infty\}$ with the classical euclidean topology. We have $\|x\|^2 = \sum_{i\geq1} x_i^2$ for $x \in V$. Note however that what follows can naturally be extended to any separable Hilbert space.

Definition 9 *For $b \in V$ such that $b_1 \geq b_2 \geq \ldots > 0$, we define the ellipsoid E_b as:*

$$E_b = \left\{x \in V : \sum_{i\geq1} \frac{x_i^2}{b_i^2} \leq 1\right\}. \tag{44}$$

We then define the ε-entropy and unit-entropy of such an ellipsoid as in the finite-dimension case.

Theorem 5 (Unit Entropy of Ellipsoids in Infinite Dimension) *Let $E_b \subset V$ an ellipsoid. Define the following quantities:*

$$
\begin{array}{rcl}
K_b & = & \sum_{i \geq 1} \ln^+(b_i), \\
m_b & = & \sum_{i \geq 1} \mathbf{1}_{b_i \geq 1/2} \\
M_b & = & \inf\{n \geq 1 : \sum_{i \geq n+1} b_i^2 \leq 1/2\}
\end{array}
$$

Then, we have for some universal constant $c > 0$:

$$
\mathcal{H}_1(E_b) \leq K_b + c \left(\sqrt{\ln^+(b_1) \ln(M_b) m_b} + \ln(M_b) \right). \tag{45}
$$

The proof follows the same steps as the one in finite dimension, replacing the global dimension in Step 1 by M_b. The interest of ellipsoids in infinite dimension lies in the appearance of another notion of dimension than the one we studied: the power-law norm decay of a vector $v \in V$.

Definition 10 (Power-Law Norm Decay) *A vector $x \in V$ is of power-law norm decay $d > 0$ if:*

$$
\sum_{i \geq n+1} x_i^2 = \mathcal{O}\left(n^{-2/d}\right).
$$

An ellipsoid E_b is said of power-law norm decay d if b is of power-law norm decay d.

The power-law norm decay d of a vector $\lambda \in V$ is closely related to the spectral dimension of infinite graphs or operators on Hilbert spaces: $(\lambda)_{i \in \mathbb{N}^*}$ usually corresponds to the eigenvalues of the Laplacian of the graph in the first case, or to the eigenvalues of the operator in the second case. The following corollary illustrates how this notion is relevant.

Corollary 2 (ε-Entropy and Power-Law Norm Decay) *Let $E_b \subset V$ be an ellipsoid of power-law norm decay $d > 0$. Then, when $\varepsilon \to 0$, we have:*

$$
\mathcal{H}_\varepsilon(E_b) \leq d \ln(\varepsilon^{-1})^2 (1 + o(1)). \tag{46}
$$

Proof We have for any ellipsoid E_b of power-law norm decay d:

$$
m_{b/\varepsilon} \leq M_{b/\varepsilon} = \mathcal{O}(\varepsilon^{-d/2}). \tag{47}
$$

Remark that $\mathcal{H}_\varepsilon(E_b) = \mathcal{H}_1(E_{b/\varepsilon})$. Then, using Theorem 5, and the previous consideration (47), we obtain our result. ∎

This needs to be put in light with the ε-entropy of the unit ball in dimension $d < \infty$, that behaves as $d \ln(\varepsilon^{-1})$ when $\varepsilon \to 0$. Despite the presence of $\ln(\varepsilon^{-1})^2$ instead of $\ln(\varepsilon^{-1})$, we have a linearity in this expression in terms of d.

Appendix B. Proof of Theorems 2 and 1 and General Considerations on these Large Deviation Bounds

B.1. Proof of Theorem 2: Bound Without Centering

In order to have lighter notations, we write $d'_{\text{eff}} = d_{\text{eff}}(r)$ in what follows. For all $j \geq 0$, let $\mathcal{N}_j$ be a covering of $\mathcal{S}_1$ satisfying the properties of Corollary 1 for $\varepsilon_j = 2^{-j}$. For all $x \in \mathcal{S}_1$, let

$\Pi_j x$ be some point in $\mathcal{N}_j$ such that (15) holds. By convention we take $\Pi_0 x = 0$. Then for all $(x_1, \ldots, x_r) \in \mathcal{S}^r$, using the chaining approach, we write

$$\frac{1}{n} \sum_{i \in [n]} \prod_{k \in [r]} f_k(a_i^\top x_k) = \sum_{j \geq 0} \frac{1}{n} \sum_{i \in [n]} \left\{ \prod_{k \in [r]} f_k(a_i^\top \Pi_{j+1} x_k) - \prod_{k \in [r]} f_k(a_i^\top \Pi_j x_k) \right\}$$

$$= \sum_{j \geq 0} \sum_{k \in [r]} \frac{1}{n} \sum_{i \in [n]} [f_k(a_i^\top \Pi_{j+1} x_k) - f_k(a_i^\top \Pi_j x_k)]$$

$$\times \left\{ \prod_{\ell=1}^{k-1} f_\ell(a_i^\top \Pi_{j+1} x_\ell) \times \prod_{\ell=k+1}^{r} f_\ell(a_i^\top \Pi_j x_\ell) \right\}.$$

Let $j \geq 0$ and $k \in [r]$ be fixed. Consider a term of the form

$$Z = \frac{1}{n} \sum_{i \in [n]} Z_i,$$

with

$$Z_i = \prod_{\ell=1}^{k-1} f_\ell(a_i^\top u_\ell)[f_k(a_i^\top u_k) - f_k(a_i^\top v_k)] \prod_{\ell=k+1}^{r} f_\ell(a_i^\top v_\ell), \tag{48}$$

where $u_\ell \in \mathcal{N}_j$, $v_\ell \in \mathcal{N}_{j+1}$, and $\|u_k - v_k\|_\Sigma \leq \sigma_1 \epsilon_j$, where we defined $\epsilon_j := 2^{-j+1}$. By the triangle inequality, for all $x_\ell \in \mathcal{S}_1$, letting $u_\ell = \Pi_j x_\ell$ and $v_\ell = \Pi_{j+1} x_\ell$, these assumptions are satisfied. Note also that $\|u_\ell\|_\Sigma$ and $\|v_\ell\|_\Sigma$ are upper-bounded by σ_1.

Clearly, $|Z_i| \leq 2B$. Also,

$$|Z_i| \leq (B/B_k)|a_i^\top(u_k - v_k)|$$
$$\leq (B/B_k)\|a_i\|_{\Sigma^{-1}} \|u_k - v_k\|_\Sigma.$$

We introduce the new parameter $R_{\Sigma^{-1}}$, that is an upper bound on the norms $\|a_i\|_{\Sigma^{-1}}$. Note that, for the Gaussian case where Σ is the covariance matrix of the Gaussian vector a_i, the natural scaling assumption is to take:

$$R_{\Sigma^{-1}} = \Theta(\sqrt{d}).$$

As we don't want any dependency on the overall dimension d, we will aim at making this quantity disappear. We then introduce the notation

$$P_{j,k} := \min(2B/\epsilon_j, (B/B_k)\sigma_1 R_{\Sigma^{-1}}).$$

For each Z_i and $t > 0$, using that $\|u_k - v_k\|_\Sigma \leq \varepsilon_j$, we then have:

$$\mathbb{P}(Z_i \geq \epsilon_j \sigma_1^r t) \leq \mathbb{P}(|f_\ell(a_i^\top u_\ell)| \geq \sigma_1 t^{1/r} \text{ for some } \ell < k,$$
$$\text{or } |f_k(a_i^\top u_k) - f_k(a_i^\top v_k)| \geq \sigma_1 \epsilon_j t^{1/r},$$
$$\text{or } |f_\ell(a_i^\top v_\ell)| \geq \sigma_1 t^{1/r} \text{ for some } \ell > k)$$
$$\leq 2r e^{-t^{2/r}/2} \text{ if } t \leq \sigma_1^{-r} P_{j,k}$$
$$= 0 \text{ if } t > \sigma_1^{-r} P_{j,k},$$

We note $P_j = \frac{2B}{\varepsilon_j} \geq P_{j,k}$. The previous bounds on the tail of Z_i's distribution allow to bound exponential moments of Z_i. Fix some $\theta > 0$. Then

$$
\begin{aligned}
\mathbb{E}e^{(\theta/n)\sigma_1^{-r}Z_i/\epsilon_j} &\leq 1 + \frac{\theta}{n}\int_0^\infty e^{(\theta/n)y}\mathbb{P}(\sigma_1^{-r}Z_i/\epsilon_j \geq y)dy \\
&\leq 1 + \frac{\theta}{n}2r\int_0^{\sigma_1^{-r}P_{j,k}} e^{(\theta/n)y}e^{-y^{2/r}/2}dy.
\end{aligned}
$$

