Sketching, Embedding, and Dimensionality Reduction for Information Spaces

Amirali Abdullah
amirali@cs.utah.edu
University of Utah

Ravi Kumar
ravi.k53@gmail.com
Google, Inc.

Andrew McGregor
mcgregor@cs.umass.edu
University of Massachusetts-Amherst

Sergei Vassilvitskii
sergeiv@google.com
Google, Inc.

Suresh Venkatasubramanian
suresh@cs.utah.edu
University of Utah

Abstract

Sketching and dimensionality reduction are powerful techniques for speeding up algorithms for massive data. However unlike the rich toolbox available for the ℓ_2^2 distance, there are no robust results of this nature known for most popular information-theoretic measures.

In this paper we show how to embed information distances like the χ^2 and Jensen-Shannon divergences efficiently in low dimensional spaces while preserving all pairwise distances. We then prove a dimensionality reduction result for the Hellinger, Jensen–Shannon, and χ^2 divergences that preserves the information geometry of the distributions, specifically, by retaining the simplex structure of the space. While our first result already implies these divergences can be explicitly embedded in the Euclidean space, retaining the simplex structure is important because it allows us to do inferences in the reduced space.

We also show that these divergences can be sketched efficiently (i.e., up to a multiplicative error in sublinear space) in the aggregate streaming model. This result is exponentially stronger than known upper bounds for sketching these distances in the strict turnstile streaming model.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

1 Introduction

The space of *information distances* includes many distances that are used extensively in data analysis, for example, the well-known Bregman divergences, the α-divergences, and the f-divergences. These divergences are used in statistical tests and estimators [7], as well as in image analysis [23], computer vision [16, 21], and text analysis [11, 12]. They were introduced by Csiszár [10], and, in the most general case, also include measures such as the Hellinger, Jensen–Shannon (JS), and χ^2 divergences. In this work we focus on a subclass of the f-divergences that admits embeddings into some (possibly infinite-dimensional) Hilbert space, with a specific emphasis on the Jensen–Shannon (JS) divergence.

To work with the geometry of these divergences effectively at scale and in high dimensions, we need algorithmic tools that can provide provably high quality approximate representations of the geometry. The techniques of sketching, embedding, and dimensionality reduction have evolved as ways of dealing with this problem.

A *sketch* for a set of points with respect to a property P is a function that maps the data to a small summary from which property P can be evaluated, albeit with some approximation error. In the context of data streams, where information arrives online, linear sketches are especially useful for estimating a derived property in a fast and compact way.[1] Complementing sketching, *embedding* techniques are one-to-one mappings that transform a collection of points lying in a

[1] Indeed Li, Nguyen, and Woodruff [20] show that any optimal one-pass streaming sketch algorithm in the turnstile model can be reduced to a linear sketch with logarithmic space overhead.

space X to another (presumably easier) space Y, while approximately preserving pairwise distances between points. *Dimensionality reduction* is a special kind of embedding where Y has two properties: it has much smaller dimension than X and it has the same structure as X. These techniques can be used in an almost "plug-and-play" fashion to speed up many algorithms in data analysis: for example for near neighbor search (and classification), clustering, and closest pair calculations.

Unfortunately, while these tools have been well developed for norms like ℓ_1 and ℓ_2, they are less studied for information distances. This is not just a theoretical concern: information distances are semantically more suited to many tasks in machine learning [11, 12, 16, 21, 23], and building the appropriate algorithmic toolkit to manipulate them efficiently would greatly expand the places where they can be used.

1.1 Our contributions

Dimensionality reduction. Our main result is a *structure-preserving* dimensionality reduction for information distances, where we wish to preserve not only the distances between pairs of points (distributions), but also the underlying *simplicial* structure of the space, so that we can continue to interpret coordinates in the new space as probabilities. The notion of a structure-preserving dimensionality reduction is implicit when dealing with normed spaces (since we always map a normed space to another), but requires an explicit mapping when dealing with more structured spaces. We prove an analog of the classical JL-Lemma for JS, Hellinger, and χ^2 divergences (Theorem 5.6): given a set of n points in a d-dimensional simplex, there is an explicit mapping of the points to an $O(\frac{1}{\epsilon^2}\log n)$-dimensional simplex that preserves these divergences between every pair of points up to a multiplicative factor of $(1+\varepsilon)$.

This result extends to "well-behaved" f-divergences (See Section 5 for a precise definition). Moreover, the dimensionality reduction is constructive for any divergence with a finite-dimensional kernel (such as the Hellinger divergence), or an infinite-dimensional Kernel that can be sketched in finite space, as we show is feasible for the JS and χ^2 divergences. We note f-divergences that are not well-behaved (e.g., ℓ_1) do not admit a similar dimensionality reduction [4, 9, 19, 25].

Sketching information divergences. In establishing the above, we show how to embed information distances into ℓ_2^2, which implies sketchability.

We first show that a set of points in the d-dimensional simplex under the JS or χ^2 divergences can be deterministically embedded into $\tilde{O}(\frac{d^2}{\varepsilon})$ dimensions under ℓ_2^2 with ε additive error (Theorem 4.4). Next, we show an analogous embedding result for the multiplicative error case; however, the embedding is randomized (Theorem 4.6). For both these results, applying the Euclidean JL-Lemma can further reduce the target dimension. We also give the first results on sketching of infinite-dimensional kernels, a challenge posed in Avron, Nguyen, and Woodruff [6].

Applications to data streams. Data streams are a standard model of computation for processing large datasets in small space, in which data is revealed incrementally. The model comes in two flavors: the aggregate model, where each element of the stream gives the ith coordinate of a point, and the turnstile model where each element of the stream serves to incrementally increase the ith coordinate of a point. The primary difference in the two models is that in the former a coordinate is updated exactly once, whereas in the latter it may receiver multiple updates. See the books by Muthukrishnan [22] or Aggarwal [2], or the article by Wright [28] for more background on the field and popular techniques. Previously Guha, Indyk, and McGregor [15] showed that these divergences cannot be sketched even up to a constant factor in sublinear space under the turnstile model, and asked whether the same lower bound holds in the aggregate model.

