ON PROJECTIONS OF METRIC SPACES

Mark Kozdoba∗

ABSTRACT. Let $X$ be a metric space and let $\mu$ be a probability measure on it. Consider a Lipschitz map $T : X \to \mathbb{R}^n$, with Lipschitz constant $\leq 1$. Then one can ask whether the image $TX$ can have large projections on many directions. For a large class of spaces $X$, we show that there are directions $\phi \in S^{n-1}$ on which the projection of the image $TX$ is small on the average (in $L_2(\mu)$), with bounds depending on the dimension $n$ and the eigenvalues of the Laplacian on $X$. Our results can be viewed as a multidimensional extension of the concentration of measure principle.

1 Introduction

Let $(X, m)$ be a metric space and let $\mu$ be a probability measure on $X$. Consider a Lipschitz map $T : X \to \mathbb{R}^n$, with $\|T\|_{Lip} \leq 1$, where $\mathbb{R}^n$ is taken with the standard inner product $\langle \cdot, \cdot \rangle$ and the corresponding Euclidean norm $|\cdot|$.

Let us define another inner product on $\mathbb{R}^n$ by setting
\[
[\phi, \theta] = \int \langle \phi, Tx \rangle \langle \theta, Tx \rangle d\mu(x)
\]
for $\theta, \phi \in \mathbb{R}^n$ and denote the associated norm by
\[
\|\theta\|_{L_2} := \left(\int \langle Tx, \theta \rangle^2 d\mu(x)\right)^{\frac{1}{2}}.
\]

This norm is known as the covariance structure of the push-forward measure $T \mu$, and we shall regard $\|\theta\|_{L_2}$ as measuring the size of the projection of the image of $X$ onto the direction $\theta$. Similar notion of size is used, for instance, in the archetypical dimensionality reduction technique in Machine Learning, the Principal Component Analysis (PCA) (see [3] for an exposition). In PCA, the data is projected onto a subspace $V \subset \mathbb{R}^n$, of a fixed dimension $k$, such that the variation of the data on $V$ is maximal among all subspaces with that dimension. One ideally expects then that the projection onto $V$ is the real source of the data, and the projection onto $V^\perp$ comes from noise, with $\theta \in S^{n-1} \cap V^\perp$ having small $\|\theta\|_{L_2}$ norm.

In this article we show that when a metric space is mapped into $\mathbb{R}^n$, the geometry of $X$ can provide bounds on the size of the projections of the image. To see why this might be the case, it is instructive first to take a look at an example. Take $X$ to be the unit interval $[0, 1]$, with Lebesgue measure, and let $T : [0, 1] \to \mathbb{R}^n$ be a Lipschitz map, i.e. we consider

∗markk@tx.technion.ac.il
curves of length at most one in $\mathbb{R}^n$ (here and in what follows, Lipschitz will mean “having Lipschitz constant $\leq 1$”). By considering a few pictures, it is clear that if $n$ is large then there should be a direction with a small projection and we are interested in quantifying this phenomenon. It turns out that this particular situation was investigated already by Gohberg and Krein, in a different formulation and context (related to spectral properties of smooth kernels), see the book [7], Chapter III.10.3.

**Theorem 1** ([7]). *There is an absolute constant $c > 0$, such that for every $n > 1$ and every $T : [0,1] \rightarrow \mathbb{R}^n$ with $\|T\|_{\text{Lip}} \leq 1$, there is $\theta \in S^{n-1}$ such that $\|\theta\|_{L^2} \leq c \cdot n^{-\frac{3}{2}}$.***

Our objective is to extend this result to a more general class of metric spaces with measure on which an appropriate notion of a derivative can be defined. Namely, we shall discuss finite spaces $X$ which are graphs with the shortest path metric and with measure which is the stationary measure of the random walk, as described below. While this setting makes the presentation simpler, the statement and the proof of our result, Theorem 2, can be repeated with straightforward modifications in the setting of Riemannian manifolds. In addition, the approach we take does not explicitly use the stationarity of the measure. This assumption is used mainly because in the stationary case the associated Laplacians are well investigated operators, for which explicit bounds can be provided. Details are given in Section 2.