We now fix $\theta_{j,k} > 0$ such that, for all $y \in [0, \sigma^{-r}P_{j,k}]$, one has:

$$
(\theta_{j,k}/n)y \leq \frac{1}{4}y^{2/r},
$$

or equivalently, we assume (recall that $r \geq 2$):

$$
\frac{\theta_{j,k}}{n} \leq \frac{1}{4}(\sigma_1^{-r}P_{j,k})^{2/r-1}. \tag{49}
$$

This entails the bound:

$$
\mathbb{E}e^{(\theta_{j,k}/n)\sigma_1^{-r}Z_i/\epsilon_j} \leq 1 + \frac{\theta_{j,k}}{n}2r\int_0^\infty e^{-y^{2/r}/4}dy =: 1 + \frac{\theta_{j,k}}{n}c_r,
$$

where we introduced notation $c_r := 2r\int_0^\infty e^{-y^{2/r}/4}dy$. Thus, for $\theta_{j,k} > 0$ satisfying (49), one has:

$$
\mathbb{E}e^{\theta_{j,k}\sigma_1^{-r}\frac{1}{n}\sum_{i\in[n]}Z_i/\epsilon_j} \leq (1 + \theta_{j,k}c_r/n)^n \leq e^{\theta_{j,k}c_r}. \tag{50}
$$

The number of possible choices for $u_\ell \in \mathcal{N}_j$ and $v_\ell \in \mathcal{N}_{j+1}$ involved in the definition of Z_i is upper-bounded by

$$
|\mathcal{N}_{j+1}|^{r+1} \leq e^{(r+1)\mathcal{H}_{j+1}},
$$

where $\mathcal{H}_j$ is defined in (16). Thus for any $t_{j,k} > 0$, the probability that for some choice of u_ℓ, v_ℓ in the corresponding nets, one has

$$
\frac{1}{n}\sum_{i\in[n]}Z_i \geq \epsilon_j\sigma^r t_{j,k}
$$

is upper bounded by:

$$
e^{(r+1)\mathcal{H}_{j+1}} \inf_{\theta\in[0,\frac{n}{4}(\sigma^{-r}P_{j,k})^{2/r-1}]} e^{-\theta t_{j,k}+\theta c_r}.
$$

We now take $\theta_{j,k} = \frac{n(\sigma_1^{-r}P_{j,k})^{2/r-1}}{4}$, and :

$$
t_{j,k} = c_r + \frac{(r+1)\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)}{\theta_{j,k}}
$$

for some $\lambda \geq 0$. This upper bound is then no more than $(1+j)^{-2}e^{-\lambda}$. We now use a union bound over $j \geq 0$ and $k \in [r]$ to obtain:

$$
\mathbb{P}(Y \geq \sigma^r \sum_{j\geq 0}\sum_{k\in[r]}\epsilon_j t_{j,k}) \leq e^{-\lambda}\, r\sum_{j\geq 0}(1+j)^{-2} = r\frac{\pi^2}{6}e^{-\lambda}.
$$

Let us bound the sum appearing in this probability. Fix some $k \in [r]$.

$$\sum_{j \geq 0} \epsilon_j t_{j,k} = \sum_{j \geq 0} 2^{-j+1} \left[c_r + \frac{(r+1)\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)}{\theta_{j,k}} \right] \tag{51}$$

with $\theta_{j,k} \geq \frac{n}{4} \left(\sigma_1^{-r} P_j \right)^{2/r-1} = \left(\sigma_1^{-r} 2B \right)^{2/r-1} 2^{-j+j\frac{2}{r}}$, for $P_j = 2B/\varepsilon_j$. Moreover, for $j^* = \frac{r}{2} \ln_2 \left(e\frac{d-1}{d'_{\text{eff}}-1} \right)$, we have:

$$j \leq j^* \implies \mathcal{H}_j \leq j \ln(2) + 2^{j\frac{2}{r}} \frac{d'_{\text{eff}} - 1}{e} + c \left[\ln(d) + \sqrt{j\ln(d)m_j} \right],$$

$$j > j^* \implies \mathcal{H}_j \leq j \ln(2) + (d-1)\ln\left(2^{j\frac{2}{r}} \frac{d'_{\text{eff}} - 1}{d-1} \right) + c \left[\ln(d) + \sqrt{j\ln(d)m_j} \right],$$

$$\text{with } m_j \leq 1 + 2^{j\frac{2}{r}} (d'_{\text{eff}} - 1).$$

The *easy* part of (51) to study:

$$\sum_{j \geq 0} 2^{-j+1} \left[c_r + \frac{\lambda + 2\ln(j+1)}{\theta_{j,k}} \right] \leq \frac{A_r \lambda + B_r}{\frac{n}{4} \left(\sigma_1^{-r} B \right)^{2/r-1}}$$

Now, let us bound $H := \sum_{j>0} 2^{-j} \mathcal{H}_j / \varepsilon_j^{1-2/r} = \sum_{j>0} 2^{-j\frac{2}{r}} \mathcal{H}_j$.

$$\sum_{1 \leq j \leq j^*} 2^{-j\frac{2}{r}} \mathcal{H}_j \leq \sum_{1 \leq j \leq j^*} 2^{-j\frac{2}{r}} \left(j\ln(2) + 2^{j\frac{2}{r}} \frac{d'_{\text{eff}} - 1}{e} + c \left[\ln(d) + \sqrt{j\ln(d)2^{j\frac{2}{r}}d'_{\text{eff}}} \right] \right)$$

$$\leq C_r j^* d'_{\text{eff}} + D_r \sqrt{\ln(d)d'_{\text{eff}}} + E_r \ln(d).$$

Using that j^* is *big* when $d'_{\text{eff}} \ll d$, we bound the sum for $j > j^*$:

$$\sum_{j>j^*} 2^{-j\frac{2}{r}} \mathcal{H}_j \leq \sum_{j>j^*} 2^{-j\frac{2}{r}} \left(j\ln(2) + (d-1)\ln\left(2^{j\frac{2}{r}} \frac{d'_{\text{eff}} - 1}{d-1} \right) + c \left[\ln(d) + \sqrt{j\ln(d)m_j} \right] \right)$$

$$\leq F'_r \sum_{j>j^*} j 2^{-j\frac{2}{r}} d + G'_r \sum_{j>j^*} 2^{-j\frac{2}{r}} \sqrt{j 2^{j\frac{2}{r}} \ln(d)}$$

$$\leq F_r d'_{\text{eff}} \ln\left(\frac{d}{d'_{\text{eff}}} \right) + G_r,$$

where we used that $2^{j^* \frac{2}{r}} = e\frac{d-1}{d'_{\text{eff}}-1}$ and $\sum_{j>k} j 2^{-j\frac{2}{r}} \leq Ck 2^{-k\frac{2}{r}}$. All in one, that leaves us with the following bound on (51):

$$\sum_{j \geq 0} \epsilon_j t_{j,k} \leq c_r \frac{A_r \lambda + B_r + C_r d'_{\text{eff}} + D_r \sqrt{\ln(d)d'_{\text{eff}}} + E_r \ln(d) + F_r d'_{\text{eff}} \ln\left(\frac{d}{d'_{\text{eff}}} \right) + G_r}{\frac{n}{4} \left(\sigma_1^{-r} B \right)^{2/r-1}}.$$

In a more synthetic formulation:

$$\sum_{j \geq 0} \epsilon_j t_{j,k} \leq C_r \left(1 + \frac{d'_{\text{eff}} \ln(d) + \lambda}{\frac{n}{4} \left(\sigma_1^{-r} B \right)^{2/r-1}} \right). \tag{52}$$

25

For some suitable constant C_r the following holds, by summing previous considerations for $1 \leq k \leq r$. One has for all $\lambda > 0$ that:

$$\mathbb{P}\left(Y \geq \sigma_1^r C_r \left(1 + \frac{d_{\text{eff}}(r)\ln(d) + \lambda}{n}(\sigma_1^{-r}B)^{1-2/r}\right)\right) \leq e^{-\lambda}\, r\frac{\pi^2}{6}. \tag{53}$$

B.2. Proof of Theorem 1: Bound With Centering

We now look for bounds on:

$$Y' := \sup_{x_1,\dots,x_r \in \mathcal{S}_1} \frac{1}{n} \sum_{i \in [n]} \left\{ \prod_{k=1}^r f_k(a_i^\top x_k) - \mathbb{E}\prod_{k=1}^r f_k(a_1^\top x_k)\right\}. \tag{54}$$