We resolve their question and show that a point set in the d-dimensional simplex under the JS or χ^2 divergences can be deterministically embedded into $\tilde{O}(\frac{d^2}{\varepsilon})$ dimensions under ℓ_2^2 with ε additive error (Corollary 4.5).

Our techniques. The unifying approach of our three results—sketching, embedding into ℓ_2^2, and dimensionality reduction—is to analyze carefully the infinite-dimensional kernel of the information divergences. Quantizing and truncating the kernel yields the sketching result, sampling repeatedly from it produces an embedding into ℓ_2^2. Finally given such an embedding, we show how to perform dimensionality reduction by proving that each of the divergences admits a region of the simplex where it is similar to ℓ_2^2. To the best of our knowledge, this is the first result that explicitly uses the kernel representation of these information distances to build approximate geometric structures. While the *existence* of a kernel for the Jensen–Shannon distance was well-known, this structure had never been exploited to give algorithms with robust theoretical guarantees.

We note that the results for the χ^2 divergence are similar to those for the JS divergence, and we defer the details to the full version [1].

2 Related Work

The works of Fuglede and Topsøe [13] and Vedaldi and Zisserman [27] study embeddings of information divergences into an infinite-dimensional Hilbert space by representing them as integrals along a one-dimensional curve in $\mathbb{C}$. Vedaldi and Zisserman give an explicit formulation of this kernel for JS and χ^2 divergences, for which a discretization (by quantizing and truncating) yields an additive error embedding into a finite-dimensional ℓ_2^2. However, they do not obtain explicit bounds on the target space dimension needed to derive a sketching algorithm; furthermore, they do not get a multiplicative approximation.

Kyng, Phillips, and Venkatasubramanian [18] show a limited structure-preserving dimensionality reduction result for the Hellinger distance. Their approach works by showing that if the input points lie in a specific region of the simplex, then a standard random projection will keep the points in a lower-dimensional simplex while preserving the distances approximately. Unfortunately, this region is a small ball centered in the interior of the simplex, which shrinks with the dimension. This is in sharp contrast to our work here, where the input points are unconstrained.

One can achieve a multiplicative approximation in the aggregate streaming model for information divergences that have a finite-dimensional embedding into ℓ_2^2. For instance, Guha et al. [14] observe that for the Hellinger distance that has a trivial such embedding, sketching is equivalent to sketching ℓ_2^2 and hence may be done up to a $(1+\varepsilon)$-multiplicative approximation in $\frac{1}{\varepsilon^2}\log n$ space. This immediately implies a constant-factor approximation of JS and χ^2 divergences in the same space, but no bounds were known prior to our work for a $(1+\varepsilon)$-sketching result for JS and χ^2 divergences in *any* streaming model.

There has been a wide range of work done on embedding in other spaces as well. Rahimi and Recht [24] embed shift-invariant kernels into ℓ_2^2 via random Fourier features; however their result does not hold for the more general kernels we consider in this paper. Avron, Nguyen, and Woodruff [6] give a sketching technique for the polynomial kernel and pose the open question of obtaining similar results for infinite-dimensional kernels; we address this question in this paper. One of the most famous results in dimension reduction is the Johnson–Lindenstrauss (JL) Lemma, which states that any set of n points in ℓ_2^2 can be embedded into $O\left(\frac{\log n}{\varepsilon^2}\right)$ dimensions in the same space while preserving pairwise distances to within $(1\pm\varepsilon)$. The general literature on sketching and embeddability in normed spaces is too extensive to be reviewed here; see, for example, Andoni et al. [5].

3 Background

In this section, we define precisely the class of information divergences that we work with, and their specific properties that allow us to obtain sketching, embedding, and dimensionality results. For what follows Δ_d denotes the *d-simplex*:

$$\Delta_d = \{(x_1,\ldots,x_d) \mid \sum_i x_i = 1 \text{ and } x_i \geq 0, \forall i\}.$$

We will assume in this paper that all distributions are defined over a finite ground set $[n] = \{1,\ldots,n\}$.

Definition 3.1. *The* Jensen–Shannon *(JS),* Hellinger*, and χ^2 divergences between distributions p and q are defined as* $JS(p,q) = \sum_i p_i \log\frac{2p_i}{p_i+q_i} + q_i\log\frac{2q_i}{p_i+q_i}$, $He(p,q) = \sum_i(\sqrt{p_i}-\sqrt{q_i})^2$ *and* $\chi^2(p,q) = \sum_i \frac{(p_i-q_i)^2}{p_i+q_i}$ *respectively.*

Definition 3.2 (Regular distance)**.** *On domain X, we call a distance function $D : X\times X \to \mathbb{R}$* regular *if there exists a feature map $\phi : X \to V$, where V is a (possibly infinite-dimensional) Hilbert space, such that:*

$$D(x,y) = \|\phi(x)-\phi(y)\|^2 \quad \forall x,y \in X.$$

The work of [13] and [27] prove that the JS divergence is regular: they give a feature map $\phi(x) = \int_{-\infty}^{+\infty}\Psi_x(\omega)\,\mathrm{d}\omega$, where $\Psi_x(\omega) : \mathbb{R}\to\mathbb{C}$ is given by

$$\Psi_x(\omega) = \exp(i\omega\ln x)\sqrt{\frac{2x\,\mathrm{sech}(\pi\omega)}{(\ln 4)(1+4\omega^2)}}. \tag{3.1}$$

Hence we have for $x, y \in \mathbb{R}$, $\mathrm{JS}(x,y) = \|\phi(x)-\phi(y)\|^2 = \int_{-\infty}^{+\infty}\|\Psi_x(\omega)-\Psi_y(\omega)\|^2\,\mathrm{d}\omega$. The infinite-dimensional "embedding" for a given distribution $p\in\Delta_d$ is then the concatenation of the functions $\phi(p_i)$, i.e., $\phi(p) = (\phi_{p_1},\ldots,\phi_{p_d})$.