Let $X$ be a finite set and let $E \subset X \times X$ be a set of edges such that $(X,E)$ is a connected undirected graph without loops. For $x,y \in X$, write $x \sim y$ iff $(x,y) \in E$ and let $d(x)$ denote the degree of $x$. We endow $X$ with the shortest path metric and we set $\mu$ to be the stationary distribution of a simple nearest neighbour random walk on $(X,E)$, $\mu(x) = \frac{d(x)}{2|E|}$. Denote by $L(X)$ the space of real-valued functions on $X$. The graph Laplacian on $L(X)$ is defined by

$$\triangle(f)(x) = 2\left(f(x) - \frac{1}{\deg(x)} \sum_{y \sim x} f(y)\right).$$

(3)

For our purposes, it suffices to note that this is a self-adjoint non-negative operator, with a one dimensional kernel consisting of constant functions. Some of these properties will be derived in Section 2, while the full details on the graph Laplacian can be found, for instance, in [4]. We denote by $\{\lambda_i\}_{i=1}^{|X|-1}$ the sequence of non-zero eigenvalues of $\triangle$ in non-decreasing order, including multiplicities.

**Theorem 2.** *There is an absolute constant $c > 0$ such that for every graph $X$ as above, every $n > 1$ and every Lipschitz map $T : X \rightarrow \mathbb{R}^n$, there is a direction $\theta \in S^{n-1}$ such that $\|\theta\|_{L^2} \leq c \cdot n^{-\frac{3}{2}} \lambda_1^{-\frac{1}{2}}$.***

**Example 3 (Discrete Space).** Let $(X,m)$ be a metric space such that $m(x,y) = 1$ for all $x \neq y$, and let $\mu$ be the uniform probability on $X$. Then for every Lipschitz $T : X \rightarrow \mathbb{R}^n$ there is $\theta \in S^{n-1}$ such that

$$\|\theta\|_{L^2} \leq \frac{c}{\sqrt{n}}.$$  

(4)
Note that, interestingly enough, the size of the space, \(|X|\), does not appear in this bound (except that, of course, for \(n > |X|\) we always have \(\theta \in S^{n-1} \) with \(\|\theta\|_{L_2} = 0\)). That is, one can not increase the minimal projection size of \(X\) in, say, \(\mathbb{R}^20\), by adding more points. To prove (4), note that \((X, m)\) corresponds to a clique graph, and the Laplacian has a single non-zero eigenvalue, which is of the order of a constant and has multiplicity \(|X| - 1\).

Example 4 (Combinatorial Cube). Here \(X\) is the set \(\mathcal{C}^d = \{0, 1\}^d\), with the Hamming metric

\[ m(x, y) = |\{i \in \{1, \ldots, d\} \mid x_i \neq y_i\}| \]

where \(x = x_1 \ldots x_d, y = y_1 \ldots y_d \in \mathcal{C}^d\) and \(\mu\) is again taken to be the uniform probability measure. The non-zero eigenvalues of the Laplacian on \(X\) are \(4k^d\) with multiplicity \(\binom{d}{k}\), \(k = 1, \ldots, d\) (See [5]). The \(n\)-th smallest eigenvalue is therefore of the order \(\frac{\log(n)}{d}\) and we obtain that for every Lipschitz \(T : X \rightarrow \mathbb{R}^n\), there is \(\theta \in S^{n-1}\) such that

\[ \|\theta\|_{L_2} \leq c \cdot \frac{\sqrt{d}}{\sqrt{n} \cdot \sqrt{\log(n)}}. \]

We shall in fact prove a slightly stronger statement then Theorem 2. To state it, we require an additional bit of notation. Define the covariance matrix of the measure \(T \mu\) in \(\mathbb{R}^n\) to be the operator \(\text{Cov}_{T \mu} : \mathbb{R}^n \rightarrow \mathbb{R}^n\),

\[ \text{Cov}_{T \mu} = \int Tx \otimes Tx \ d\mu(x), \]

where for \(y \in \mathbb{R}^n\), \(y \otimes y\) denotes an operator on \(\mathbb{R}^n\) which acts by \((y \otimes y)(x) = \langle x, y \rangle y\). Note that for every \(\theta, \phi \in \mathbb{R}^n\),

\[ [\theta, \phi] = \langle \text{Cov}_{T \mu} \theta, \phi \rangle. \]

