Conditional Distribution Compression via the Kernel Conditional Mean Embedding

Dominic Broadbent [1] *Nick Whiteley* [1] *Robert Allison* [1] *Tom Lovett* [2]

(1) *University of Bristol, School of Mathematics;* **(2)** *University of Oxford, Mathematical Institute*

Motivation

- Distribution compression seeks to replace large datasets with smaller representative sets that preserve their key statistical properties, reducing the financial, environmental, and time costs of storage and computation.

- Existing methods have been developed for unlabelled data, targeting the distribution $\mathbb{P}_X$ [1, 2, 3]. However, many real-world datasets are labelled, where preserving relationships between inputs and outputs is essential.

- Depending on the downstream task, one may wish to preserve the joint distribution $\mathbb{P}_{X,Y}$, which captures dependencies between features and labels, or the conditional distribution $\mathbb{P}_{Y|X}$ which governs predictive behaviour.

Motivation

- Distribution compression seeks to replace large datasets with smaller representative sets that preserve their key statistical properties, reducing the financial, environmental, and time costs of storage and computation.

- Existing methods have been developed for unlabelled data, targeting the distribution $\mathbb{P}_X$ [1, 2, 3]. However, many real-world datasets are labelled, where preserving relationships between inputs and outputs is essential.

- Depending on the downstream task, one may wish to preserve the joint distribution $\mathbb{P}_{X,Y}$, which captures dependencies between features and labels, or the conditional distribution $\mathbb{P}_{Y|X}$ which governs predictive behaviour.

Motivation

University of BRISTOL NEURAL INFORMATION PROCESSING SYSTEMS UNIVERSITY OF OXFORD MATHEMATICAL INSTITUTE

- Distribution compression seeks to replace large datasets with smaller representative sets that preserve their key statistical properties, reducing the financial, environmental, and time costs of storage and computation.
- Existing methods have been developed for unlabelled data, targeting the distribution $\mathbb{P}_X$ [1, 2, 3]. However, many real-world datasets are labelled, where preserving relationships between inputs and outputs is essential.
- Depending on the downstream task, one may wish to preserve the joint distribution $\mathbb{P}_{X,Y}$, which captures dependencies between features and labels, or the conditional distribution $\mathbb{P}_{Y|X}$ which governs predictive behaviour.

Distribution Compression

University of Bristol · Neural Information Processing Systems · Mathematical Institute, University of Oxford

- Distribution compression algorithms optimise the compressed set $\mathcal{C} = \{\boldsymbol{z}_i\}_{i=1}^m$ to minimise the MMD to the empirical distribution $\hat{\mathbb{P}}_X$ of the target dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$\begin{aligned}\mathrm{MMD}^2(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Z) &:= \|\hat{\mu}_X - \hat{\mu}_Z\|^2_{\mathcal{H}_k} \\ &= \sum_{i,j=1}^{n} k(\boldsymbol{x}_i, \boldsymbol{x}_j) - 2\sum_{i,j=1}^{n,m} k(\boldsymbol{x}_i, \boldsymbol{z}_j) + \sum_{i,j=1}^{m} k(\boldsymbol{z}_i, \boldsymbol{z}_j),\end{aligned}$$

where $m \ll n$, and we denote μ_X as the *kernel mean embedding* of the distribution $\mathbb{P}_X$. The KME μ_X lies in the *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}_k$ induced by the positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which is defined on the feature space $\mathcal{X}$.

