Bilateral Distribution Compression: Reducing Both Data Size and Dimensionality

Dominic Broadbent *Nick Whiteley* *Robert Allison* *Tom Lovett*

Introduction

- Training AI models often now requires enormous datasets, computation, and energy, with serious financial and environmental costs. A key challenge is thus how to reduce data without losing the essential information it carries.
- Existing *distribution compression* methods focus exclusively on reducing the number of observations, while preserving the essential statistical properties of the original data. However, this overlooks the fact that modern datasets are often large both in sample size *and* in dimensionality
- To address this gap, we propose *Bilateral Distribution Compression* (**BDC**), highlighting that compression acts simultaneously on both the number of samples *and* their dimensionality, i.e. *bilaterally*.

Maximum Mean Discrepancy

- Given a dataset $\mathcal{D} := \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ with $n, d \gg 1$ sampled i.i.d. from a distribution $\mathbb{P}_X$ we construct a compresed set $\mathcal{C} := \{\boldsymbol{z}_j\}_{j=1}^m \subset \mathbb{R}^p$ with $m \ll n$ and $p \ll d$ that preserves the distribution of the original data.
- To measure distributional fidelity, we use the *Maximum Mean Discrepancy* (**MMD**). Given an additional distribution $\mathbb{P}_Y$ the **MMD** is given as

$$\mathrm{MMD}^2(\mathbb{P}_X, \mathbb{P}_Y) = \mathbb{E}[k(x, x')] - 2\,\mathbb{E}[k(x, y)] + \mathbb{E}[k(y, y')]$$

  where $x, x' \sim \mathbb{P}_X$, $y, y' \sim \mathbb{P}_Y$ and $k(\cdot, \cdot)$ is a user-specified *kernel function* which determines which aspects of the two distributions are compared. It can be shown that for certain kernels, the **MMD** is zero if and only if $\mathbb{P}_X = \mathbb{P}_Y$.
- As a notion of discrepancy between the empirical distribution $\hat{\mathbb{P}}_X$ in ambient space $\mathbb{R}^d$, and the empirical distribution of the compressed set $\hat{\mathbb{P}}_Z$ in latent space $\mathbb{R}^p$ we propose the *Decoded MMD* (**DMMD**):

$$\mathrm{DMMD}^2(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_{\phi(Z)}) = \sum_{i,j=1}^{n} k(\boldsymbol{x}_i, \boldsymbol{x}_j) - 2\sum_{i,j=1}^{n,m} k(\boldsymbol{x}_i, \phi(\boldsymbol{z}_j)) + \sum_{i,j=1}^{m} k(\phi(\boldsymbol{z}_i), \phi(\boldsymbol{z}_j))$$

where $\phi : \mathbb{R}^p \to \mathbb{R}^d$ is a *decoder* that maps points back into ambient space.