Hence \(\text{Cov}_{T \mu}\) is the non-negative matrix which connects the bilinear form \([\cdot, \cdot]\) with the natural inner product \(\langle \cdot, \cdot \rangle\). Let \(\{\alpha_i\}_{i=1}^n\) be the eigenvalues of \(\text{Cov}_{T \mu}\) in non-increasing order. Then, with the rest of the notation as in Theorem 2, we have:

**Theorem 5.** There is an absolute constant \(c > 0\) such that for every graph \(X\) as above, every \(n > 1\) and every Lipschitz map \(T : X \rightarrow \mathbb{R}^n\), which satisfies in addition that

\[ \int Tx \ d\mu(x) = 0, \]

for every \(i = 1, \ldots, n\),

\[ \alpha_i \leq c \cdot i^{-1} \lambda_{[n/2]}^{-1}. \]

Note that for \(T\) satisfying the mean-zero assumption (6), Theorem 5 indeed implies Theorem 2. This can be seen by choosing \(\theta\) of Theorem 2 to be the eigenvector of \(\text{Cov}_{T \mu}\) corresponding to the eigenvalue \(\alpha_n\). The removal of assumption (6) is simple and will be discussed in Section 2.

Our approach to Theorem 2 extends the argument in [7] and, naturally, Theorem 1 can also be seen as a consequence of the principle of Theorem 2. Indeed, consider the
Laplacian on $[0, 1]$, $\Delta(f) = -\frac{\partial^2 f}{\partial x^2}$. The minus sign in front of $\frac{\partial^2 f}{\partial x^2}$ is taken for consistency with definition (3), making $\Delta(f)$ a non-negative operator. This Laplacian has eigenspaces spanned by $\{\sin 2\pi k x, \cos 2\pi k x \mid k \in \mathbb{N}\}$, with eigenvalues $\lambda_k \approx k^2$, up to multiplicative constants. Hence the decay rate of $n^{-\frac{3}{2}} \cdot (n^2)^{-\frac{1}{2}} = n^{-\frac{3}{2}}$ in Theorem 1.

Finally, we note that Theorem 5 can be viewed as a multidimensional extension of the (weak) concentration of measure phenomenon. The concentration of measure phenomenon states, roughly, that the distribution of a Lipschitz function on a metric measure space tends to be more concentrated around the mean than what would be suggested by space’s diameter. Many results, utilizing different geometric parameters of the space were developed to quantify this phenomenon. We refer to [10] and [9] for further details on the concentration of measure. For the purposes of the present article, let us state a classic result by Alon and Milman, [1], which derives concentration of measure on a graph in terms of the first eigenvalue of the Laplacian.

**Theorem 6** (Alon and Milman, 1985). There exist constants $c, C > 0$, such that for every graph $X$, every Lipschitz $f : X \to \mathbb{R}$, and every $t > 0$:

$$
\mu \left( |f - \mathbb{E}f| > \sqrt{\lambda_1^{-1}} \cdot t \right) \leq Ce^{-ct}.
$$

(8)

The sub-exponential inequality (8) implies in particular the following, weaker, $L_2$ form: There is a constant $C > 0$ such that for every $X$ and every Lipschitz $f : X \to \mathbb{R}$,

$$
(\text{Var}_\mu(f))^{\frac{1}{2}} \leq C \cdot \sqrt{\lambda_1^{-1}}.
$$

(9)

In Riemannian manifolds, this inequality is a special case of the Poincaré inequality (see [6]) and it has a short independent proof. We give a short argument for the graph case in Section 2.

In M. Gromov’s terminology, [8], both inequalities (8) and (9) state that the observable diameter of $X$, viewed through Lipschitz functions, is $\sqrt{\lambda_1^{-1}}$. In other words, a Lipschitz image of a space $X$ in $\mathbb{R}$ will have most of its points within an interval of diameter of order $\sqrt{\lambda_1^{-1}}$, which is typically much smaller then the diameter of $X$. In this light, one can ask what can be said about observations of $X$ with values in higher dimensional Euclidean spaces. Theorem 5 provides one answer to this question, in terms of the Euclidean structure of the image. Thus Theorem 5 and can be viewed as an extension of inequality (9) to higher dimensional valued observations.