Distribution Compression

- Given an additional kernel $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ defined on the response space $\mathcal{Y}$ we induce the RKHS $\mathcal{H}_k \otimes \mathcal{H}_l$. We can then extend existing distribution compression algorithms to optimise a compressed set $\mathcal{C} = \{(\boldsymbol{z}_i, \boldsymbol{w}_i)\}_{i=1}^m$ which minimises the Joint MMD [5] to the empirical distribution of the target dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$:

$$\mathrm{JMMD}^2(\hat{\mathbb{P}}_{X,Y}, \hat{\mathbb{P}}_{Z,W}) := \|\hat{\mu}_{X,Y} - \hat{\mu}_{Z,W}\|^2_{\mathcal{H}_{k\otimes l}}$$

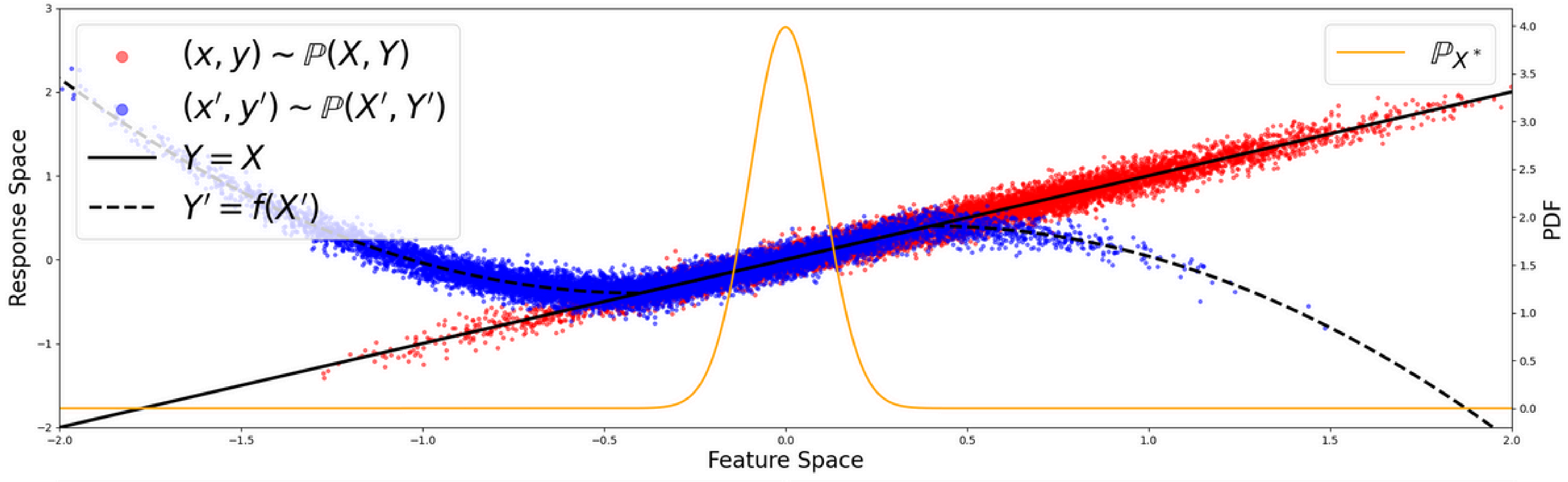
$$= \sum_{i,j=1}^{n} k(\boldsymbol{x}_i, \boldsymbol{x}_j) l(\boldsymbol{y}_i, \boldsymbol{y}_j) - 2 \sum_{i,j=1}^{n,m} k(\boldsymbol{x}_i, \boldsymbol{z}_j) l(\boldsymbol{y}_i, \boldsymbol{w}_j) + \sum_{i,j=1}^{m} k(\boldsymbol{z}_i, \boldsymbol{z}_j) l(\boldsymbol{w}_i, \boldsymbol{w}_j).$$

Distribution Compression

- In order to extend distribution compression to the conditional distribution, we first require a notion of conditional disrepancy, for this we introduce the AMCMD:

$$\mathrm{AMCMD}\left(\mathbb{P}_{X^*}, \mathbb{P}_{Y|X}, \mathbb{P}_{Y'|X'}\right) := \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{X^*}}\left[\left\|\mu_{Y|X=\boldsymbol{x}} - \mu_{Y'|X'=\boldsymbol{x}}\right\|^2_{\mathcal{H}_l}\right]}$$

where $\mathbb{P}_{X^*}$ is a weighting distribution, and $\mu_{Y|X} : \mathcal{X} \to \mathcal{H}_l$ is the *kernel conditional mean embedding* (KCME). The KCME is a *vector-valued* function, which takes as inputs conditioning values $\boldsymbol{x} \in \mathcal{X}$, and outputs KMEs $\mu_{Y|X=\boldsymbol{x}}$ lying in $\mathcal{H}_l$.



