The Rare Eclipse Problem in Quantised Random Embeddings: A Matter of Consistency?

Valerio Cambareri, Chunlei Xu and Laurent Jacques
ISG-Group, ICTEAM/ELEN, Université catholique de Louvain, Louvain-la-Neuve, Belgium.
E-mail: {valerio.cambareri, chunlei.xu, laurent.jacques}@uclouvain.be

Abstract—We study the problem of verifying when two disjoint closed convex sets remain separable after the application of a quantised random embedding, as a means to ensure exact classification from the signatures produced by this non-linear dimensionality reduction. An analysis of the interplay between the embedding, its quantiser resolution and the sets’ separation is presented in the form of a convex problem; this is completed by its numerical exploration in a special case, for which the phase transition corresponding to exact classification is easily computed.

I. PROBLEM STATEMENT

Non-linear dimensionality reduction techniques play an important role in simplifying statistical learning on very large-scale datasets. Among such techniques, we focus on quantised random embeddings obtained by a non-linear map \( A \) applied to \( x \in \mathbb{K} \subset \mathbb{R}^d \), that is

\[
y = A(x) := \Phi \delta(A(x) + \xi)
\]

with \( \Phi \in \mathbb{R}^{m \times n} \) a random sensing matrix, \( \Phi(\cdot) := \delta(x) \) a uniform scalar quantiser of resolution \( \delta > 0 \) (applied component-wise), and the signature \( y \in \delta \mathbb{Z}^m \). In (1), the divergence \( \xi \sim \mathcal{U}(0, \delta) \) is a well-known means to stabilise the action of the quantiser [1], [2].

The non-linear map (1) is a non-adaptive dimensionality reduction that yields compact signatures for storage and transmission, while retaining a notion of quasi-isometry that enables the approximation of \( x \) [2], [3]. Consequently, distance-based learning tasks preserve their accuracy if run on \( A(K) \) rather than \( K \), provided some requirements are met on \( m, \delta \), the distribution of \( \Phi \) and the “dimension” of \( K \) as measured, e.g., by its Gaussian mean width \( w(K) := \sup_{x \in \mathbb{K}} ||q(x)|| \) with \( q \sim \mathcal{N}(0,1) \) (see, e.g., [2]). In this context we aim to show that, given two classes described by some sets \( C_1, C_2 \subset \mathbb{K} : C_1 \cap C_2 = \emptyset \) and \( x \in C_1 \cup C_2 \subset \mathbb{K} \), classifying whether \( x \) belongs to \( C_1 \) or \( C_2 \) is still possible from \( y = A(x) \).

The authors are funded by the F.R.S.-FNRS and by the project ALTERSense (MIS-FNRS). All authors have equally contributed to the realisation of this work.

We here leverage the quantised restricted isometry property (QRIP) introduced in [2] to estimate \( \eta \) and the conditions on \( m \). The QRIP establishes some conditions on \( m \) that ensure \( 1 \geq \frac{1}{\delta} ||A(x') - A(x'')|| \geq (\varepsilon' - \varepsilon)||x' - x''|| - c \delta \geq (\varepsilon' - \varepsilon) \sigma - c \delta \varepsilon'' := \frac{H}{\delta} \) for some controllable distortions \( \varepsilon, \varepsilon' > 0 \), with \( \sigma \leq ||x|| \) and some constants \( c, \delta > 0 \). Thus \( A(x') \neq A(x'') \) if \( H > 0 \). In particular, we deduce the following proposition whose proof is postponed to an extended version of this work.

Proposition 1. In the setup of Prob. 2, let \( r_1 := \text{rad}(C_1), r_2 := r_1 + r_2, \) and \( A \) defined in (1) with \( \delta > 0 \). Given \( \eta \in (0, 1) \), if

\[
m \geq \left( \frac{w^2_0 + \eta \delta^2}{\varepsilon'} \right)(1 + \log(1 + \frac{m}{\eta^2 \delta})) + \frac{w^2_0 \log \frac{1}{\eta}}{\varepsilon^2} \]  

then \( \psi_3 \geq 1 - \eta \).

Numerically testable but stronger conditions ensuring \( \psi_3 > 1 - \eta \) in Prob. 2 can be deduced as follows. We first note that if \( \Phi z = 0 \) for a given \( \Phi \) and any \( z \in C' \), i.e., \( \ker(\Phi) \cap C' \neq \emptyset \), then \( \psi_5 = 0 \) for all \( \delta > 0 \) since then \( A(x') = A(x'') \) induceds \( ||\Phi z||_\infty \leq \delta \) for \( z := x' - x'' \in C' \), proving \( \psi_5 = 0 \) in \( C' \), \( ||\Phi z||_\infty < \delta \geq 1 - \eta \) will solve Prob. 2 since \( \psi_3 \geq \psi_5 \).

We define accordingly a consistency margin \( \tau := ||\Phi z||_\infty \) with \( z := \min_{z \in C' : \Phi z \neq 0} ||\Phi z||_\infty \) s.t. \( z \in C' := C_1 - C_2 \),

i.e., as a property of \( \Phi \) and \( C' \) so that, if \( \tau > 0 \), we necessarily have \( A(x') \neq A(x'') \) for \( x', x'' \) in different classes, i.e., \( \psi_5 = 0 \).

