Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Alec Barber, Anthony Quinn

# Bayesian transfer learning between autoregressive inference tasks

No. 2389 2020

**GACR 18-15970S**

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: (+420)266052422, Fax: (+420)286890378, Url: http://www.utia.cas.cz, E-mail: utia@utia.cas.cz

Any opinions and conclusions expressed in this report are those of the authors and do not necessarily represent the views of the Institute.

# Contents

***Abstract***

Bayesian transfer learning typically relies on a complete stochastic dependence specification between source and target learners which allows the opportunity for Bayesian conditioning. We advocate that any requirement for the design or assumption of a full model between target and sources is a restrictive form of transfer learning.

Motivated to identify a knowledge transfer strategy which does not require any model assumptions/design, this document identifies a transfer learning scheme via adoption of the Fully Probabilistic Design (FPD) distributional design framework. More specifically, this document approaches transfer learning in a distributed autoregressive setting where multiple source learners provide a target learner knowledge in the form of a distribution at each time step. We investigate the feasibility of using parameter knowledge transfer over the more typical data prediction transfer but finds no convincing method to do so. Motivated by a desire to achieve robustness to negative transfer, we propose a binary "*obstinate*" acceptance criterion describing when the target inference task should accept or reject source knowledge.

Experimentation was conducted using Monte Carlo simulations with artificially generated parameters. Results showed good performance with the transfer learning algorithm providing an increase in predictive power over the isolated learner scenario. Interestingly, results are independent of the source inference task's generative model variance when considering the mean absolute error metric. When considering a more atypical and novel metric, the Kullback-Leibler divergence between the targets one step ahead data predictor and the true generative function, results show that increasing source variance reduces the prediction capabilities of target inference machine with negative transfer becoming possible for high variance values. Experiments on food stuffs price economic data was also conducted which showed modest improvements of food price prediction power when adopting the FPD transfer learning algorithm (with and without adoption of obstinate acceptance criterion).

***Keywords:*** autoregression, transfer learning, Fully Probabilistic Design, FPD, food-commodities price prediction

# 1 Introduction

Transfer learning (multi-task learning)[1] [2] is the process of adopting knowledge from a set of source learning tasks to increase the learning rate of a given target task [3]. This document is specifically concerned with transfer learning in a Bayesian context and proposes a scheme for knowledge transfer between multiple source autoregressive (AR) inference machines via the sharing of one step-ahead data predictors (probability distribution) at each time-step. The objective of Bayesian Transfer Learning (BTL) is to identify an optimal target distribution $\mathsf{M}^{\circ}$ conditioned on provided source knowledge $\mathsf{F}_{\mathsf{S}}$ alongside $\mathsf{n}$ locally observed data $\mathbf{z}_{\mathsf{n}}$.

BTL in literature typically is either approached by the design of a prior for the target via knowledge provided by source learners [4][5], or by the assumption/design of a joint model between the target and source learners [6], allowing Bayesian conditioning to take place. Prior design for knowledge transfer can prove effective, but is restrictive due to its inability to incorporate synchronous on-line learning. The design or assumption of a complete model is argued to be inconsistent with a general solution to BTL and is rather is a restrictive special case.

Motivated to avoid specifying a complete model between any target and source learner, Fully Probabilistic Design (FPD) [7][8] is adopted as an axiomatically justified [9] optimal distribution design framework, rooted in the minimum cross-entropy principle and is employed to elicit optimal distributions conditioned on source knowledge. FPD has previously been applied to transfer learning applications such as Gaussian process regression [10], Kalman filtering [11] and Student-t filtering [12]. This document aims to extend FPD to an autoregressive setting.

The specific setting for this research is as shown in figure 1 where multiple source and one target learner conduct standard Bayesian autoregressive inference on local data synchronously. At each time step, each of the source learners makes available a distribution $\mathsf{F}_{\mathsf{S}}$. This distribution is taken as the local one step ahead data predictor, although this was a free decision and could theoretically be any other probabilistic object.

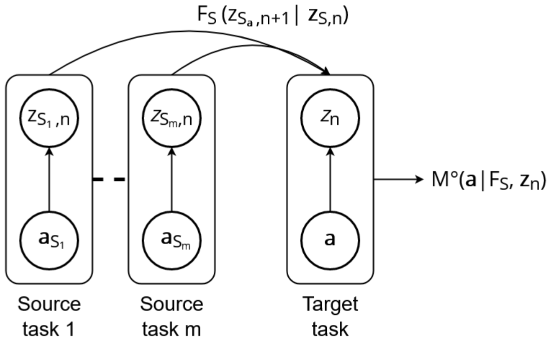

Figure 1: Source and Target AR inference tasks process local data $\mathbf{z}_{\boldsymbol{S_i,n}}$ and $\mathbf{z_n}$ to make inferences of respective local parameters $\mathbf{a}_{\mathsf{S_i}}$ and $\mathbf{a}$ respectively. At each time-step the source makes available its one step ahead data predictor $\mathsf{F_S}$ to the target. The target improves its performance by conditioning on $\mathsf{F_S}$ and eliciting the FPD-optimal knowledge conditional distribution $\mathsf{M}^{\circ}$.