Let us illustrate this on the example of the cube $C^d$. The diameter of $C^d$ is $d$ and we have $\sqrt{\lambda_1^{-1}} = \sqrt{d}$. Let $T : X \to \mathbb{R}^n$ be a Lipschitz mapping and assume for simplicity that $\int T x \ d\mu(x) = 0$. For every $\theta \in S^{n-1}$, define a function $f_\theta : X \to \mathbb{R}$ by

$$
f_\theta(x) = \langle Tx, \theta \rangle.
$$

Since for every $\theta \in S^{n-1}$, $f_\theta$ is a Lipschitz function, we can conclude from inequality (9) that

$$
\|f_\theta\|_{L_2} = \text{Var}_\mu f_\theta \leq C\sqrt{d}
$$
and this is the best possible bound which is uniform over all $\theta$. However, from Theorem 5 we know, for instance, that this bound can not be sharp for all $\theta$. In particular, from Theorem 2, there exists $\theta \in S^{n-1}$ with $\text{Var}_{\mu} f_\theta \leq c \cdot \frac{\sqrt{d}}{\sqrt{n} \sqrt{\log(n)}}$.

Our results in this article deal with the $L_2$ theory. That is, we give bounds on the second moments of the functions of the form $f_\theta$ under the distribution induced by $T$, and these bounds are better then what can be obtained by application of the usual, one dimensional concentration inequalities. Whether similar results can be obtained for higher moments, or for full sub-exponential inequalities of the form (8) remains an interesting open question.

2 Proof

Let us begin by introducing additional some notions and notation that is used in the proof of Theorem 2. Fix a graph $(X, E)$ with the stationary measure $\mu$ on it. The notion of a gradient on a graph is folklore, although not particularly frequent in the literature and the notation varies considerably. We thus recall the definitions and some important properties.

Define an inner product on $L(X)$ by

$$[f, g]_X = \int f(x) g(x) d\mu(x).$$

Denote by $L(E)$ the set of real valued functions on the set of edges, $E$, and let $\nu$ be the uniform probability measure on $E$. We equip $L(E)$ with the inner product

$$[f, g]_E = \int f(e) g(e) d\nu.$$

Fix an arbitrary orientation on $E$, i.e. for each edge $e = \{x, y\}$ choose in an arbitrary way an enumeration $v^1_e, v^2_e$ of the vertices of the edge. We can express this enumeration as a $\{+1, -1\}$-valued function $o$,

$$o(x, y) = \begin{cases} 
+1 & \text{if } x = v^1_e \text{ and } y = v^2_e \\
-1 & \text{if } x = v^2_e \text{ and } y = v^1_e 
\end{cases}$$

defined for pairs of vertices $x, y$ which are connected by an edge $e$. The gradient operator on $X$ is defined by $\nabla : L(X) \rightarrow L(E)$,

$$(\nabla f)(e) = f(v^1_e) - f(v^2_e).$$

Equivalently, for $e = \{x, y\}$,

$$(\nabla f)(e) = o(x, y) (f(x) - f(y)),$$

The Laplacian on the graph $(X, E)$ is defined as $\triangle = \nabla^* \circ \nabla$, where $\nabla^*$ is the Hilbert space adjoint of the $\nabla$. Let us first compute the explicit expression for the Laplacian for arbitrary
measures $\mu$ and $\nu$ on $X$ and $E$. By the definition of adjoints, the Laplacian needs to satisfy

$$[\nabla^* \circ \nabla f, g]_X = [\nabla f, \nabla g]_E \quad (10)$$

for every $f, g \in L(X)$. By choosing $g$ to be the delta function at vertex $x$, we get

$$(\triangle f)(x) \cdot \mu(x) = \sum_e o(x, y) \cdot (f(x) - f(y)) \cdot o(x, y) \nu(e) = \sum_e (f(x) - f(y)) \nu(e),$$

where the sum is over all edges $e = \{x, y\}$ connected to the vertex $x$. Note, in particular, that the Laplacian does not depend on the orientation, but does depend on the choice of the measures $\mu$ and $\nu$. We shall be interested in the case where $\mu(x) = \frac{deg(x)}{2|E|}$ is the stationary measure of the graph, and $\nu(e) = \frac{1}{|E|}$ is the uniform measure. These measures produce the classical definition (3), which has the advantage that this particular Laplacian is well-studied and its eigenvalues are known in many cases (see, for instance, [4] and [5]).