Fixing $j \geq 0$, $k \in [r]$, we again consider the random variables Z_i as previously defined in (48). Write:

$$\mathbb{E}e^{(\theta/n)\sigma_1^{-r}[Z_i - \mathbb{E}Z_i]/\epsilon_j} = 1 + \left(\frac{\theta}{n}\right)^2 \mathbb{E}\left[(\epsilon_j^{-1}\sigma_1^{-r}(Z_i - \mathbb{E}Z_i))^2 F((\theta/n)(\epsilon_j^{-1}\sigma_1^{-r}(Z_i - \mathbb{E}Z_i)))\right],$$

where :

$$F(x) := x^{-2}[e^x - x - 1] \leq e^{|x|}.$$

Thus using this bound and the inequality $xy \leq x^2 + y^2$:

$$\mathbb{E}e^{(\theta/n)\sigma_1^{-r}[Z_i - \mathbb{E}Z_i]/\epsilon_j} \leq 1 + \left(\frac{\theta}{n}\right)^2 \left[\mathbb{E}((\epsilon_j^{-1}\sigma_1^{-r}(Z_i - \mathbb{E}Z_i))^4 + \mathbb{E}e^{2(\theta/n)\sigma_1^{-r}|Z_i - \mathbb{E}Z_i|/\epsilon_j}\right].$$

By the sub-gaussian tail assumption, $\mathbb{E}(\epsilon_j^{-1}\sigma_1^{-r}(Z_i - \mathbb{E}Z_i))^4$ is bounded by a constant κ_r dependent on r. By the same arguments as above, for:

$$\frac{\theta}{n} \leq \frac{(\sigma_1^{-r}P_{j,k})^{2/r-1}}{8},$$

then $\mathbb{E}e^{2(\theta/n)\sigma_1^{-r}|Z_i - \mathbb{E}Z_i|/\epsilon_j}$ is also bounded by another constant κ_r' dependent on r. Indeed, by the sub-gaussian tail assumption, $|\mathbb{E}Z_i| \leq \sigma_1^r \epsilon_j s_r$ for some r-dependent constant, and we can then use the upper bound:

$$\begin{aligned}\mathbb{E}e^{2\theta/n\sigma_1^{-r}|Z_i - \mathbb{E}Z_i|/\epsilon_j} &\leq e^{2(\theta/n)s_r}[1 + \tfrac{\theta}{n}\int_0^\infty e^{2(\theta/n)y}[\mathbb{P}(Z_i \geq y\sigma_1^r\epsilon_j) + \mathbb{P}(-Z_i \geq y\sigma_1^r\epsilon_j)]dy]\\ &\leq e^{2(\theta/n)s_r}[1 + \tfrac{\theta}{n}2r\int_0^{\sigma^{-r}P_{j,k}} e^{2(\theta/n)y - y^{2/r}/2}dy].\end{aligned}$$

Thus the probability that for some choices of u_ℓ, v_ℓ, $\ell \in [r]$ in the suitable ϵ-nets, one has:

$$\frac{1}{n}\sum_{i \in [n]} Z_i - \mathbb{E}Z_i \geq \sigma_1^r \epsilon_j t_{j,k}$$

is upper-bounded, for all $\theta \in [0, n(\sigma_1^{-r}P_{j,k})^{2/r-1}/8]$ by:

$$\exp\left((r+1)\mathcal{H}_{j+1} - \theta t_{j,k} + \kappa_r''\theta^2/n\right),$$

26

where we defined $\kappa_r'' = \kappa_r + \kappa_r'$. Let now $\theta_{j,k} = \min(n, n(\sigma_1^{-r} P_{j,k})^{2/r-1}/8)$, and :

$$t_{j,k} = \kappa_r'' \frac{\theta_{j,k}}{n} + \frac{1}{\theta_{j,k}}[(r+1)\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)],$$

where $\lambda > 0$ is a free parameter. We have, for $t_j = \sum_k t_{j,k}$, if all $\theta_{j,k}$ have the same value θ_j:

$$t_j \leq c_r \left(\frac{\theta_j}{n} + \frac{(r+1)\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)}{\theta_j} \right)$$

We now take $\theta_j = \min \left(\frac{n(\sigma_1^{-r} P_{j,k})^{2/r-1}}{8}, \sqrt{n\left[(r+1)\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)\right]} \right)$, leading to:

$$\varepsilon_j t_j \leq a_r \varepsilon_j^{2/r} \frac{(B\sigma_1)^{1-\frac{2}{r}}}{n} [\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)] + b_r \varepsilon_j \sqrt{\frac{\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)}{n}}.$$

The first term is treated in the non-centered case:

$$a_r \varepsilon_j^{2/r} \frac{(B\sigma_1)^{1-\frac{2}{r}}}{n} [\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)] \leq A_r \frac{1}{n} \frac{d_{\mathrm{eff}}' \ln(d) + \lambda}{\left(\sigma_1^{-r} B\right)^{1-2/r}}.$$

The second one follows the same lines:

$$b_r \varepsilon_j \sqrt{\frac{\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)}{n}} \leq b_r' \varepsilon_j \frac{1}{\sqrt{n}} \left(\sqrt{\mathcal{H}_{j+1}} + \sqrt{\lambda} \right).$$

In the same way that we proved $\sum_j \varepsilon_j \mathcal{H}_j \leq C_r^{te} d_{\mathrm{eff}}(r) \ln(d)$, we have that, for $C > 0$ a constant:

$$\sum_{j \geq 0} \varepsilon_j \sqrt{\mathcal{H}_j} \leq C \sqrt{d_{\mathrm{eff}}(1) \ln(d)},$$

giving us:

$$\sum_{j \geq 0} b_r \varepsilon_j \sqrt{\frac{\mathcal{H}_{j+1} + \lambda + 2\ln(j+1)}{n}} \leq B_r \frac{\sqrt{\lambda} + \sqrt{d_{\mathrm{eff}}(1) \ln(d)}}{\sqrt{n}}.$$

We then have:

$$\mathbb{P} \left(Y' \geq \sigma^r \sum_{j \geq 0} \epsilon_j t_j \right) \leq r \frac{\pi^2}{6} e^{-\lambda},$$

for:

$$\sum_{j \geq 0} \epsilon_j t_j \leq A_r \frac{1}{n} \frac{d_{\mathrm{eff}}' \ln(d) + \lambda}{\left(\sigma_1^{-r} B\right)^{2/r-1}} + B_r \frac{\sqrt{\lambda} + \sqrt{d_{\mathrm{eff}}(1) \ln(d)}}{\sqrt{n}}.$$

for some suitable constant A_r dependent only on r. We hence end up with the same computation as in the non-centered case (up to constants), leading to the following result. For suitable constant C_r', for all $\lambda > 0$, one has that:

$$\mathbb{P} \left(Y' \geq C_r' \sigma_1^r \left(\frac{1}{n} \frac{\lambda + d_{\mathrm{eff}}' \ln(d)}{\left(\sigma_1^{-r} B\right)^{2/r-1}} + \frac{\sqrt{\lambda} + \sqrt{d_{\mathrm{eff}}(1) \ln(d)}}{\sqrt{n}} \right) \right) \leq r \frac{\pi^2}{6} e^{-\lambda}, \tag{55}$$

Y' defined in (54).

B.3. Proof of Remark 1

Lemma 1 (Maximum of n *i.i.d.* Subagaussian Random Variables) *Let $a_1, ..., a_n$ be i.i.d. Σ-subgaussian random variables. Denote $R = \max_{i=1,...,n} \|a_i\|$. There exists a (universal) constant $C > 0$ such that, with probability $1 - \delta$:*

$$R^2 \leq 4\sigma_1^2 \left(\ln(\delta^{-1}) + 2d_{\text{eff}}(1) + \ln(n) \right). \tag{56}$$

Proof Let $t \geq 0$. Using a classical Markov-Chernoff approach, for some $\lambda > 0$:

$$\mathbb{P}(R \geq t) \leq n\mathbb{P}(\|a_1\|^2 \geq t^2)$$
$$\leq ne^{-\lambda t^2/2}\mathbb{E}e^{\lambda\|a_1\|^2/2}.$$

Then, writing $a_1 = \sum_{j=1}^d \rceil_j \sigma_j X_j$ where $(\rceil_\mid)_j$ is an orthonormal basis of eigenvectors of Σ, and $X_1, ..., X_d$ are *i.i.d.* standard gaussian variables $\mathcal{N}(0,1)$, yields, using independence:

$$\mathbb{E}e^{\lambda\|a_1\|^2/2} = \prod_{j=1}^d \mathbb{E}e^{\lambda\sigma_j^2 X_j^2/2}$$
$$= \prod_{j=1}^d \frac{1}{\sqrt{1 - \lambda\sigma_j^2}},$$

where we assume that $\lambda < \sigma_1^{-2}$. We take $\lambda = \frac{1}{2\sigma_1^2}$. Now, using that $\frac{1}{\sqrt{1-u}} \leq e^{2u}$ for $0 \leq u \leq 1/2$ yields for this particular λ:

$$\mathbb{E}e^{\lambda\|a_1\|^2/2} \leq e^{2\sum_{j=1}^d \frac{\sigma_j^2}{\sigma_1^2}}$$
$$= e^{2d_{\text{eff}}(1)}.$$