4 Embedding JS into ℓ_2^2

We present two algorithms for embedding JS into ℓ_2^2. The first is deterministic and gives an additive error approximation whereas the second is randomized but yields a multiplicative approximation in an offline setting. The advantage of the first algorithm is that it can be realized in the streaming model, and if we make the standard assumption of polynomial precision in the streaming input, yields a $(1+\varepsilon)$-multiplicative approximation as well in this setting.

We derive some terms in the kernel representation of $\mathrm{JS}(x,y)$ that will be convenient. First, the explicit

formulation in (3.1) yields that for x, $y \in \mathbb{R}$:

$$\begin{aligned}\mathrm{JS}(x,y) &= \int_{-\infty}^{+\infty} \left\| e^{i\omega \ln x}\sqrt{\frac{2x\,\mathrm{sech}(\pi\omega)}{(\ln 4)(1+4\omega^2)}} \right. \\ &\qquad \left. - e^{i\omega \ln y}\sqrt{\frac{2y\,\mathrm{sech}(\pi\omega)}{(\ln 4)(1+4\omega^2)}}\right\|^2 \mathrm{d}\omega \\ &= \int_{-\infty}^{+\infty} \left(\frac{2\,\mathrm{sech}(\pi\omega)}{(\ln 4)(1+4\omega^2)}\right) \\ &\qquad \cdot \|\sqrt{x}e^{i\omega\ln x} - \sqrt{y}e^{i\omega \ln y}\|^2\,\mathrm{d}\omega.\end{aligned}$$

For convenience, we now define

$$\begin{aligned}h(x,y,\omega) &= \|\sqrt{x}e^{i\omega\ln x} - \sqrt{y}e^{i\omega\ln y}\|^2 \\ &= (\sqrt{x}\cos(\omega\ln x) - \sqrt{y}\cos(\omega\ln y))^2 \\ &\quad + (\sqrt{x}\sin(\omega\ln x) - \sqrt{y}\sin(\omega\ln y))^2,\end{aligned}$$

and $\kappa(\omega) = \frac{2\,\mathrm{sech}(\pi\omega)}{(\ln 4)(1+4\omega^2)}$.

We can then write $\mathrm{JS}(p,q) = \sum_{i=1}^d f_J(p_i,q_i)$ where

$$\begin{aligned}f_J(x,y) &= \int_{-\infty}^{\infty} h(x,y,\omega)\kappa(\omega)\,\mathrm{d}\omega \\ &= x\log\left(\frac{2x}{x+y}\right) + y\log\left(\frac{2y}{x+y}\right).\end{aligned}$$

It is easy to verify that $\kappa(\omega)$ is a distribution, i.e., $\int_{-\infty}^{\infty}\kappa(\omega)d\omega = 1$.

4.1 Deterministic embedding

We will produce an embedding $\phi(p) = (\phi_{p_1},\ldots,\phi_{p_d})$, where each ϕ_{p_i} is an integral that can be discretized by quantizing and truncating carefully.

Algorithm 1: Embed $p \in \Delta_d$ under JS into ℓ_2^2.

Input: $p = \{p_1,\ldots,p_d\}$ where coordinates are ordered by arrival, ε

Output: A vector of length $O\left(\frac{d^2}{\varepsilon}\log\frac{d}{\varepsilon}\right)$

$J \leftarrow \lceil \frac{32d}{\varepsilon}\ln\left(\frac{8d}{\varepsilon}\right)\rceil$

for $-J \le j \le J$ **do** $\quad w_j \leftarrow j\times\varepsilon/32d$

for $1 \le i \le d, -J \le j \le J$ **do**

$a^p_{(2J+1)i+j} \leftarrow \sqrt{p_i}\cos(\omega_j\ln p_i)\sqrt{\int_{\omega_j}^{\omega_{j+1}}\kappa(\omega)d\omega}$

$b^p_{(2J+1)i+j} \leftarrow \sqrt{p_i}\sin(\omega_j\ln p_i)\sqrt{\int_{\omega_j}^{\omega_{j+1}}\kappa(\omega)d\omega}$

return a^p *concatenated with* b^p

To analyze Algorithm 1, we first obtain bounds on the function h and its derivative.

Lemma 4.1. *For* $0 \le x,y, \le 1$, *we have* $0 \le h(x,y,\omega) \le 2$ *and* $\left|\frac{\partial h(x,y,\omega)}{\partial\omega}\right| \le 16$.

Proof. Clearly $h(x,y,\omega) \ge 0$. Furthermore, since $0 \le x,y \le 1$, we have

$$h(x,y,\omega) \le \left|\sqrt{x}e^{i\omega\ln x}\right|^2 + \left|\sqrt{y}e^{i\omega\ln y}\right|^2 = x+y \le 2.$$

$$\begin{aligned}\text{Next, } &\left|\frac{\partial h(x,y,\omega)}{\partial\omega}\right| \\ &= \left|2\left(\sqrt{x}\cos(\omega\ln x) - \sqrt{y}\cos(\omega\ln y)\right)\right. \\ &\qquad \left.\left(-\sqrt{x}\sin(\omega\ln x)\ln x + \sqrt{y}\sin(\omega\ln y)\ln y\right)\right. \\ &\qquad +2\left(\sqrt{x}\sin(\omega\ln x) - \sqrt{y}\sin(\omega\ln y)\right) \\ &\qquad \left.\left(\sqrt{x}\cos(\omega\ln x)\ln x - \sqrt{y}\cos(\omega\ln y)\ln y\right)\right| \\ &\le \left|2\left(\sqrt{x}+\sqrt{y}\right)\left(\sqrt{x}\ln x + \sqrt{y}\ln y\right)\right| \\ &\qquad +2\left|\left(\sqrt{x}+\sqrt{y}\right)\left(\sqrt{x}\ln x + \sqrt{y}\ln y\right)\right| \;\le\; 16,\end{aligned}$$

where the last inequality follows since $\max_{0\le x\le 1}|\sqrt{x}\ln x| < 1$. □

The next two steps are useful to approximate the infinite-dimensional continuous representation by a finite-dimensional discrete representation by appropriately truncating and quantizing the integral.

Lemma 4.2 (Truncation)**.** *For* $t \ge \ln(4/\varepsilon)$, $f_J(x,y) \ge \int_{-t}^{t} h(x,y,\omega)\kappa(\omega)d\omega \ge f_J(x,y) - \varepsilon$.