Distribution Compression

University of BRISTOL · NEURAL INFORMATION PROCESSING SYSTEMS · UNIVERSITY OF OXFORD MATHEMATICAL INSTITUTE

Theorem - *The AMCMD is a proper metric*

Suppose the response kernel $l(\cdot,\cdot)$ is characteristic, that $\mathbb{P}_X$, $\mathbb{P}_{X'}$, and $\mathbb{P}_{X^*}$ are absolutely continuous with respect to eachother, and that $\mathbb{P}(\cdot \mid X)$ and $\mathbb{P}(\cdot \mid X')$ admit regular versions. Then, $\mathrm{AMCMD}\left(\mathbb{P}_{X^*}, \mathbb{P}_{Y|X}, \mathbb{P}_{Y'|X'}\right) = 0$ if and only if, for almost all $\boldsymbol{x} \in \mathcal{X}$ wrt $\mathbb{P}_{X^*}$, $\mathbb{P}_{Y|X=\boldsymbol{x}}(A) = \mathbb{P}_{Y'|X'}(A)$ for all $A \in \mathscr{Y}$.

Moreover, assuming the Radon-Nikodym derivatives $\frac{\mathrm{d}\mathbb{P}_{X^*}}{\mathrm{d}\mathbb{P}_X}$, $\frac{\mathrm{d}\mathbb{P}_{X^*}}{\mathrm{d}\mathbb{P}_X}$, and $\frac{\mathrm{d}\mathbb{P}_{X^*}}{\mathrm{d}\mathbb{P}''_X}$ are bounded, then the triangle inequality is satisfied, i.e.

$$\mathrm{AMCMD}\left(\mathbb{P}_{Y|X}, \mathbb{P}_{Y''|X''}\right) \leq \mathrm{AMCMD}\left(\mathbb{P}_{Y|X}, \mathbb{P}_{Y'|X'}\right) + \mathrm{AMCMD}\left(\mathbb{P}_{Y'|X'}, \mathbb{P}_{Y''|X''}\right).$$