We thus get $\mathbb{P}(R \geq t) \leq n\exp(-\frac{t^2}{4\sigma_1^2} + 2d_{\text{eff}}(1))$. For $\delta \in (0,1)$, we hence have that, with probability $1 - \delta$:

$$R^2 \leq 4\sigma_1^2(\ln(\delta^{-1}) + 2d_{\text{eff}}(1) + \ln(n)). \tag{57}$$

■

B.4. Eventual Tightness

B.4.1. WITHOUT CENTERING

We want to derive possible tightness for our probability bounds. Let $A_1, ..., A_n$ *i.i.d.* centered gaussians of covariance Σ. Then, for $x_1, ..., x_r \in \mathcal{S}$:

$$\mathbb{E}\left[\prod_{k=1}^r A_1^\top x_k \right] \leq O(\sigma_1^r).$$

We then take $x_1 = ... = x_r = A_1 / \|A_1\|$ in order to have:

$$\frac{1}{n} \sum_i \prod_{k=1}^{r} A_i^\top x_k = \frac{\|A_1\|^r}{n} + O(\sigma_1^r) \tag{58}$$

$$= O\left(\frac{\sigma_1^r d_{\text{eff}}(1)^{r/2}}{n} + \sigma_1^r \right). \tag{59}$$

Cotrast this we the results obtained in Theorem 2: we require n to be of order $d_{\text{eff}}(r) \ln(d) d_{\text{eff}}(1)^{r/2-1}$ (Remark 1) for our bound to be of order $O(1)$. Considerations just above, and in particular (59) require $d_{\text{eff}}(1)^{r/2} = O(n)$. Our lower and upper bounds match only up to a factor $\frac{d_{\text{eff}}(r) \ln(d)}{d_{\text{eff}}(1)}$, that should not be too large. However, our dependency in n seems optimal $(1/n)$. We believe that $d_{\text{eff}}(r) \ln(d)$ instead of $d_{\text{eff}}(1)$ is is simply an artifact of the proof.

B.4.2. WITH CENTERING

We now consider the centered case. The non-centered case suggest that we are not tight in terms of dimension-dependency, we thus restrict ourselves to the dependency in n. Consider the same random variables $A_1, ..., A_n$ as above. Let, for $x_1, ..., x_k \in \mathcal{B}$:

$$X = \frac{1}{n} \sum_{i=1}^{n} \left(\prod_{k=1}^{r} A_i^\top x_k - \mathbb{E}\left[\prod_{k=1}^{r} A_i^\top x_k \right] \right). \tag{60}$$

We have:

$$\mathbb{E}[X^2] = \frac{1}{n} \mathbb{E}\left[\left(\prod_{k=1}^{r} A_1^\top x_k - \mathbb{E} \prod_{k=1}^{r} A_1^\top x_k \right)^2 \right].$$

We thus observe a dependency in $1/n$ on the second moment. That leads to an optimal dependency in n in our centered bound. Indeed, we have a $1/\sqrt{n}$, but we cannot gain any order of magnitude: if we have something of the form $\mathbb{P}(X \geq \gamma \frac{\lambda+\beta}{n^\alpha}) \leq e^{-\lambda}$ for all $\lambda > 0$, we get $E[X^2] \leq \frac{\gamma^2 \beta^2}{n^{2\alpha}}$, leading to an optimal exponent α of $1/2$, which we have.

Appendix C. Bounding Random Tensors

C.1. General Considerations: Isotropic Bound VS Non-Isotropic Bound

We wish to bound the operator norm of tensors of the form $T = \frac{1}{n} \sum_{i=1}^{n} T_i$, where $T_1, ..., T_n$ are *i.i.d.* subgaussian random tensors of rank 1 and order p, using a dependency in an effective dimension rather than the global one. We have:

$$\|T\|_{\text{op}} = \frac{1}{n} \sup_{x_1, ..., x_p \in \mathcal{S}} \sum_{i=1}^{n} \prod_{k=1}^{p} \langle a_i, x_k \rangle \tag{61}$$

$$= \frac{1}{n} \sup_{x \in \mathcal{S}} \sum_{i=1}^{n} \langle a_i, x \rangle^p. \tag{62}$$

Bubeck et al. (2020) use tools from Paouris et al. (2017) and version (62) of the operator norm in the following way: for fixed $x \in \mathcal{S}$, $\sum_{i=1}^{n} \langle a_i, x \rangle^p$ is bounded by $c_p \sigma_1^p \sqrt{\frac{\tau}{n}}$ with probability $c_p' e^{-c_p'' \tau}$. Then, a union bound on $\mathcal{N}$ a $\frac{1}{2p}$-net of balls ($\mathcal{N}$ of size $(1 + 4p)^d$) covering $\mathcal{S}$ is made. Finally, as $\|T\|_{\mathrm{op}} \leq \max_{x \in \mathcal{N}} \langle T, x^{\otimes p} \rangle + \frac{1}{2} \|T\|_{\mathrm{op}}$ yields the following.

Proposition 8 (Isotropic Bound on Random Tensors) *Let $T_1, ..., T_n$ be i.i.d. random tensors of order p, rank 1, symmetric and $\sigma_1^2 I_d$-subgaussian. Let $T = \frac{1}{n} \sum_{i=1}^{n} T_i$. Then, for some universal constant $C_p > 0$ we have with probability $1 - \delta$:*

$$\|T - \mathbb{E}T\|_{\mathrm{op}} \leq C_p \sigma_1^p \sqrt{\frac{d + \ln(\delta^{-1})}{n}}. \tag{63}$$

However, if we desire to use the non-isotropic properties of the random variables $a_1, ..., a_n$, some changes need to be done. Still using formulation (62), we get:

$$\|T\|_{\mathrm{op}} = \frac{1}{n} \sup_{x \in \mathcal{S}} \sum_{i=1}^{n} \langle \sqrt{\Sigma}^{-1} a_i, \sqrt{\Sigma} x \rangle^p \tag{64}$$

$$= \frac{1}{n} \sup_{\sqrt{\Sigma}^{-1} y \in \mathcal{S}} \sum_{i=1}^{n} \langle \sqrt{\Sigma}^{-1} a_i, y \rangle^p. \tag{65}$$

(65) can then be upper-bounded with high probability in the same way as in the isotropic case. Then, instead of covering $\mathcal{S}$ with balls of radii $\varepsilon_p = 1/(2p)$, we need to cover $\sqrt{\Sigma}\mathcal{S}$ with these balls. Our hope would be that the logarithm of the size of this ε_p-net would be linear in an effective dimension rather than a global dimension. However, as highlighted in Corollary 1, that is the case only if ε_p is not too small in front of $\frac{d_{\mathrm{eff}}(r) - 1}{d - 1}$ for any effective dimension as in Theorem 2.

Proposition 9 (Non-Isotropic Bound on Random Tensors, Version 1) *Let $T_1, ..., T_n$ be i.i.d. random tensors of order p, rank 1, symmetric and Σ-subgaussian. Let $T = \frac{1}{n} \sum_{i=1}^{n} T_i$. Assume that their exists some $r \in \mathbb{N}^*$ such that:*

$$(2p)^{\frac{2}{r}} \frac{d_{\mathrm{eff}}(r) - 1}{d - 1} \leq e. \tag{66}$$

Then, for some universal constant $C_{p,r} > 0$ we have with probability $1 - \delta$:

$$\|T - \mathbb{E}T\|_{\mathrm{op}} \leq C_{p,r} \sigma_1^p \sqrt{\frac{d_{\mathrm{eff}}(r) + \ln(d) + \ln(\delta^{-1})}{n}}. \tag{67}$$

Corollary 3 *Let $T_1, ..., T_n$ be i.i.d. random tensors of order p, rank 1, symmetric and Σ-subgaussian. Let $T = \frac{1}{n} \sum_{i=1}^{n} T_i$. Then, for some universal constant $C_p > 0$ we have with probability $1 - \delta$:*

$$\|T - \mathbb{E}T\|_{\mathrm{op}} \leq C_p \sigma_1^p \sqrt{\frac{d_{\mathrm{eff}}(1) + \ln(d) + \ln(\delta^{-1})}{n}}. \tag{68}$$

Using Theorem 1, the chaining argument uses an infinite sequence of ε-coverings and gives the following.