Bilateral Distribution Compression

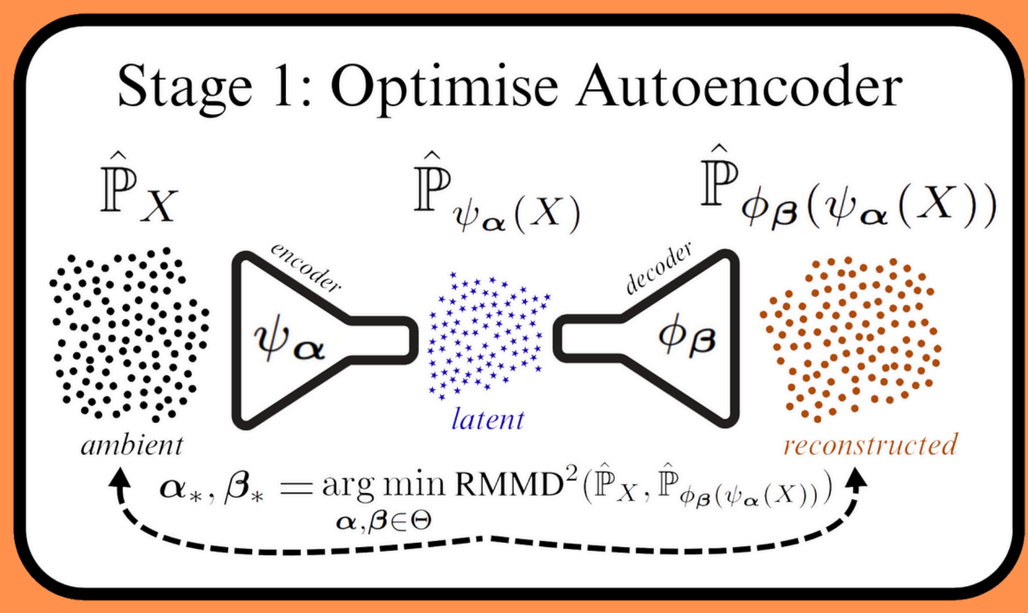
- The **DMMD** is a natural target, however joint optimisation over both the decoder and the compressed set is challenging. Their roles are tightly coupled within the **DMMD**, leading to a highly non-convex and entangled optimisation landscape.
- The key idea of **BDC** is thus to break a hard problem into two easier stages:
  - **Stage 1**: Train an encoder $\psi : \mathbb{R}^d \to \mathbb{R}^p$ and decoder $\phi : \mathbb{R}^p \to \mathbb{R}^d$ by minimising the **MMD** between the original data $\mathcal{D}$ and its reconstruction $\phi(\psi(\mathcal{D}))$, denoted as ***Reconstruction*** **MMD**.
  - **Stage 2**: With ϕ, ψ fixed, optimise a compressed set by minimising the **MMD** between $\mathcal{D}$ projected into latent space, i.e. $\psi(\mathcal{D})$, and the compresseed set $\mathcal{C} \subset \mathbb{R}^p$, denoted as ***Encoded*** **MMD**.
- The **RMMD** checks if an autoencoder's reconstructions preserve the distribution of the data and the **EMMD** checks if the compressed latent set represents the encoded data well in latent space.