Proof. The first inequality follows since $h(x,y,\omega) \ge 0$. For the second inequality, we use $h(x,y,\omega) \le 2$:

$$\begin{aligned}&\int_{-\infty}^{-t} h(x,y,\omega)\kappa(\omega)d\omega + \int_{t}^{\infty} h(x,y,\omega)\kappa(\omega)d\omega \\ &\le 4\int_t^{\infty}\kappa(\omega)d\omega < 4\int_t^{\infty}\frac{4e^{-\pi\omega}}{\ln 4}d\omega < 4e^{-t} \le \varepsilon,\end{aligned}$$

where the last line follows if $t \ge \ln(4/\varepsilon)$. □

Define $\omega_i = \varepsilon i/16$ for $i \in \pm\mathbb{Z}$ and $\tilde{h}(x,y,\omega) = h(x,y,\omega_i)$ where $i = \max\{j \mid \omega_j \le \omega\}$.

Lemma 4.3 (Quantization)**.** *For any* a,b, $\int_a^b h(x,y,\omega)\kappa(\omega)d\omega = \int_a^b \tilde{h}(x,y,\omega)\kappa(\omega)d\omega \pm \varepsilon$.

Proof. First note that

$$|\tilde{h}(x,y,\omega) - h(x,y,\omega)| \le \left(\frac{\varepsilon}{16}\right)\cdot\max_{x,y\in[0,1],\omega}\left|\frac{\partial h(x,y,\omega)}{\partial\omega}\right| \le \varepsilon\,.$$

Hence, $\left|\int_{-a}^{b}\tilde{h}(x,y,\omega)\kappa(\omega)d\omega - \int_{-a}^{b}h(x,y,\omega)\kappa(\omega)d\omega\right| \le \left|\int_{-a}^{b}\varepsilon\kappa(\omega)d\omega\right| \le \varepsilon$. □

Given a real number z, define vectors $\mathbf{v}^z$ and $\mathbf{u}^z$ indexed by $i \in \{-i^*,\ldots,-2,-1,0,1,2,\ldots i^*\}$ where

$i^* = \lceil 16\varepsilon^{-1}\ln(4/\varepsilon)\rceil$ by:

$$\mathbf{v}^z = \sqrt{z}\cos(\omega_i \ln z)\sqrt{\int_{\omega_i}^{\omega_{i+1}} \kappa(\omega)d\omega},$$

$$\mathbf{u}^z = \sqrt{z}\sin(\omega_i \ln z)\sqrt{\int_{\omega_i}^{\omega_{i+1}} \kappa(\omega)d\omega},$$

and note that $(\mathbf{v}_i^x - \mathbf{v}_i^y)^2 + (\mathbf{u}_i^x - \mathbf{u}_i^y)^2 = h(x,y,\omega_i)\int_{\omega_i}^{\omega_{i+1}}\kappa(\omega)d\omega$. Thus, $\|\mathbf{v}^x - \mathbf{v}^y\|_2^2 + \|\mathbf{u}^x - \mathbf{u}^y\|_2^2 = \int_{w_{-i^*}}^{w_{i^*+1}} \tilde{h}(x,y,\omega)\kappa(\omega)d\omega = \int_{w_{-i^*}}^{w_{i^*+1}} h(x,y,\omega)\kappa(\omega)d\omega \pm \varepsilon = \int_{-\infty}^{\infty} h(x,y,\omega)\kappa(\omega)d\omega \pm 2\varepsilon = f_J(x,y) \pm 2\varepsilon$, where the second to last equality follows from Lemma 4.3 and the last equality follows from Lemma 4.2, since $\min(|w_{-i^*}|, w_{i^*+1}) \geq \ln(4/\varepsilon)$.

Define the vector $\mathbf{a}^p$ to be the vector generated by concatenating $\mathbf{v}^{p_i}$ and $\mathbf{u}^{p_i}$ for $i \in [d]$. Then it follows that $\|\mathbf{a}^p - \mathbf{a}^q\|_2^2 = \mathrm{JS}(p,q) \pm 2\varepsilon d$. Hence we have reduced the problem of estimating $\mathrm{JS}(p,q)$ to ℓ_2 estimation. Rescaling $\varepsilon \leftarrow \varepsilon/(2d)$ ensures the additive error is ε while the length of the vectors $\mathbf{a}^p$ and $\mathbf{a}^q$ is $O\left(\frac{d^2}{\varepsilon}\log\frac{d}{\varepsilon}\right)$.

Theorem 4.4. *Algorithm 1 embeds a set P of points in Δ_d under JS into $O\left(\frac{d^2}{\varepsilon}\log\frac{d}{\varepsilon}\right)$ dimensions under ℓ_2^2 with ε additive error.*

Note that using the JL-Lemma, the dimensionality of the target space can be reduced to $O\left(\frac{\log|P|}{\varepsilon^2}\right)$. Theorem 4.4, along with the AMS sketch of [3], and the standard assumption of polynomial precision immediately implies:

Corollary 4.5. *There is an algorithm that works in the aggregate streaming model to approximate JS to within $(1+\varepsilon)$-multiplicative factor using $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\log d\right)$ space.*

As noted earlier, this is the first algorithm in the aggregate streaming model to obtain a $(1+\varepsilon)$-multiplicative approximation to JS, which contrasts against linear space lower bounds for the same problem in the turnstile streaming model [15].

4.2 Randomized embedding

In this section we show the following result for multiplicative error.

Theorem 4.6. *A set P of points under the JS or χ^2 divergence can be embedded into $\ell_2^{\bar{d}}$ with $\bar{d} = O\left(\frac{n^2d^3}{\varepsilon^2}\right)$ with $(1+\varepsilon)$ multiplicative error.*

This can be reduced further to $O(\frac{\log n}{\varepsilon^2})$ dimensions by simply applying the Euclidean JL-Lemma. Furthermore, if we ignore precision constraints on sampling from a continuous distribution in a streaming algorithm, then this also would yield a sketching bound of $O(d^3\varepsilon^{-2})$ for a $(1+\varepsilon)$ multiplicative approximation.