30

Proposition 10 (Non-Isotropic Bound on Random Tensors, Version 2) *Let $T_1, ..., T_n$ be i.i.d. random tensors of order p, symmetric and Σ-subgaussian. Let $T = \frac{1}{n}\sum_{i=1}^n T_i$. With probability $1 - \delta$ for any $\delta > 0$, we have:*

$$\|T\|_{\mathrm{op}} \leq c_p \sigma_1^p \left(1 + \frac{d_{\mathrm{eff}}(p)\ln(d) + \ln(\delta^{-1})}{n}(d_{\mathrm{eff}}(1) + \ln(\delta^{-1}))^{p/2-1}\right),$$

and:

$$\|T - \mathbb{E}T\| \leq C_p' \sigma_1^p \left(\frac{d_{\mathrm{eff}}(p)\ln(d) + \ln(\delta^{-1})}{n}(d_{\mathrm{eff}}(1) + \ln(\delta^{-1}))^{p/2-1}\right.$$
$$\left. + \frac{\sqrt{\ln(\delta^{-1})} + \sqrt{d_{\mathrm{eff}}(1)\ln(d)}}{\sqrt{n}}\right).$$

C.2. Proof of the Isotropic Large Deviation Bound (Proposition 8)

Let $x \in \mathcal{S}$. Let us bound deviations from its mean with high probability of $Y_x = \sum_{i=1}^n \langle a_i, x\rangle^p$. $\langle a_i, x\rangle$ for $i = 1, ..., n$ are *i.i.d.* distributed according to $\mathcal{N}(0, \sigma_1^2)$. From Paouris et al. (2017) (their Theorem 1.1, concentration of ℓ^p norms of gaussians), we get:

$$\mathbb{P}\left(\frac{1}{n}(Y_x - \mathbb{E}Y_x) \geq c_p \sigma_1^p \sqrt{\frac{\tau}{n}}\right) \leq c_p' e^{-c_p''\tau}.$$

Let $\mathcal{N}$ be $\frac{1}{2p}$-covering of $\mathcal{S}$. We know that one can achieve $|\mathcal{N}| \leq (1 + 4p)^d$. An union bound on $\mathcal{N}$ yields:

$$\mathbb{P}\left(\exists x \in \mathcal{N} : \frac{1}{n}(Y_x - \mathbb{E}Y_x) \geq c_p \sigma_1^p \sqrt{\frac{\tau}{n}}\right) \leq c_p' e^{-c_p''\tau + d\ln(1+p)},$$

giving:

$$\mathbb{P}\left(\sup_{x\in\mathcal{N}}\langle(T - \mathbb{E}T), x^{\otimes p}\rangle \geq c_p \sigma_1^p \sqrt{\frac{\tau}{n}}\right) \leq c_p' e^{-c_p''\tau + d\ln(1+p)}.$$

Now, let $y \in \mathcal{S}$. There exists $x \in \mathcal{N}$ such that $\|x - y\| \leq 1/(2p)$.

$$\langle(T - \mathbb{E}T), y^{\otimes p}\rangle \leq \langle(T - \mathbb{E}T), x^{\otimes p}\rangle + \|T - \mathbb{E}T\|_{\mathrm{op}}\|(x - y)^{\otimes p}\|$$
$$\leq \sup_{x\in\mathcal{N}}\langle(T - \mathbb{E}T) + \frac{1}{2}\|T - \mathbb{E}T\|_{\mathrm{op}},$$

as we have $\|(x - y)^{\otimes p}\| \leq p\|x - y\| \leq 1/2$. We then have our result by taking the supremum over all $x \in \mathcal{S}$.

C.3. Proof Of the Non-Isotropic Large Deviation Bounds

Here, the a_i's are assumed to be Σ-subgaussian. We begin by proving Proposition 9. Remind that we have:

$$\|T\|_{\mathrm{op}} = \frac{1}{n}\sup_{x\in\mathcal{S}}\sum_{i=1}^n \langle\sqrt{\Sigma}^{-1}a_i, \sqrt{\Sigma}x\rangle^p$$
$$= \frac{1}{n}\sup_{\sqrt{\Sigma}^{-1}y\in\mathcal{S}}\sum_{i=1}^n \langle\sqrt{\Sigma}^{-1}a_i, y\rangle^p,$$

and the $\sqrt{\Sigma^{-1}}a_i, i = 1, ..., n$ are *i.i.d.* distributed according to $\mathcal{N}(0, I_d)$. Denote as above $Z_y = \sum_{i=1}^{n}\langle\sqrt{\Sigma}^{-1}a_i, y\rangle^p$, for any $y \in \sqrt{\Sigma}\mathcal{S}$. We know that for such y, we have $\|y\| \leq \sigma_1$. Using again results from Paouris et al. (2017) for $y \in \sqrt{\Sigma}\mathcal{S}$:

$$\mathbb{P}\left(\frac{1}{n}(Z_y - \mathbb{E}Z_y) \geq c_p\sigma_1^p\sqrt{\frac{\tau}{n}}\right) \leq c_p'e^{-c_p''\tau}.$$

Let now $\mathcal{N}$ be an $1/(2p)$-covering of $\sqrt{\Sigma}\mathcal{S}$. We thus have:

$$\mathbb{P}\left(\sup_{x\in\mathcal{N}}\langle(T - \mathbb{E}T), x^{\otimes p}\rangle \geq c_p\sigma_1^p\sqrt{\frac{\tau}{n}}\right) \leq c_p'e^{-c_p''\tau+\ln(|\mathcal{N}|)}.$$

Let us use Corollary 1 in order to bound $\ln(|\mathcal{N}|)$:

$$\ln(|\mathcal{N}|) \leq C_{p,r}'\left[\min\left(d-1, (2p)^{2/r}(d_{\text{eff}}(r)-1)/e\right)\ln\left(\max\left(e, (2p)^{2/r}\frac{d_{\text{eff}}(r)-1}{d-1}\right)\right) + \ln(d) + d_{\text{eff}}(r)\right],$$

and thus, as we assume that $(2p)^{2/r} \leq e\frac{d-1}{d_{\text{eff}}(r)-1}$, we have:

$$\ln(|\mathcal{N}|) \leq C_{p,r}'\left((2p)^{2/r}\frac{d_{\text{eff}}(r)}{e} + \ln(d)\right).$$

Back to our probabilistic bound:

$$\mathbb{P}\left(\sup_{x\in\mathcal{N}}\langle(T - \mathbb{E}T), x^{\otimes p}\rangle \geq c_p\sigma_1^p\sqrt{\frac{\tau}{n}}\right) \leq c_p'e^{-c_p''\tau+C_{p,r}(d_{\text{eff}}(r)+\ln(d))},$$

yielding the stated result, using the same argument as in the isotropic case: for some constant $C_{p,r}$, for any $\delta > 0$, with probability $1 - \delta$:

$$\|T - \mathbb{E}T\|_{\text{op}} \leq C_{p,r}\sqrt{\frac{d_{\text{eff}}(r) + \ln(d) + \ln(\delta^{-1})}{n}}.$$

Let now take $r = 1$. If $(2p)^2 \leq e\frac{d-1}{d_{\text{eff}}(1)-1}$, we have with probability $1 - \delta$:

$$\|T - \mathbb{E}T\|_{\text{op}} \leq C_{p,1}\sqrt{\frac{d_{\text{eff}}(1) + \ln(d) + \ln(\delta^{-1})}{n}}.$$

If $(2p)^2 > e\frac{d-1}{d_{\text{eff}}(1)-1}$, we have:

$$\|T - \mathbb{E}T\|_{\text{op}} \leq C_{p,\infty}\sqrt{\frac{d + \ln(d) + \ln(\delta^{-1})}{n}}.$$

In the latter case, since $d \leq c_pd_{\text{eff}}(1)$, we still have

$$\|T - \mathbb{E}T\|_{\text{op}} \leq C_p'\sqrt{\frac{d_{\text{eff}}(1) + \ln(d) + \ln(\delta^{-1})}{n}}.$$

Finally, with $C_p = \max(C_p', C_{p,1})$, in both cases, with probability $1 - \delta$

$$\|T - \mathbb{E}T\|_{\text{op}} \leq C_p\sqrt{\frac{d_{\text{eff}}(1) + \ln(d) + \ln(\delta^{-1})}{n}}.$$

Proving Proposition 10 simply requires to use Theorems 2 and 1, and Remark 1.