Intuitively, \( z' \) is related to the minimal separation \( \sigma \) between \( C_1 \) and \( C_2 \).

Note that (3) is clearly convex if \( C' \) and \( C'' \) are convex. We anticipate that the construction of a certificate for this problem will provide a bound on \( \tau > 0 \) when \( C' \) is known, and analyse an exemplary case afterwards.

II. NUMERICAL TEST FOR TWO DISJOINT \( \ell_2 \)-BALLS

We consider the simple, yet broadly applicable convex case of two balls \( C_1 = r_1 \mathbb{B}^n + c_1, C_2 = r_2 \mathbb{B}^n + c_2 \). Then, \( C' = r \mathbb{B}^n + c \) with \( c := c_1 - c_2 \) and \( r := r_1 + r_2 \).\textsuperscript{3} In this context \( ||c|| = \sigma = r \) and \( \eta \approx \frac{\pi}{2} \leq \frac{1}{\pi} < \frac{1}{\tau} \). (1) [4].

By solving random instances of this problem\textsuperscript{2} w.r.t. \( \Phi \sim \mathcal{N}(0, n, 1) \), with \( n = 2^n \) and \( m \in \{2^n, 2^{n+1} \} \), we are able to compute the consistency margin for each \( \Phi \) on \( C' \), which is varied by fixing \( r = 2 \) and taking \( \sigma = ||c|| = \frac{r}{\tau} \). Then, we collect \( \tau_{\text{min}} \), i.e., the smallest \( \tau \) resulting from 2\textsuperscript{n} trials for each configuration (Fig. 2a), and also estimate on the same trials the probability \( \bar{p}_5 = \mathbb{P}[\tau > \delta = 1] \) in Fig. 2b.

Fig. 2a reports several level curves of \( \tau_{\text{min}} \). For each curve, the event \( A(C_1)' \cap A(C_2) = \emptyset \) holds if \( \delta = \tau_{\text{min}} \). While this condition is necessary but not sufficient, these level curves are compatible with the points \( \left( \frac{m}{\eta^2 \delta}, \sigma \right) \) that satisfy the rule \( m \approx \frac{\pi^2}{4} \eta n \) (up to log factors) induced by (2) in Prob. 1. Fig. 2b displays a sharp phase transition in the contours of \( \bar{p}_5 \). Despite the fact that \( \bar{p}_5 \approx \bar{p}_5 \), the contours are also approximately aligned with the iso-probability curves that can be deduced from (2), i.e.,

\[ m - \frac{\pi^2}{4} \eta n \approx C \log \left( \frac{1}{\eta} \right), \]

with \( \bar{p}_5 \approx 1 - \eta \) for some \( C, \sigma > 1 \).

III. CONCLUSION AND OPEN QUESTIONS

The fundamental limits of learning tasks with embeddings are being tackled in several studies [5]–[8]. Our contribution expands the requirements for exact classification from the signatures produced by two closed convex sets after quantised random embeddings. We shall also specify this analysis to low-complexity structured sets \( K \) (e.g., selecting disjoint “clusters” of sparse signals).

\textsuperscript{2}By uniformity of \( \ker(\Phi) \), \( \Phi \sim \mathcal{N}(0, 1) \) over the Grassmannian at the origin, it is legitimate to fix a randomly drawn direction \( c/||c|| \) for the simulations.
Figure 1. Geometrical intuition on the quantised eclipse problem for two disjoint $\ell_2$-balls and $n = 3$, $m = 2$: (left) $C_1$ and $C_2$ are projected on $\Phi$, identified by the unit vectors $\varphi_1, \varphi_2$: on these directions, we construct the lattice $\delta \mathbb{Z}^m$, with a shift $\xi$ of the origin due to dithering; the finite sets $A(C_1), A(C_2)$ are also reported, along with the consistency margin $\tau$; (right) ensuring that $A(C_1) \cap A(C_2) = \emptyset$ requires that any $z \in C^-$ is so that its image under $\Phi$ has $\|\Phi z\|_{\infty} > \tau$; taking the smallest of such values on the difference set yields the consistency margin, which is $\tau = 0$ when $\text{Ker}(\Phi) \cap C^- \neq \emptyset$.

Figure 2. Empirical phase transitions of the quantised eclipse problem for the case of two disjoint $\ell_2$-balls; for several random instances of $\Phi$ and as a function of $\sigma$ and the dimensionality reduction rate $\frac{m}{n}$, we report (a) the contours of $\log_2 \tau_{\text{min}}$; (b) the contours of $\bar{\rho}_\delta = P[\tau > \delta] \approx 1 - \eta$ for $\delta = 1$. In (a), the level curves of $\tau_{\text{min}}$ are compatible, up to log factors, with the points $\{ (\frac{m}{n}, \sigma) : m \approx \sigma^2 n/\delta^2 \}$ deduced from (2) in Prop. 1. In (b), the level curves of $\bar{\rho}_\delta$ are also approximately aligned with the iso-probability curves $m - c \frac{\log^2 n}{\sigma^2} \approx C \log\left(\frac{1}{\eta}\right)$, also deduced from (2), once we set $\bar{\rho}_\delta \approx 1 - \eta \in \{0.25, 0.5, 0.75, 0.9, 0.95\}$ for some $C, c > 1$.

REFERENCES