Our proof of Theorem 2, however, uses the Laplacian only through equation (10), and does not explicitly assume stationarity.

Denote by $L_0(X)$ the subspace of $L(X)$ that is orthogonal to the constant functions. Since $\text{Ker} \triangle$ is the space of the constant functions, the Laplacian is invertible on $L_0(X)$ and hence the gradient is also an invertible operator from $L_0(X)$ onto its image. The inverse of the gradient will be of importance in what follows and we denote it by $\Gamma : \text{Im}(\nabla) \to L_0(X)$.

With the preliminaries on the Laplacian in place, we give a short argument for the proof of inequality (9). This inequality, however, will not be used in the rest of this section. First note that for every $f \in L(X)$,

$$\text{Var}_\mu f = [(f - \mathbb{E}f), (f - \mathbb{E}f)]_X \leq \lambda_1^{-1} [\triangle f, f]_X. \quad (11)$$

Indeed, write $f$ as an orthogonal decomposition $f = \sum_{i \geq 0} f_i$ where $f_i$ is contained in the eigenspace of the eigenvalue $\lambda_i$ of $\triangle$ (where $f_0 = \mathbb{E}f$ is the mean, corresponding to eigenvalue 0). Then the left hand-side of (11) is $\sum_{i > 0} [f_i, f_i]$ and the right hand-side is $\sum_{i > 0} \lambda_i [f_i, f_i]$. Since $\lambda_1$ is the smallest non-zero eigenvalue, (11) follows. Now, since the norm of the derivative satisfies by definition

$$[\nabla f, \nabla f]_E = [\triangle f, f]_X \quad (12)$$

and since $[\nabla f, \nabla f]_E \leq 1$ for Lipschitz functions, inequality (9) follows.

We now briefly recall the notion of singular values of an operator and some related singular value inequalities. We refer to [2] or [7] for full details and proofs. If $V$ and $W$ are (say, finite dimensional) Hilbert spaces, and $A : V \to W$ is a linear operator, then $A^*A$ is a non-negative self-adjoint operator. Let $\{\lambda_i(A^*A)\}_{i=1}^{\dim V}$ be the eigenvalues of $A^*A$, in non-increasing order, with multiplicities. Then singular values of the operator $A$ are the non-increasing sequence $s_i(A) = \sqrt{\lambda_i(A^*A)}$.

The singular values of $A$ have a simple geometric interpretation. If $B_2(V)$ denotes the unit ball of $V$, then the image $A(B_2(V))$ is an ellipsoid in $W$ and $s_i(A)$ are precisely the lengths of the principal axes of this ellipsoid.
The singular values satisfy $s_i(A) = s_i(A^*)$ and singular values of a composition can be bounded by the following inequality due to Ky Fan. Let $A : V \to W$ and $B : W \to U$ be two operators on Hilbert spaces. Then, for every $i, j \geq 1$,

$$s_{i+j-1}(BA) \leq s_i(A)s_j(B). \quad (13)$$

Finally, the Hilbert-Schmidt norm of an operator is defined by

$$\|A\|_{HS} = \sqrt{\text{tr}A^*A} = \left(\sum_is_i^2(A)\right)^{\frac{1}{2}}.$$ 

Note that since $s_i$ is a non-increasing sequence,

$$s_i(A) \leq \frac{\|A\|_{HS}}{\sqrt{i}} \quad (14)$$

for all $i$.

Before proceeding with the proof of Theorem 2, it will be convenient to introduce an additional assumption on the Lipschitz map $T : X \to \mathbb{R}^n$, namely that

$$\int T(x)d\mu = 0. \quad (15)$$

With conditions of Theorem 2 and this assumption, we will show that there is $\theta \in S^{n-1}$ such that

$$\|\theta\|_{L_2} = c \cdot n^{-\frac{1}{2}} \lambda_{[n/2]}^{-\frac{1}{2}}. \quad (16)$$

This implies the result for arbitrary Lipschitz $T : X \to \mathbb{R}^n$. Indeed, assume that $v := \int T(x)d\mu \neq 0$. Denote by $P$ the orthogonal projection onto the $n-1$ dimensional space orthogonal to $v$. Then the composition $P \circ T$ is a Lipschitz map into that space and

$$\int (P \circ T)(x)d\mu = 0,$$ 

so (16) can be applied in dimension $n-1$.