Appendix D. Statistical Preconditioning

D.1. Large Deviation of Hessians and Proposition 2

Proof We first apply Theorem 1 and Remark 1 with $f_1 = f_2 = Id$ and $f_3 = \ell'' - \ell''(0)$, giving a bound on the following quantity:

$$M := \sup_{x,y,z \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \left\{ (a_i^\top x)(a_i^\top y)(\ell''(a_i^\top z) - \ell''(0)) - \mathbb{E}\left[(a_i^\top x)(a_i^\top y)(\ell''(a_i^\top z) - \ell''(0)) \right] \right\}.$$

Now, notice that:

$$\left\| H_x - \bar{H}_x \right\|_{\mathrm{op}} \leq M + N,$$

where

$$M' := \sup_{x,y \in \mathcal{B}} \frac{\ell''(0)}{n} \sum_{i=1}^{n} \left\{ (a_i^\top x)(a_i^\top y) - \mathbb{E}\left[(a_i^\top x)(a_i^\top y) \right] \right\}.$$

Again, M' can be bounded by Theorem 1 and Remark 1. ∎

D.2. Bregman Gradient Descent: Algorithms and Theoretical Guarantees

Problem Formulation: As mentioned in Section 4.3, we aim at solving the following problem:

$$\min_{x \in \mathbb{R}^d} \Phi(x) := F(x) + \psi(x), \tag{69}$$

for some convex regularizer ψ on a convex domain Dom_ψ, and F $\sigma_{F/\phi}$ relatively strongly convex and $L_{F/\phi}$ relatively smooth with respect to some strongly convex function ϕ (named the preconditioner). We still denote $\kappa_{F/\phi} = \frac{\sigma_{F/\phi}}{L_{F/\phi}}$ their relative condition numbers.

Bregman Gradient Descent: The most classical algorithm in order to solve this optimization problem is *Bregman Gradient Descent* or *Mirror Gradient Descent*. The algorithm is the following, as sketched in Section 4.3.

1. Start from $x_0 \in \mathrm{Dom}_\psi$;

2. For $t \in \mathbb{N}$ and some stepsize $\eta_t > 0$, perform the update:

$$x_{t+1} \in \arg\min_{x \in \mathbb{R}^d} \left\{ \langle \nabla F(x_t), x \rangle + \psi(x) + \frac{1}{\eta_t} D_\phi(x, x_t) \right\}. \tag{70}$$

For $\phi = \frac{\|.\|^2}{2}$, we get classical proximal gradient descent.

Proposition 11 (Bregman Gradient Descent: Convergence Guarantees) *For stepsizes* $\eta_t = \frac{1}{L_{F/\phi}}$ *and* x^* *the minimizer,* (70) *yields:*

$$D_\phi(x_t, x^*) \leq \left(1 - \kappa_{F/\phi}^{-1}\right)^t D_\phi(x_0, x^*).$$

If ϕ *is* μ-*strongly convex, one has:*

$$\|x_t - x^*\|^2 \leq \frac{1}{\mu} \left(1 - \kappa_{F/\phi}^{-1}\right)^t D_\phi(x_0, x^*).$$

Proof For simplicity, we only assume that $\psi = 0$ (no regularization). Let $V_t(x) = \langle \nabla F(x_t), x \rangle + \frac{1}{\eta_t} D_\phi(x, x_t)$. One has $\nabla V_t(x) = \nabla F(x_t) + \frac{1}{\eta_t}(\nabla \phi(x) - \nabla \phi(x_t))$. As $\nabla V_t(x_{t+1}) = 0$, we have:

$$\eta_t \nabla F(x_t) + \nabla \phi(x_{t+1}) - \nabla \phi(x_t) = 0.$$

Moreover:

$$V_t(x^*) - V_t(x_{t+1}) = D_\phi(x_{t+1}, x^*),$$

leading to:

$$D_\phi(x_{t+1}, x^*) = \eta_t \nabla F(x_t)^\top (x^* - x_{t+1}) + D_\phi(x_t, x^*) - D_\phi(x_{t+1}, x_t).$$

In order to study $\eta_t \nabla F(x_t)^\top (x^* - x_{t+1})$, we write:

$$\eta_t \nabla F(x_t)^\top (x^* - x_{t+1}) = \eta_t \nabla F(x_t)^\top (x^* - x_t) + \eta_t \nabla F(x_t)^\top (x_t - x_{t+1}).$$

We have $\eta_t \nabla F(x_t)^\top (x^* - x_t) = \eta_t(F(x^*) - F(x_t) - D_F(x_t, x^*))$. For the second term, we remark that;

$$D_\phi(x_{t+1}, x_t) + D_\phi(x_t, x_{t+1}) = \langle \nabla \phi(x_t) - \nabla \phi(x_{t+1}), x_t - x_{t+1} \rangle$$
$$= \eta_t \langle \nabla F(x_t), x_t - x_{t+1} \rangle.$$

Plugging all this leads to:

$$D_\phi(x_{t+1}, x^*) = D_\phi(x_t, x^*) - \eta_t(D_F(x^*, x_t) + F(x_t) - F(x^*)) + D_\phi(x_t, x_{t+1})$$
$$\leq (1 - \eta_t \sigma_{F/\phi}) D_\phi(x^*, x_t) + D_\phi(x_t, x_{t+1}) - \eta_t D_F(x_t, x^*).$$

Finally, using Bregman co-coercivity yields $D_\phi(x_t, x_{t+1}) \leq \eta_t D_F(x_t, x^*)$ for $\eta_t = 1/L_{F/\phi}$, hence the result. ∎

Acceleration and *SPAG* Algorithm: although Dragomir et al. (2019) prove that the rate of convergence $(1 - \kappa_{F/\phi})$ above is optimal, Hendrikx et al. (2020) propose an acceleration (*SPAG* algorithm) in the sense that asymptotically, one can reach a rate of convergence $(1 - \sqrt{\kappa_{F/\phi}})$. Using stochastic gradients is also possible (Dragomir et al., 2021), with or without variance reduction.

D.3. Bounding Condition Numbers and Consequences on Statistical Preconditioning

For $\mu > 0$ such that $\forall x \in \mathrm{Dom}_\psi, \left\| \nabla^2 f(x) - \nabla^2 F(x) \right\|_{\mathrm{op}} \leq \mu$, we have, for all $x \in \mathrm{Dom}_\psi$ and for $\phi(x) = f(x) + \frac{\|x\|^2}{2}$ (inequalities are taken in terms of symmetric matrices):

$$\nabla^2 F(x) \leq \nabla^2 f(x) + \mu I_d = \nabla^2 \phi(x),$$

giving us $L_{F/\phi} \leq 1$. Then, for relative strong*convexity, as F is λ-strongly convex:

$$\nabla^2 f(x) + \mu I_d \leq \nabla^2 F(x) + 2\mu I_d$$
$$\leq (1 + 2\mu/\lambda) \nabla^2 F(x).$$

Hence, we obtain:

$$\kappa_{F/\phi} \leq 1 + \frac{2\mu}{\lambda}. \tag{71}$$

Proposition 3 bounds this μ with high probability using large deviations on Hessians, in order to apply these considerations.

D.4. Assumption 1 Encompasses Logistic and Ridge Regressions

The statistician has access to feature vectors $a_1, ..., a_n$, and corresponding labels $b_1, ..., b_n$. Linear models (including logistic and ridge regression) take the form $\ell(x, (a_i, b_i)) = \ell_i(a_i^\top x) + \frac{\lambda}{2}\|x\|^2$. Linear regression problems then reduces to the minimization of:

$$\frac{1}{n}\sum_{i=1}^{n} \ell_i(a_i^\top x) + \frac{\lambda}{2}\|x\|^2. \tag{72}$$

It is then to be noticed that for logistic and ridge regressions, functions ℓ_i verify $\ell_i'' = \ell_j''$ for $i, j \in [n]$.

Appendix E. Randomized Smoothing

E.1. Randomized Smoothing: Detailed Algorithm and Convergence Guarantees for General Smoothing Distributions

Detailed Algorithm: we here describe in details how the algorithm works. We recall that $f(x) = \mathbb{E}_a[F(x, a)]$ for $x \in \mathbb{R}^d$. μ is the smoothing distribution, ϕ the known regularizing function. The algorithm uses three sequences of points $(x_t, y_t, z_t)_t$, where y_t is the *query* point: at iteration t, stochastic gradients are computed using y_t. The three sequences evolve according to a dual-averaging algorithm, involving three scalars L_t, θ_t, η_t to control the stepsizes. The smoothed gradients use a sequence of scalars $(u_t)_t$. The algorithm:

1. Computes $y_t = (1 - \theta_t)x_t + \theta_t z_t$.

2. Draws $Z_{1,t}, ..., Z_{m,t}$ *i.i.d.* random variables according to the smoothing distribution μ, for m a fixed integer.