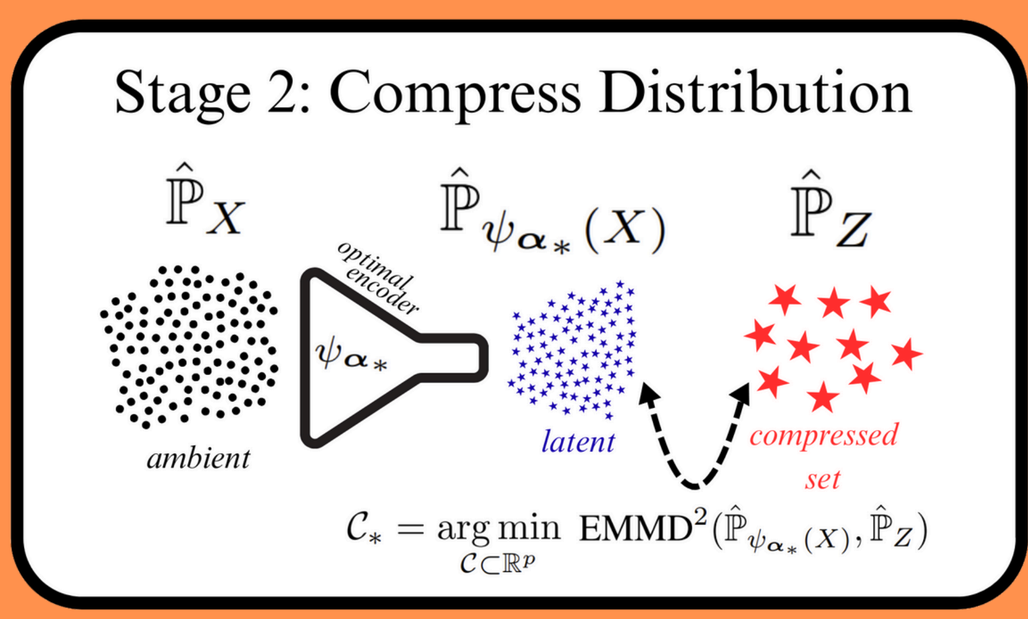
Theorem 1

Under mild conditions on the choice of kernel functions, suppose that $\mathbb{P}_X$ and $\mathbb{P}_Z$ are such that $\mathrm{RMMD}\left(\mathbb{P}_X, \mathbb{P}_{\phi(\psi(X))}\right) = 0$ and $\mathrm{EMMD}\left(\mathbb{P}_{\psi(X)}, \mathbb{P}_Z\right) = 0$ then we have that $\mathrm{DMMD}\left(\mathbb{P}_X, \mathbb{P}_{\psi(Z)}\right) = 0$.

- Theorem 1 states that if both **RMMD** and **EMMD** vanish, then the **DMMD** is also exactly zero, motivating the two-stage process. Thus, the distribution is preserved despite the reduction in both sample size and dimensionality.

Two-Stage Optimisation





Downstream Tasks

- **BDC** can be used to construct both labelled and unlabelled compressed sets, opening up acceleration of a variety of important supervised and unsupervised downstream tasks such as (but not limited to) regression, classification and clustering. We now present a clustering problem.
- We construct a synthetic dataset of $n = 100{,}000$ two-dimensional points arranged in 10 clusters of varying shapes, with an additional $2{,}000$ points of uniform noise. The data are non-linearly projected to $d = 500$ dimensions.
- Using **BDC** we then compress back to $p = 2$ dimensions with $m = 300$ points, achieving 99.999% compression, taking just over a minute.
- Using the clustering method HDBSCAN, we are then able to cluster held out test points with accuracy comparable to the full dataset, and standard (non-bilateral) distribution compression methods, at a fraction of the cost.

Original Intrinisic Data **BDC**

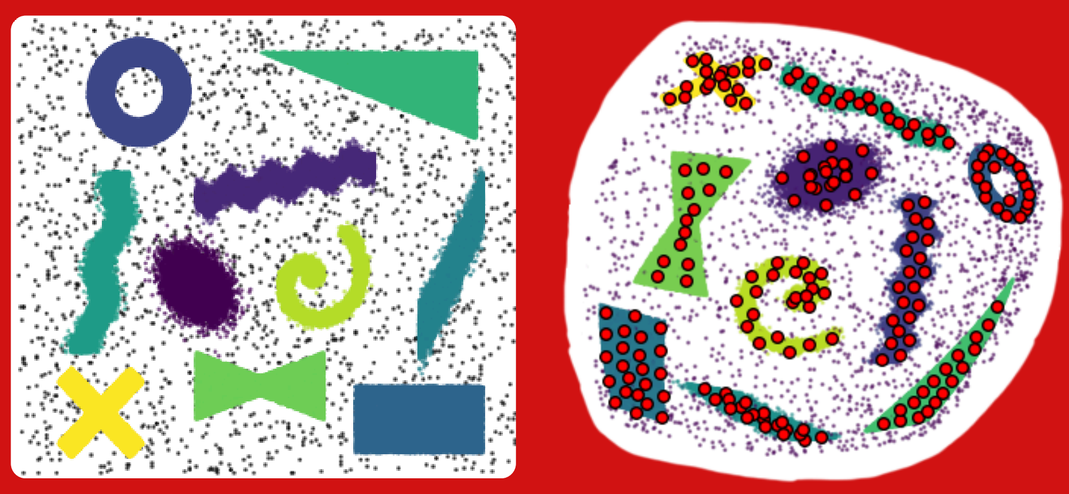


Figure 1: Original intrinsic data before projection, next to a compressed set constructed by **BDC** (red). The encoder recovers the intrinsic space and the compressed set clearly delineates the clusters, ignoring noise.

4th Compass Conference, October 2025

University of BRISTOL
School of Mathematics
dominic.broadbent@bristol.ac.uk