Our approach is to sample from the kernel repeatedly to obtain each coordinate of the embedding so that the final ℓ_2^2 distance is an unbiased estimate of the divergence. While this is similar in spirit to Rahimi and Recht [24], proving this estimate is sufficiently concentrated is our main technical challenge.

For fixed $x, y, \in [0,1]$, we consider the random variable T that takes the value $h(x,y,\omega)$ with probability $\kappa(\omega)$. (Recall that $\kappa(\cdot)$ is a distribution.) We compute the first and second moments of T.

Theorem 4.7. $E[T] = f_J(x,y), \mathrm{var}[T] \leq 36(E[T])^2$.

Proof. The expectation follows immediately from the definition, $E[T] = \int_{-\infty}^{\infty} h(x,y,\omega)\kappa(\omega)d\omega = f_J(x,y)$. To bound the variance it will be useful to define the function $f_H(x,y) = (\sqrt{x}-\sqrt{y})^2$ corresponding to the one-dimensional Hellinger distance that is related to $f_J(x,y)$. We now state two claims regarding $f_H(x,y)$ and $f_\chi(x,y)$:

Claim 4.1. *For all $x, y \in [0,1]$, $f_H(x,y) \leq 2f_J(x,y)$.*

Proof. Let $f_\chi(x,y) = \frac{(x-y)^2}{x+y}$ correspond to the one-dimensional χ^2 distance. Then, we have

$$\frac{f_\chi(x,y)}{f_H(x,y)} = \frac{(x-y)^2}{(x+y)(\sqrt{x}-\sqrt{y})^2} = \frac{(\sqrt{x}+\sqrt{y})^2}{x+y} = \frac{x+y+2\sqrt{xy}}{x+y} \geq 1\ .$$

This shows that $f_H(x,y) \leq f_\chi(x,y)$. To show $f_\chi(x,y) \leq 2f_J(x,y)$ we refer the reader to [26, Section 3]. Combining these two relationships gives us our claim. □

Claim 4.2. *For all $x, y \in [0,1], \omega \in \mathbb{R}$, $h(x,y,\omega) \leq f_H(x,y)(1+2|\omega|)^2$.*

Proof. Without loss of generality, assume $x \geq y$.

$$\begin{aligned}
&\sqrt{h(x,y,\omega)} \\
&= |\sqrt{x}\cdot e^{i\omega\ln x} - \sqrt{y}\cdot e^{i\omega\ln y}| \\
&\leq |\sqrt{x}\cdot e^{i\omega\ln x} - \sqrt{y}\cdot e^{i\omega\ln x}| \\
&\quad + |\sqrt{y}\cdot e^{i\omega\ln x} - \sqrt{y}\cdot e^{i\omega\ln y}| \\
&= |\sqrt{x}-\sqrt{y}| + \sqrt{y}\cdot|e^{i\omega\ln x} - e^{i\omega\ln y}| \\
&= |\sqrt{x}-\sqrt{y}| + \sqrt{y}\cdot 2\cdot|\sin(\omega\ln(x/y)/2)| \\
&\leq \sqrt{f_H(x,y)} + \sqrt{y}\cdot 2\cdot|\omega\ln(\sqrt{x/y})| \\
&\leq \sqrt{f_H(x,y)} + \sqrt{y}\cdot 2\cdot|\sqrt{x/y}-1|\cdot|\omega| \\
&= \sqrt{f_H(x,y)} + 2\sqrt{f_H(x,y)}\cdot|\omega|. \quad \square
\end{aligned}$$

These claims allow us to bound the variance:

$$\begin{aligned}\text{var}[T] \le E[T^2] &\le f_H(x,y)^2 \int_{-\infty}^{\infty} (1+2|\omega|)^4 \kappa(\omega) d\omega \\ &= f_H(x,y)^2 \cdot 8.94 \;\; < \;\; 36 f_J(x,y)^2. \; \square\end{aligned}$$

Algorithm 2: Embeds point $p \in \Delta_d$ under JS into ℓ_2^2.

Input: $p = \{p_1, \ldots, p_d\}$, ε
Output: A vector of length $O\left(n^2 d^3 \varepsilon^{-2}\right)$
$s \leftarrow \lceil 36 n^2 d^2 \varepsilon^{-2} \rceil$
for $1 \le j \le s$ **do** $\quad \omega_j \sim \kappa(\omega)$
for $1 \le i \le d, 1 \le j \le s$ **do**
$\quad a^p_{s(i-1)+j} \leftarrow \left(\sqrt{p_i}\cos(\omega_j \ln p_i)/\sqrt{s}\right)$
$\quad b^p_{s(i-1)+j} \leftarrow \left(\sqrt{p_i}\sin(\omega_j \ln p_i)/\sqrt{s}\right)$
return a^p *concatenated with* b^p

Let $\omega_1, \ldots, \omega_s$ be s independent samples chosen according to $\kappa(\omega)$. For any distribution p on $[d]$, define vectors $\mathbf{v}^p, \mathbf{u}^p \in \mathbb{R}^{sd}$ where, for $i \in [d], j \in [s]$,

$$\mathbf{v}^p_{i,j} = \sqrt{p_i}\cdot\cos(\omega_j \ln p_i)/\sqrt{s}; \mathbf{u}^p_{i,j} = \sqrt{p_i}\cdot\sin(\omega_j \ln p_i)/\sqrt{s}.$$

Let $\mathbf{v}^p_i$ be a concatenation of $\mathbf{v}^p_{i,j}$ and $\mathbf{u}^p_{i,j}$ over all $j \in [s]$. Then note that $E[\|\mathbf{v}^p_i - \mathbf{v}^q_i\|_2^2] = f_J(p_i, q_i)$ and $\text{var}[\|\mathbf{v}^p_i - \mathbf{v}^q_i\|_2^2] \le 36(f_J(p_i,q_i))^2/s$. Hence, for $s = 36n^2d^2\varepsilon^{-2}$, by the Chebyshev bound,

$$\Pr[|\|\mathbf{v}^p_i - \mathbf{v}^q_i\|_2^2 - f_J(p_i,q_i)| \ge \varepsilon f_J(x,y)] \le \frac{36}{\varepsilon^2 s} = (nd)^{-2}.$$