3. Queries the oracle at the m points $y_t + u_t Z_{i,t}, i = 1, ..., m$, yielding stochastic gradients $g_{i,t} \in \partial F(y_t + u_t Z_{i,t}, a_{i,t})$.

4. Computes the average $g_t = \frac{1}{m}\sum_{i=1}^{m} g_{i,t}$.

5. Performs the update:

$$\begin{aligned}
z_{t+1} &= \arg\min_x \left\{ \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}\langle g_\tau, x \rangle + \sum_{\tau=1}^{t} \frac{1}{\theta_\tau}\phi(x) + \frac{1}{2}(L_{t+1} + \frac{\eta_{t+1}}{\theta_{t+1}})\|x\|^2 \right\}, \\
x_{t+1} &= (1 - \theta_t)x_t + \theta_t z_{t+1}.
\end{aligned} \tag{73}$$

Duchi et al. (2012) obtain the following result.

Proposition 12 (Convergence Guarantees for General Smoothing) *Assume that there exist constants L_0 and L_1 such that for all $u > 0$, we have $\mathbb{E}_{Z\sim\mu}[f(x + uZ)] \leq f(x) + L_0 u$, and $\mathbb{E}_{Z\sim\mu}[f(x + uZ)]$ has L_1-Lispchitz continuous gradient. Set $u_t = \theta_t u$, $L_t = L_1/u_t$, and assume that η_t is non-decreasing. Set $\theta_0 = 1$, and $\theta_{t+1} = \frac{2}{1+\sqrt{1+4/\theta_{t-1}^2}}$. Assume that $\|x^*\| \leq R$. Then, for all $T > 0$:*

$$\mathbb{E}[f(x_T) + \phi(x_T) - f(x^*) - \phi(x^*)] \leq \frac{6L_1 R^2}{Tu} + \frac{2\eta_T R^2}{T} + \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{\eta_t}\mathbb{E}[\|e_t\|^2] + \frac{4L_0 u}{T}, \tag{74}$$

where $e_t = \nabla f_{\mu_t}(y_t) - g_t$ is the error in the gradient estimate.

E.2. Isotropic Smoothing: Proof of both Propositions 4 and 5

In the isotropic case, $\mu = \mathcal{N}(0, I_d)$ is the smoothing distribution. We now assume that $F(., a)$ and thus f are L-Lipschitz. Proposition 4 leads to explicit constants L_0 and L_1 in Proposition 12 just above. We prove Proposition 4 here.

Proof For all $x \in \mathbb{R}^d$, one has with Jensen inequality:

$$f(x) \le f^\gamma(x).$$

Then, we obtain $f^\gamma(x) \le f(x) + \gamma L \sqrt{d}$ using Lipschitz continuity of f and $\mathbb{E}\|Z\| \le \sqrt{d}$. In order to prove that f^γ is L/γ-smooth, we need to compute its gradient:

$$\frac{f^\gamma(x+h) - f^\gamma(x)}{h} = \int_{\mathbb{R}^d} \mathrm{d}z f(z)(\mu(z - h/\gamma) - \mu(z))/h$$

$$= -\frac{1}{\gamma}\mathbb{E}_Z[f(x + \gamma Z)Z]$$

when $h \to 0$, where $\mu(z)$ is the density of the smoothing distribution. As in Duchi et al. (2012), we then have that:

$$\|\nabla f^\gamma(x) - \nabla f^\gamma(y)\| \le \frac{1}{\gamma} \int_{\mathbb{R}^d} \mathrm{d}z L_0 |\mu(z - x) - \mu(z - y)|.$$

The end of the proof follows as in their Lemma 10. ∎

Using these properties, and setting $\eta_t = \frac{L\sqrt{t+1}}{R\sqrt{m}}$, $u = Rd^{-1/4}$ and $L_t = L/u_t$, Proposition 5 is obtained by simplifying the expression in Proposition 12 (Duchi et al., 2012).

E.3. Non-Isotropic Smoothing

We now focus on non-isotropic smoothing distributions: $\mu = \mathcal{N}(0, \Sigma')$, for Σ' a symmetric definite positive matrix to determine. We start by proving Proposition 6.
Proof Let $X \sim \mathcal{N}(0, \Sigma')$. Denote, for $i \in [N]$:

$$f_i^\gamma(x) = \mathbb{E}[\ell(a_i^\top(x + \gamma X))]. \tag{75}$$

We have using L-Lipschitz continuity of f:

$$f \le f^\gamma \le f + \gamma L \sqrt{d_{\mathrm{eff}}(1)\sigma_1^2}\sqrt{\mathrm{Tr}(\Sigma')}. \tag{76}$$

Some computations lead to: f_i^γ is differentiable and

$$\nabla f_i^\gamma(x) = -\frac{1}{\gamma}\mathbb{E}[\ell_i(a_i^\top(x + \gamma X))\Sigma'^{-1}X] \tag{77}$$

$$= -\frac{1}{\gamma}\mathbb{E}[\ell_i(a_i^\top(x + \gamma \Sigma'Y))Y] \text{ where } Y \sim \mathcal{N}(0, \Sigma'^{-1}) \tag{78}$$

$$= -\frac{1}{\gamma}\mathbb{E}[\ell_i(a_i^\top(x + \gamma \Sigma'Y))a_i \times \frac{\langle Y, a_i \rangle}{\|a_i\|^2}], \tag{79}$$

as only the contribution of Y in $\mathbb{R}a_i$ is to be taken into account. The form (8) begins to appear here.

$$
\begin{aligned}
\|\nabla f^\gamma(x) - \nabla f^\gamma(y)\| &= \frac{1}{\gamma} \sup_{\|v\|\leq 1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}_Y[(\ell_i(a_i^\top(x+\gamma\Sigma'Y)) - \ell_i(a_i^\top(y+\gamma\Sigma'Y)))\frac{a_i\langle Y, a_i\rangle}{\|a_i\|^2}]^\top v \\
&\leq \frac{1}{\gamma} \sup_{\|v\|\leq 1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}_Y[L|a_i^\top(x-y)||a_i^\top v|\frac{|\langle Y, a_i\rangle|}{\|a_i\|^2}] \\
&= \frac{1}{\gamma} \sup_{\|v\|\leq 1} \frac{1}{N} \sum_{i=1}^N L|a_i^\top(x-y)||a_i^\top v|\frac{\mathbb{E}_Y[|\langle Y, a_i\rangle|]}{\|a_i\|^2} \\
&\leq \frac{1}{\gamma} \frac{\max_i \|a_i\|_{\Sigma'^{-1}}}{\min_i \|a_i\|^2} \sup_{\|v\|\leq 1} \frac{1}{N} \sum_{i=1}^N L|a_i^\top(x-y)||a_i^\top v|.
\end{aligned}
$$

Using Theorem 3, we have that:

$$
\left\| \frac{1}{N} \sum_i a_i a_i^\top - \mathbb{E}a_i a_i^\top \right\|_{\text{op}} \leq C\sqrt{\frac{d_{\text{eff}}(1) + \ln(\delta^{-1})}{N}}.
$$

Thus, with probability $1 - \delta$:

$$
\sup_{\|v\|\leq 1, \|x-y\|\leq 1} \frac{1}{N} \sum_{i=1}^N L|a_i^\top(x-y)||a_i^\top v| \leq C\sigma_1^2\sqrt{\frac{d_{\text{eff}}(1) + \ln(\delta^{-1})}{N}} + \sigma_1^2.
$$

Furthermore, tightness of norms of gaussian random variables around their mean lead to $\min_i \|a_i\|^2 \approx \sigma_1^2 d_{\text{eff}}(1)$. Then, $\max_i \|a_i\|_{\Sigma'^{-1}} \approx \text{Tr}(\Sigma\Sigma'^{-1})$. Minimizing this under $\|\Sigma'\| = C^{te}$ leads to $\Sigma' = \sqrt{\Sigma}$.

All in one, with $\Sigma' = \sqrt{\Sigma}$, we end up with:

$$
f^\gamma \leq f + \gamma L\sqrt{\sigma_1^3 d_{\text{eff}}(1) d_{\text{eff}}(2)},
$$

$$
\|\nabla f^\gamma\|_{\text{Lip}} \leq \frac{L\sigma_1^{1/2} d_{\text{eff}}(2)^{1/2}}{\gamma d_{\text{eff}}(1)} \left(1 + C\sqrt{\frac{d_{\text{eff}}(1)\ln(d) + \ln(\delta^{-1})}{N}} \right).
$$

∎

Then, Proposition 7 is obtained in the same way as Proposition 5, setting $\eta_t = \frac{L\sqrt{t+1}}{R\sqrt{m}}$, $u = \frac{R}{\frac{L\sigma_1^{1/2} d_{\text{eff}}(2)^{1/2}}{\gamma d_{\text{eff}}(1)}\left(1+C\sqrt{\frac{d_{\text{eff}}(1)\ln(d)+\ln(\delta^{-1})}{N}}\right)}$ and $L_t = L/u_t$. We just replaced $d^{-1/4}$ by a less ergonomic, yet smaller expression.