Distribution Compression

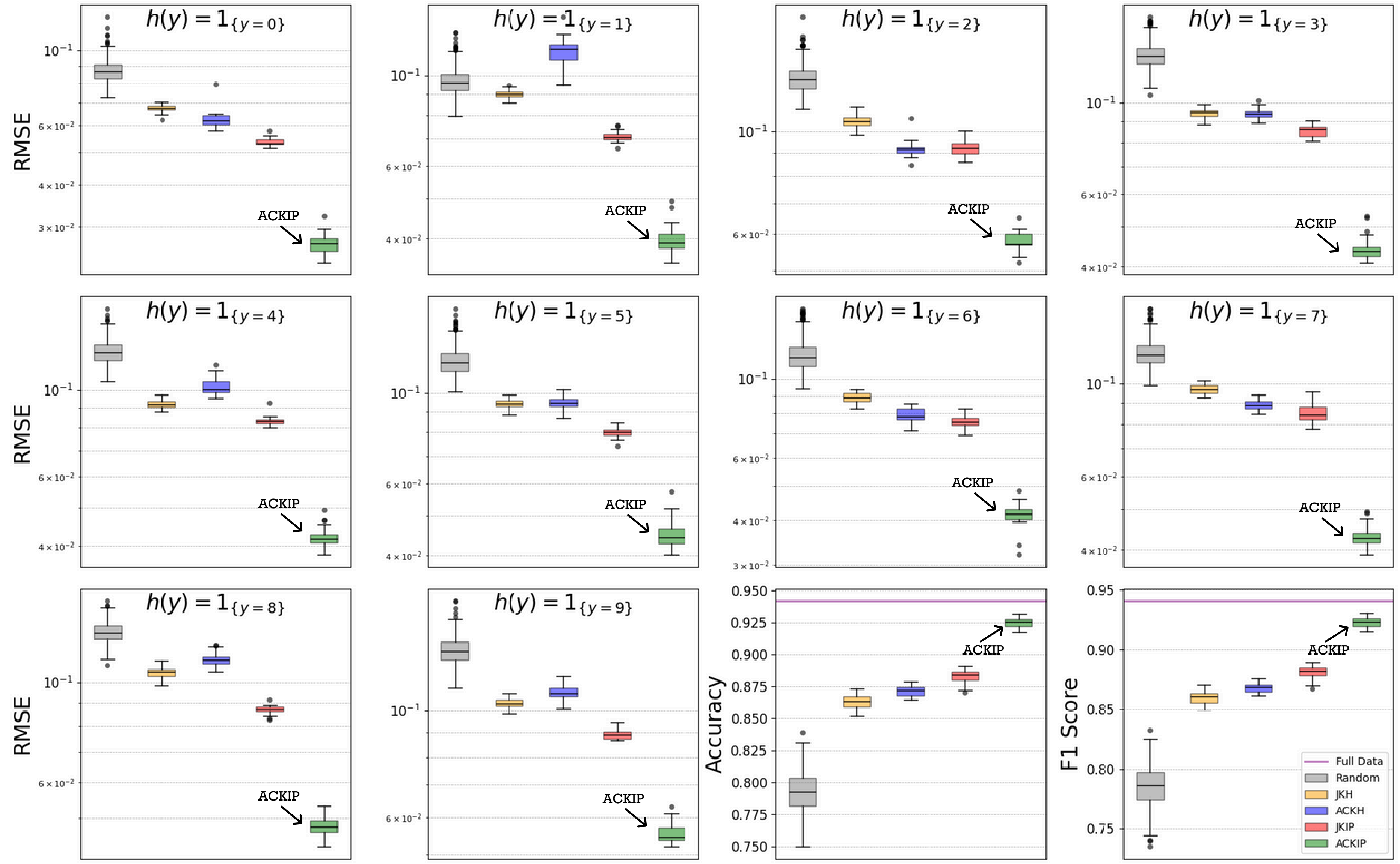
- We can now optimise a compressed set $\mathcal{C} = \{(\boldsymbol{z}_i, \boldsymbol{w}_i)\}_{i=1}^{m}$ which minimises the AMCMD to the empirical conditional distribution of the target dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$:

$$\mathrm{AMCMD}^2\left(\hat{\mathbb{P}}_{X^*}, \hat{\mathbb{P}}_{Y|X}, \hat{\mathbb{P}}_{Z|W}\right) = \frac{1}{q}\sum_{i=1}^{q}\left\|\hat{\mu}_{Y|X=\boldsymbol{x}_i^*} - \hat{\mu}_{Z|W=\boldsymbol{x}_i^*}\right\|_{\mathcal{H}_l}^2.$$

- We can obtain a closed-form representation of this, however it has $\mathcal{O}(n^3)$ cost. For distribution compression, it is natural to choose $\mathbb{P}_{X^*} = \mathbb{P}_X$, then by applying the tower property, we can reduce to $\mathcal{O}(n)$ cost, enabling linear-time conditional distribution compression.

Distribution Compression

- The KCME has many important applications. In particular it may be used as a regressor and classifier. In our work, we investigate how compression effects these downstream tasks. Below, we show results on MNIST after 98% compression:



References

[1] - Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. UAI 2010

[2] - Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow, NeurIPS 2019

[3] - Raaz Dwivedi and Lester Mackey. Kernel thinning. COLT 2021

[4] - Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test, JMLR 2012

[5] - Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks, ICML 2017