## 2 Autoregressive parameter inference overview

AR models are a flexible and frequently adopted model with uses in the fields of speech synthesis [13][14] and finance [15][16] amongst others. AR modelling allows for the representation of dynamic relations between successive data points.

When conducting Bayesian AR inference, the objective is typically to infer model parameters $\boldsymbol{a}$ and variance $\mathsf{r}$ [17].

A univariate time-invariant AR model of order $\mathsf{p}$ is described via the Wold representation in (1). $\mathbf{I}$ is the identity matrix, $\mathbf{e_n}$ describes a white Gaussian noise random variable, $z_{\mathsf{n}} \in \mathbb{R}$ is the scalar observation datum at time $\mathsf{n}$ and $\boldsymbol{\psi}_{\mathsf{n}} \in \mathbb{R}^{\mathsf{p}}$ is the regression vector (vector of previous observations). This paper adopts $^{\mathsf{T}}$ to denote the transposition operator.

$$
\begin{gathered}
z_{\mathsf{n}} = \boldsymbol{\psi}_{\mathsf{n}}^{\mathsf{T}} \mathbf{a} + \mathbf{e}_{\mathsf{n}}, \\
\mathbf{e}_{\mathsf{n}} \sim \mathcal{N}(0, \mathsf{r}\mathbf{I}), \\
\psi_{\mathsf{n}} = [z_{\mathsf{n-1}} \dots z_{\mathsf{n-p}}]^{\mathsf{T}}
\end{gathered} \tag{1}
$$

$$
\mathsf{F}(z_{\mathsf{n}} | \mathbf{a}, \mathsf{r}, \mathbf{z}_{\mathsf{n-1}}) = \mathcal{N}(\boldsymbol{\psi}_{\mathsf{n}}^{\mathsf{T}} \mathbf{a}, \mathsf{r}) \tag{2}
$$

The normal observation model (2) for the process is a member of the exponential family and guaranteed to have a conjugate prior distribution and sufficient statistics. Accordingly, we adopt the *Normal-inverse-Gamma* ($\mathcal{N}$iG) conjugate distribution [?] to model our beliefs in the AR model parameters.

$$
\begin{gathered}
\mathsf{F}(\mathbf{a}, \mathsf{r} | \mathbf{V}, \nu) = \mathcal{N}\mathrm{iG}_{\mathbf{a},\mathsf{r}}(\mathbf{V}, \nu), \\
\mathcal{N}\mathrm{iG}_{\mathbf{a},\mathsf{r}} = \mathrm{K}^{-1} \mathsf{r}^{-\frac{\nu}{2}} \exp\left[ -\frac{1}{2\mathsf{r}} \begin{bmatrix} -1 & \mathbf{a}^T \end{bmatrix} \mathbf{V} \begin{bmatrix} -1 \\ \mathbf{a} \end{bmatrix} \right], \\
\mathsf{K} = \Gamma\left(\frac{\nu}{2}\right) \lambda^{-\frac{\nu}{2}} \left|\mathbf{V}_{\mathsf{aa}}\right|^{-\frac{1}{2}} (2\pi)^{\frac{1}{2}}, \\
\lambda = \mathsf{v}_{11} - \mathbf{v}_{\mathsf{a1}}^{\mathsf{T}} \mathbf{V}_{\mathsf{aa}}^{-1} \mathbf{v}_{\mathsf{a1}}
\end{gathered} \tag{3}
$$

$$
\mathbf{V} = \begin{bmatrix} v_{11} & \mathbf{v}_{a1}^T \\ \mathbf{v}_{a1} & \mathbf{V}_{aa} \end{bmatrix} \tag{4}
$$

Equation (3) describes the sufficient statistic parameterisation of $\mathcal{N}$iG where $\mathbf{V}$ is the Extended Information Matrix (EIM), $\nu$ is the degrees of freedom parameter and $\Gamma(\cdot)$ denotes the Gamma function. The EIM $\mathbf{V}$ must be a positive and semi-definite (4) shows the partitioning of the EIM into its sub-components, isolating the scalar $v_{11}$, the vector $\mathbf{v}_{a1} \in \mathbb{R}^{\mathsf{p}}$ and the matrix $\mathbf{V}_{aa} \in \mathbb{R}^{\mathsf{p} \times \mathsf{p}}$.