Appendix F. Robustness of Two-Layered Neural Networks with Polynomial Activation

In this section, we present two applications of our chaining bounds, that played a role of toy problem. Bubeck et al. (2020) conjecture that two-layered neural networks interpolating *generic*

data (defined below)) have a Lipshitz constant that must be lower bounded by $\sqrt{\frac{n}{k}}$ where n is the number of data points, and k the number of neurons.

Definition 11 (Generic Data 1) *Data $(x_i, y_i)_{1 \le i \le n}$ are generic if they are* i.i.d. *and if y_i are centered random signs, x_i centered gaussians of covariance I_d/d.*

We aim at generalizing some of their results in a non-isotropic framework. We thus define in another way generic data.

Definition 12 (Generic Data 2) *Data $(x_i, y_i)_{1 \le i \le n}$ are generic if they are* i.i.d. *and if y_i are centered random signs, x_i centered gaussians of covariance $\Sigma/(\sigma_1^2 d_{\text{eff}}(1))$.*

Definition 13 (Tensor) *A tensor of order $p \in \mathbb{N}^*$ is an array $T = (T_{i_1,\ldots,i_p})_{i_1,\ldots,i_p \in [d]} \in \mathbb{R}^{dp}$. T is said to be of rank 1 if it can be written as:*

$$T = u_1 \otimes \cdots \otimes u_p$$

for some $u_1, \ldots, u_p \in \mathbb{R}^p$.
Scalar product is defined as:

$$\langle T, S \rangle = \sum_{i_1,\ldots,i_p} T_{i_1,\ldots,i_p} S_{i_1,\ldots,i_p}.$$

We define the operator norm of a tensor as:

$$\|T\|_{\text{op}} = \sup_{\|x_1 \otimes \ldots \otimes x_p\| \le 1} \langle T, x_1 \otimes \ldots \otimes x_p \rangle.$$

Definition 14 (Two Layered Neural Network) *A two-layered neural network with inputs in $\mathbb{R}^d$, k neurons and Lipschitz non-linearity ψ is a function of the form:*

$$f(x) = \sum_{\ell=1}^{k} a_\ell \psi(w_\ell^\top x + b_\ell). \tag{80}$$

Conjecture: A two-layered neural network f that fits generic data $(x_i, y_i)_{1 \le i \le n}$ must satisfy, with high probability when $d \to \infty$, for some constant $c > 0$ (Bubeck et al., 2020):

$$\text{Lip}_{\mathcal{S}}(f) \ge c\sqrt{\frac{n}{k}}. \tag{81}$$

That conjecture is not proven (just in some very particular cases and regimes). However, we propose to adapt considerations made with polynomial activation functions ψ in the isotropic regime, to the non-isotropic one. Our aim is however not to link k the number of neurons, to n the number of observations. Indeed, we believe that in this model of *generic data* (both isotropic and non-isotropic ones), dimensionality plays a core role in the Lipschitz constant of f. If one considers different dimensions, n and k being fixed, it is natural to believe that, due to the concentration of gaussians in small dimensions, the Lipschitz constant will be bigger for smaller dimensions. Furthermore, adding non-isotropy and introducing effective dimensions should not change this replacing dimensions by effective ones, hence the following proposition, which aims at giving insights on the impact of (effective) dimension on the Lipschitz constant of f.

Proposition 13 *Assume that $(x_i, y_i)_{1 \leq i \leq n}$ are generic data (Definition 12). Let $\psi(t) = \sum_{q=0}^{p} \alpha_q t^q$ and f a two-layered neural network with activation function ψ such that $\forall i, f(x_i) = y_i$. Then, if $d_{\mathrm{eff}}(1) \leq c_0 n$ for some $c_0 > 0$, with probability $1 - a \exp(-b d_{\mathrm{eff}}(1))$:*

$$\mathrm{Lip}_{\mathcal{S}}(f) \geq C_p \frac{n}{d^{p-1} d_{\mathrm{eff}}(1)}. \tag{82}$$

Either the bound is not tight (likely), or achieving better Lipschitz constants for f is easier with non isotropic data. Both are possible however, and suggest the importance of effective dimensions in the robustness of neural networks. The proof below follows the same steps as in Bubeck et al. (2020).

Proof Note that there exist $T_0, ..., T_p$ tensors such that T_q is of order q and:

$$f(x) = \sum_q \langle T_q, x^{\otimes q} \rangle.$$

Let $\Omega_q = \sum_{i=1}^{n} y_i x_i^{\otimes q}$. We have:

$$n = \sum_i y_i f(x_i) = \sum_{q=0}^{p} \langle T_q, \Omega_q \rangle.$$

Hence, there exists $q \geq 1$ such that $\langle T_q, \Omega_q \rangle \geq c_p n$.

$$
\begin{aligned}
c_p n &\leq \langle T_q, \Omega_q \rangle \\
&\leq \|\Omega_q\|_{\mathrm{op}} \|T_q\|_{op,*} \\
&\leq d^{q-1} \|\Omega_q\|_{\mathrm{op}} \|T_q\|_{op},
\end{aligned}
$$

using that $\|T_q\|_{op,*} \leq d^{q-1} \|T_q\|_{op}$. We then notice that:

$$\frac{1}{n} \|\Omega_q\|_{\mathrm{op}} = \sup_{z_1,...,z_q \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} y_i \prod_{k=1}^{q} x_i^\top z_k,$$

which is exactly the same form as Y in (8), except for the y_i's. However, we need a centered bound here: we will use the y_i for this. Let $n^+ = \{i : y_i = 1\}$, and $n^- = \{i : y_i = -1\}$. We have:

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} y_i \prod_{k=1}^{q} x_i^\top z_k &= \frac{1}{n} \left(\sum_{i \in n^+} y_i \prod_{k=1}^{q} x_i^\top z_k - \mathbb{E} \left[\prod_{k=1}^{q} x_1^\top z_k \right] \right) \\
&+ \frac{1}{n} \left(\sum_{i \in n^-} y_i \prod_{k=1}^{q} x_i^\top z_k + \mathbb{E} \left[\prod_{k=1}^{q} x_1^\top z_k \right] \right) \\
&+ \frac{n^+ - n^-}{n} \mathbb{E} \left[\prod_{k=1}^{q} x_1^\top z_k \right].
\end{aligned}
$$

With probability $1 - C \exp(-c\tau)$ (with respect to the y_i's):

$$|n^+ - n^-| \leq \sqrt{n\tau}.$$

We then have with probability $1 - Ce^{-c\tau}$:

$$\frac{n^+ - n^-}{n}\mathbb{E}\left[\prod_{k=1}^{q} x_1^\top z_k\right] \leq \sqrt{\frac{\tau}{n}}.$$

The first two terms can be bounded using Theorem 3 and the fact that under the event $|n^+ - n^-| \leq \sqrt{n\tau}$, we have $n^+, n^- \geq \frac{n - \sqrt{n\tau}}{2} \geq n/3$ if $\tau \leq c_0 n$. Using these considerations gives with probability $1 - Ce^{-c\tau} - 4\delta$:

$$\|\Omega_q\|_{\text{op}} \leq \sqrt{\frac{\tau}{n}} + 2c'\sqrt{\frac{\ln(\delta^{-1}) + d_{\text{eff}}(1)}{n}}.$$

Taking $\delta = \exp(-d_{\text{eff}}(1))$ and $\tau = d_{\text{eff}}(1)$ yields:

$$\frac{1}{n}\|\Omega_q\|_{\text{op}} \leq C'\sqrt{\frac{d_{\text{eff}}(1)}{n}}.$$

Finally, we have, with probability $1 - a\exp(-d_{\text{eff}}(1))$, if $d_{\text{eff}}(1) \leq c_0 n$:

$$\|T_q\|_{\text{op}} \geq \frac{c_p n}{d^{q-1}\|\Omega_q\|_{\text{op}}} \geq C_q \frac{n}{d^{q-1} d_{\text{eff}}(1)}.$$

Then, by observing that the Lipschitz constant of f on the unit ball is lower bounded by $\|T_q\|_{\text{op}}$ for any q (with a constant multiplicative factor, using Markov brother's inequality), we obtain with probability $1 - a\exp(-bd_{\text{eff}}(p)\ln(d))$:

$$\text{Lip}_{\mathcal{S}}(f) \geq C_p \frac{n}{d^{p-1} d_{\text{eff}}(1)}. \tag{83}$$

■