By a union bound over all pairs of points,

$$\Pr\left[\exists i,\; p,q \in P \;\mid\; \|\mathbf{v}^p_i - \mathbf{v}^q_i\|_2^2 - f_J(p_i,q_i)| \ge \varepsilon f_J(p_i,q_i)\right]$$

is at most $1/d$. And hence, if $\mathbf{v}^p$ is a concatenation of $\mathbf{v}^p_i$ over all $i \in [d]$, then with probability at least $1 - 1/d$ it holds for all $p,q \in P$ that $(1-\varepsilon)\text{JS}(p,q) \le \|\mathbf{v}^p - \mathbf{v}^q\| \le (1+\varepsilon)\text{JS}(p,q)$. The final length of the vectors is then $sd = 36n^2d^3\varepsilon^{-2}$ for approximately preserving distances between every pair of points with probability at least $1 - \frac{1}{d}$.

5 Dimensionality Reduction

The JL-Lemma has been instrumental for improving the speed and approximation ratios of learning algorithms. In this section, we give a proof of the JL-analogue for a general class of divergences that includes the information divergences studied here. Specifically, we show that a set of n points lying on a high-dimensional simplex can be embedded to a $k = O(\log n/\varepsilon^2)$-dimensional simplex, while approximately preserving the information distances between all pairs of points. This dimension reduction amounts to reducing the support of the distribution from d to k, while approximately maintaining the divergences.

Our proof uses ℓ_2^2 as an intermediate space. On a high level, we first embed the points into a high (but finite) dimensional ℓ_2^2 space, using the techniques we developed in Section 4.2. We then use the Euclidean JL-Lemma to reduce the dimensionality, and remap the points into the interior of a simplex. Finally, we show that far away from the simplex boundaries, this class of divergences has the same structure as ℓ_2^2, hence the embedding back into information spaces can be done with a simple translation and rescaling. Note that for divergences that have an embedding into finite-dimensional ℓ_2^2, the proof is constructive.

Definition 5.1 (f-divergence)**.** *Let p and q be two distributions on $[n]$. A convex function $f : [0,\infty) \to \mathbb{R}$ such that $f(1) = 0$ gives rise to an f-divergence $D_f : \Delta_d \to \mathbb{R}$ as:*

$$D_f(p,q) = \textstyle\sum_{i=1}^d p_i \cdot f\left(\frac{q_i}{p_i}\right),$$

where we define $0 \cdot f(0/0) = 0$, $a \cdot f(0/a) = a \cdot \lim_{u\to 0} f(u)$, and $0 \cdot f(a/0) = a \cdot \lim_{u\to\infty} f(u)/u$.

Definition 5.2 (Well-behaved divergence)**.** *A well-behaved f-divergence is a regular f-divergence such that $f(1) = 0$, $f'(1) = 0$, $f''(1) > 0$, and $f'''(1)$ exists.*

Algorithm 3: Dimension Reduction for D_f

Input: Set $P = \{p_1, \ldots, p_n\}$ of points on Δ_d, error parameter ε, constant $c_0(\varepsilon, f)$
Output: A set $\bar{P}$ of points on Δ_k where $k = O\left(\frac{\log n}{\varepsilon^2}\right)$

1. Embed P into ℓ_2^2 to obtain P_1 with error parameter $\varepsilon/4$.
2. Apply Euclidean JL–Lemma with error $\frac{\varepsilon}{4}$ to obtain P_2 in dimension $k = O\left(\frac{\log n}{\varepsilon^2}\right)$.
3. Remap P_2 to the plane $L = \{x \in \mathbb{R}^{k+1} \mid \sum_i x_i = 0\}$ to obtain P_3.
4. Scale P_3 to a ball of radius $c_0 \cdot \frac{\varepsilon}{k+1}$ and center at the centroid of Δ_{k+1} to obtain $\bar{P}$.

To analyze the above algorithm, we recall the JL–Lemma [8, 17]:

Lemma 5.3 (JL Lemma)**.** *For any set P of points in a (possibly infinite-dimensional) Hilbert space H, there exists a randomized map $f\colon H \to \mathbb{R}^k$, $k = O(\frac{\log n}{\varepsilon^2})$ such that whp, $\forall p,q \in P$*
$(1-\varepsilon)\|p-q\|_2^2 \le \|f(p) - f(q)\|_2^2 \le (1+\varepsilon)\|p-q\|_2^2$.

Corollary 5.4. *For any set of points P in H there exists a constant t and a randomized map $f\colon H \to \Delta_{k+1}$, $k = O(\frac{\log n}{\varepsilon^2})$ such that $\forall p,q \in P$:*

$(1-\varepsilon)\|p-q\|_2^2 \le t\|f(p)-f(q)\|_2^2 \le (1+\varepsilon)\|p-q\|_2^2$. *Furthermore for any small enough constant r, we may bound the domain of f to be a ball B of radius r centered at the simplex centroid,* $(1/k+1, \ldots, 1/k+1)$.

Proof. Consider first the map of Lemma 5.3 from $\mathbb{R}^d \to \mathbb{R}^k$. Now note that any set of points in R^k can be isometrically embedded into the hyperplane $L = \{x \in \mathbb{R}^{k+1} \mid \sum_i x_i = 0\}$. This follows by remapping the basis vectors of $\mathbb{R}^k$ to those of L. Finally since L is parallel to the simplex plane, the entire point set may be scaled by some factor t and then translated to fit in Δ_{k+1}, or indeed in any ball of radius r centered at the simplex centroid. □

We now show that any well-behaved f divergence is nearly Euclidean near the simplex centroid.