**Proof of Theorem 2.** As mentioned above, we assume that condition (15) holds. Let $D_T$ denote the gradient of the map $T$, i.e.

$$D_T(e) = T(v^1_e) - T(v^2_e).$$

Since $T$ is Lipschitz, $D_T$ is bounded, i.e. $|D_T(e)| \leq 1$ for all $e \in E$.

For every $\theta \in \mathbb{R}^n$ consider the function on $X$, $f_\theta(x) = \langle \theta, Tx \rangle$. By (15), $f_\theta \in L_0(X)$ and clearly $(\nabla f_\theta)(e) = \langle \theta, D_T(e) \rangle$ and

$$f_\theta = \Gamma(\langle \theta, D_T \rangle).$$

Let

$$B_{L_2}(X) = \left\{ f : X \to \mathbb{R} \mid [f, f]_X \leq 1 \right\}$$

with $[\cdot, \cdot]_X$ denoting the $L_2$ inner product.
denote the unit ball in $L(X)$ and write $\|\theta\|_{L_2}$ in a dual form:

$$\|\theta\|_{L_2} = \sup_{g \in B_2(X)} \int g(x) \langle \theta, Tx \rangle \, d\mu.$$  

Then

$$\int g(x) \langle \theta, Tx \rangle \, d\mu = [g, f_\theta]_X = [g, \Gamma \circ \nabla f_\theta]_X = [\Gamma^* g, \nabla f_\theta]_E.$$  

Next, write

$$[\Gamma^* g, \nabla f_\theta]_E = \int (\Gamma^* g)(e) \cdot \langle \theta, DT(e) \rangle \, d\nu(e) = \langle \theta, \int (\Gamma^* g)(e) \cdot DT(e) \, d\nu(e) \rangle.$$  

Denote by $\widetilde{D} : L(E) \to \mathbb{R}^n$ the operator that acts by

$$\widetilde{D}u = \int u(e) \cdot DT(e) \, d\nu.$$  

With this notation,

$$\|\theta\|_{L_2} = \sup_{g \in B_2} \langle \theta, \widetilde{D} \circ \Gamma^* g \rangle = \sup_{\phi \in \mathcal{E}} \langle \theta, \phi \rangle$$  

where $\mathcal{E} = \widetilde{D} \circ \Gamma^* B_2$ is the dual (in $\mathbb{R}^n$) ellipsoid of the $\|\cdot\|_{L_2}$ norm.

Our aim is to bound the quantity

$$\inf_{\theta \in S^{n-1}} \|\theta\|_{L_2}.$$  

By (17), this quantity equals the length of the smallest principal axis of the ellipsoid $\mathcal{E}$. Since the lengths of the principal axes are the singular values of $\widetilde{D} \circ \Gamma^*$, we bound the singular values of this operator.

The singular values of $\Gamma^*$ are given, $s_i(\Gamma^*) = \lambda_i^{-\frac{1}{2}}$, where $\lambda_i$ are the non-zero eigenvalues of the Laplacian (in non-decreasing order, so that $s_i(\Gamma^*)$ do not increase). To bound the singular values of $\widetilde{D}$, we show that

$$\|\widetilde{D}\|_{HS} \leq 1.$$  

Indeed, one readily verifies that

$$\widetilde{D} \circ \widetilde{D}^*(\theta) = \int \langle \theta, DT(e) \rangle DT(e) \, d\nu.$$  

For fixed $e \in E$, the trace of an operator $\theta \mapsto \langle \theta, DT(e) \rangle DT(e)$ is $|DT(e)|^2$, which satisfies

$$|DT(e)| \leq 1$$  

by the Lipschitz assumption. Therefore (18) follows by averaging. Now, by (14), $s_i(\widetilde{D}) \leq \frac{1}{\sqrt{i}}$, and an application of (13) (with $i = j = n/2$) completes the proof.
References