Lemma 5.5. *Consider any well-behaved f divergence D_f, and let B_r be a ball of radius r such that $B_r \subset \Delta_k$ and B_r is centered at the simplex centroid. Then for any fixed $0 < \varepsilon < 1$, there exists a choice of r and scaling factor t (both dependent on k) such that $\forall p, q \in B$:*

$$(1-\varepsilon)\|p-q\|_2^2 \le tD_f(p,q) \le (1+\varepsilon)\|p-q\|_2^2.$$

Proof. We consider arbitrary p, $q \in B_r$ and note that the assumptions imply each coordinate lies in the interval $I = [\frac{1}{k}-r, \frac{1}{k}+r]$. Let $rk = \varepsilon'$, then $I = [\frac{1-\varepsilon'}{k}, \frac{1+\varepsilon'}{k}]$. We now prove the lemma for $p, q \in I$, the main result follows by considering D_f and $\|\cdot\|^2$ coordinate by coordinate. By the definition of well-behaved f-divergences and Taylor's theorem, there exists a neighborhood N of 1, and function ϕ with $\lim_{x\to 1}\phi(1) = 0$ such that for all $x \in N$:

$$\begin{aligned} f(x) =& f(1) + (x-1)f'(1) + \frac{(x-1)^2}{2}f''(1) + \\ & (x-1)^3\phi(x) \\ =& \frac{(x-1)^2}{2}f''(1) + (x-1)^3\phi(x). \end{aligned} \tag{5.1}$$

Therefore:

$$\begin{aligned} \frac{D_f(p,q)}{\|p-q\|_2^2} &= \frac{p\cdot f\left(\frac{q}{p}\right)}{(p-q)^2} \\ &= \frac{p\left(\left(\frac{q-p}{p}\right)^2\frac{f''(1)}{2} + \left(\frac{q-p}{p}\right)^3\phi\left(\frac{q}{p}\right)\right)}{(q-p)^2} \\ &= \frac{f''(1)}{2p} + \frac{q-p}{p^2}\phi\left(\frac{q}{p}\right). \end{aligned}$$

Recall again that $p \in [\frac{1-\varepsilon'}{k}, \frac{1+\varepsilon'}{k}]$ so the first term converges to the constant $2kf''(1)$ as r grows smaller (and hence ε' decreases). Note also that the second term goes to 0 with r, i.e., given a suitably small choice of r we can make the term smaller than any desired constant. Hence, for every dimension k and $0 < \varepsilon < 1$, there exists a radius of convergence r such that for all $p, q \in B_r$:

$$(1-\varepsilon)\|p-q\|_2^2 \le \frac{1}{2kf''(1)}D_f(p,q) \le (1+\varepsilon)\|p-q\|_2^2. \ \square$$

We note that the required value of r can be computed easily for the Hellinger and χ^2 divergence, and that r behaves as $\frac{1}{k}\cdot c$ where $c = c(f,\epsilon)$ is a sufficiently small constant depending only on ε and the function f and not on k or n .

To conclude the proof observe that the overall distortion is bounded by the combination of errors due to the initial embedding into P_1, the application of JL-Lemma, and the final reinterpretation of the points in Δ_{k+1}. The overall error is thus bounded by, $(1+\varepsilon/4)^3 \le 1+\varepsilon$.

Theorem 5.6. *Consider a set $P \in \Delta_d$ of n points under a well-behaved f-divergence D_f. Then there exists a $(1+\varepsilon)$ distortion embedding of P into Δ_k under D_f for some choice of k bounded as $O\left(\frac{\log n}{\varepsilon^2}\right)$.*

Furthermore this embedding can be explicitly computed even for a well-behaved f-divergence with an infinite-dimensional kernel, if the kernel can be approximated in finite dimensions within a multiplicative error as we show for JS and χ^2.

6 Experiments

We analyze the empirical performance of our algorithms and demonstrate the effect that each parameter has on the quality of the final solution. We show there is minute loss incurred both in sampling from the kernel (embedding the points into ℓ_2^2), and in remapping the points to lie on the d-dimensional simplex.

Recall that the dimension reduction procedure in Algorithm 3 has three parameters: s, the number of samples used to embed the points into ℓ_2^2, k, the target dimension of the Euclidean JL-Lemma, and c_0, the scaling parameter used to embed the points in the final simplex.

Synthetic data. To study the quality of the embedding with respect to these three parameters, we generated distributions on $d = 100, 1000$, and $10,000$ dimensions. We used the Cauchy distribution and the Log-Factorial distribution[2] as the seeds. To generate a point in the dataset, we randomly permuted the coordinates in one of these distributions.

[2]Defined on $[d]$, the pdf is given by $\Pr[i] \propto \log(1+i)$.

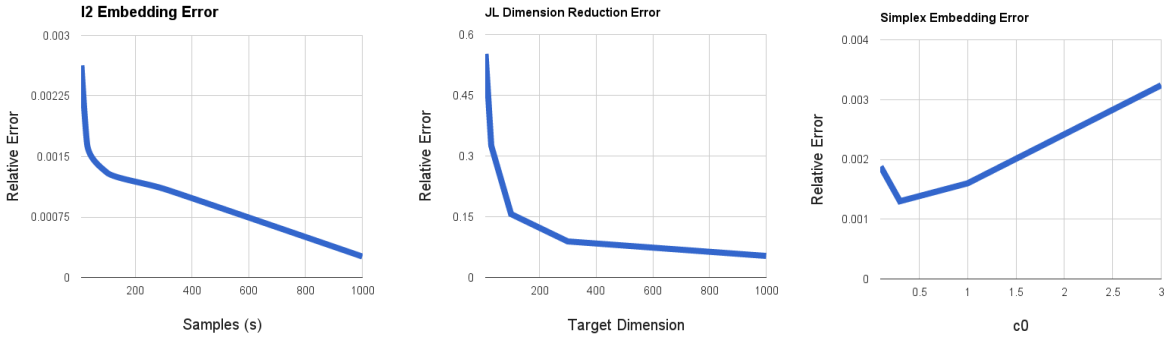

Figure 1: The breakdown of error due to the number of samples, the JL-Lemma, and the simplex embedding.

	Spoken	Fiction	Popular	Newspaper	Academic
Spoken	-	1.5%	3.3%	1.87%	0.3%
Fiction	-	-	1.4%	3.75%	2.95%
Popular	-	-	-	5.2%	5.15%
Newspaper	-	-	-	-	1.25%

Table 1: Relative error of the JS divergence after embedding into 300 dimensions.

To explore the dependence on the parameters, we set the defaults to $s = 100$, $k = 300$, and $c_0 = 0.1$. Note that the value of s is *far lower* than that implied by the analysis (the value of c_0 is far higher). In the three panels in Figure 1 we vary one of the parameters while keeping the others fixed; all of these are averaged over 100 pairwise computations. We track the error introduced by embedding into ℓ_2^2, reducing the dimension to d, and re-embedding back into the simplex Δ_d. As expected, we observe that the overall error decreases with increasing the number of samples, and with lowering c_0. These contributions are on the order of 0.075% to 0.3%, and are far outweighed by the error introduced by the JL-Lemma step itself, which is on the order of 9–10% and forms the core of the reduction.

Baselines. The *only* known method for dimension reduction in the simplex is due to [18], which essentially eschews the kernel embedding into ℓ_2^2 and proceeds to apply the JL-lemma directly on the distribution points. While it provably works only in a limited domain, we nevertheless investigate its performance. While the error of our method on the synthetic dataset ranges from 5–30% depending on the value of the target dimension, the error produced by the baseline method ranges from 95–430%.

Real data. To further demonstrate the efficacy of our approach we create a dataset where each point is the word distributions found in a specific book genre, as gathered from the free sample on `www.wordfrequency.info`. We then compute the pairwise divergences between the distributions on the full dataset and on the embedded dataset. We use Algorithm 3 to reduce the dimensionality twenty-fold from 6000 to just 300 dimensions. Using the same fixed set of parameters, we show the average (over 10 runs) error between the different genres in Table 1. Even as we reduce the dimensionality of the space 20-fold while maintaining the simplicial structure, the average distortion remains very small, only between 0.3% and 5.15%.

7 Conclusions

We present a simple, practical, and theoretically sound dimension reduction algorithm for information spaces. Our algorithm reduces the dimensionality of the space while maintaining the simplicial structure, and can be used in a black-box manner to speed up algorithms that operate in information divergence spaces.

The embedding and sketching algorithmic results we show here complement the known impossibility results for sketching information distances in the strict turnstile model, thus providing a more complete picture of how these distances can be estimated in a stream.

Notice that the mappings that we present are contractive, i.e., forced to be near the center of the simplex. We conjecture that for non-contractive mappings, the Hellinger distance will *not* admit a $(1 + \epsilon)$ dimension reduction.

8 Acknowledgments

This research was partially supported by the NSF under grants CCF-0953066, IIS-1251049, CCF-0953754, IIS-1251110, CCF-1320719, and a Google Research Award. Andrew McGregor and Suresh Venkatasubramanian was partially supported by Google during the period this research was conducted.

References

[1] A. Abdullah, R. Kumar, A. McGregor, S. Vassilvitskii, and S. Venkatasubramanian. Sketching, embedding, and dimensionality reduction for information spaces. *CoRR*, abs/1503.05225, 2015. URL `http://arxiv.org/abs/1503.05225`.

[2] C. C. Aggarwal. *Data Streams: Models and Algorithms*, volume 31. Springer Science & Business Media, 2007.

[3] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *JCSS*, 58(1):137–147, 1999.

[4] A. Andoni, M. Charikar, O. Neiman, and H. Nguyen. Near linear lower bound for dimension reduction in ℓ_1. In *FOCS*, pages 315–323, 2011.

[5] A. Andoni, R. Krauthgamer, and I. Razenshteyn. Sketching and embedding are equivalent for norms. In *STOC*, pages 479–488, 2015.

[6] H. Avron, H. Nguyen, and D. Woodruff. Subspace embeddings for the polynomial kernel. In *NIPS*, pages 2258–2266, 2014.

[7] R. Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, pages 445–463, 1977.

[8] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE TOIT*, 54(2):781–790, 2008.

[9] B. Brinkman and M. Charikar. On the impossibility of dimension reduction in l_1. *JACM*, 52(5): 766–788, 2005.

[10] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[11] I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information theoretic feature clustering algorithm for text classification. *JMLR*, 3:1265–1287, 2003.

[12] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *SIGIR*, pages 459–460, 2003.

[13] B. Fuglede and F. Topsøe. Jensen–Shannon divergence and Hilbert space embedding. In *ISIT*, 2004.

[14] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *SODA*, pages 733–742, 2006.

[15] S. Guha, P. Indyk, and A. McGregor. Sketching information divergences. *Machine Learning*, 72(1-2):5–19, 2008.

[16] X. Huang, S. Z. Li, and Y. Wang. Jensen–Shannon boosting learning for object recognition. In *CVPR*, pages 144–149, 2005.

[17] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[18] R. J. Kyng, J. M. Phillips, and S. Venkatasubramanian. Johnson–Lindenstrauss dimensionality reduction on the simplex. In *20th Fall Workshop on Computational Geometry*, 2010.

[19] J. R. Lee and A. Naor. Embedding the diamond graph in L_p and dimension reduction in L_1. *GAFA*, 14(4):745–747, 2004.

[20] Y. Li, H. L. Nguyen, and D. P. Woodruff. Turnstile streaming algorithms might as well be linear sketches. In *STOC*, pages 174–183, 2014.

[21] M. Mahmoudi and G. Sapiro. Three-dimensional point cloud recognition via distributions of geometric distances. *Graphical Models*, 71(1):22–31, 2009.

[22] S. Muthukrishnan. *Data Streams: Algorithms and Applications*. Now Publishers Inc, 2005.

[23] A. M. Peter and A. Rangarajan. Maximum likelihood wavelet density estimation with applications to image and shape matching. *IEEE TIP*, 17(4): 458–468, 2008.

[24] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007.

[25] O. Regev. Entropy-based bounds on dimension reduction in L_1. *Israel J. Math.*, pages 1–8, 2012.

[26] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE TOIT*, 46(4):1602–1609, 2000.

[27] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE PAMI*, 34 (3):480–492, 2012.

[28] A. Wright. Data streaming 2.0. *Communications of the ACM*, 53(4):13–14, 2010.