The update procedures for both on-line and batch inference at each time step is as equation (5) where $\mathbf{V}_0$ and $\nu_0$ are the model hyperparameters describing the prior belief in the AR parameters respectively. $\nu_0$ and $\mathbf{V}_0$ are the hyper parameters describing the prior belief in the AR parameters $\mathsf{a}, r$.

$$
\begin{gathered}
\mathbf{V}_{\mathsf{n}} = \mathbf{V}_{\mathsf{n-1}} + \begin{bmatrix} z_{\mathsf{n}} \\ \psi_{\mathsf{n}} \end{bmatrix} \begin{bmatrix} z_{\mathsf{n}} & {\psi_{\mathsf{n}}}^T \end{bmatrix} = \mathbf{V}_0 + \sum_{\mathsf{i}=1}^{\mathsf{n}} \begin{bmatrix} z_{\mathsf{i}} \\ \psi_{\mathsf{i}} \end{bmatrix} \begin{bmatrix} z_{\mathsf{i}} & \psi_{\mathsf{i}}^{\mathsf{T}} \end{bmatrix} \\
\nu_{\mathsf{n}} = \nu_{\mathsf{n-1}} + 1 = \nu_0 + \mathsf{n}
\end{gathered} \tag{5}
$$

# 3 Transfer learning between distributed AR inference tasks

## 3.1 Note on parameter inference knowledge transfer

In previous work which adopts principles of FPD, transfer learning was achieved by the target adopting the source's one step ahead data predictor $\mathsf{F_S}(z_{n+1}|\mathbf{z}_n)$ in its own local parameter inferences $\mathsf{F}(\theta|\mathbf{z}_n, \mathsf{F_S})$ [10][11][12]. Some time was allocated to investigations into an alternative method where the source makes available it's local parameter inferences $\mathsf{F_S}(\theta_\mathsf{S}|\mathbf{z}_{\mathsf{n,S}})$ to the target instead of a data predictor. Despite allocating time on this problem, no satisfactory or meaningful method was identified to transfer this knowledge without creating / designing a joint model of some kind. One approach which was investigated and initially suggested promise, was the design of an optimal likelihood function $\mathsf{L}^\circ(\theta|\mathbf{z}_\mathsf{n}, \mathsf{F_S})$, conditioned on knowledge from the source. Unfortunately this method led to no practical or non-trivial solution.

## 3.2 Data predictor knowledge transfer

The following subsection outlines the setting for knowledge transfer via the sharing of the source's one step ahead data predictor with the target inference task.

$$\mathsf{M_I}(\mathbf{a}, \mathsf{r}, z_\mathsf{n}|\ \mathbf{z}_{\mathsf{n}-1}) \equiv \mathsf{F}(\mathbf{a}, \mathsf{r}, z_\mathsf{n}|\ \mathbf{z}_{\mathsf{n}-1}) = \mathsf{F}(z_\mathsf{n}|\ \boldsymbol{a}, \mathsf{r})\mathsf{F}\,(\mathbf{a}, \mathsf{r}\mid \mathbf{z}_{\mathsf{n}-1}) \tag{6}$$

$$\mathsf{M}(\mathbf{a}, \mathsf{r}, z_\mathsf{n}|\ \mathsf{F_S}, \mathbf{z}_{\mathsf{n}-1}) \equiv \mathsf{F_S}(z_\mathsf{n}|\ \boldsymbol{\psi}_{\mathsf{S,n}})\Big|_{\boldsymbol{\psi}_{\mathsf{S,n}}=\boldsymbol{\psi}_\mathsf{n}} \mathsf{M}\,(\mathbf{a}, \mathsf{r}\mid \mathsf{F_S}, \mathbf{z}_{\mathsf{n}-1}) \tag{7}$$

Equation (6) defines $\mathsf{M_I}$, the chosen ideal model. It was chosen as it describes the full state of the process at time $n$. Equation (7) defines how the targets model is constrained by source knowledge. The target accepts the source's one-step-ahead data predictor $\mathsf{F_S}$ in place of $\mathsf{F}(z_\mathsf{n}|\mathbf{a}, \mathsf{r})$, asserting that the source distribution is predictive of the target's data stream. Note that distributions described with $\mathsf{F}$ are fixed form, while $\mathsf{M}$ are variational (i.e. must be designed).

Further note the choice of ideal distribution and specification on how the target accepts the source's one-step-ahead data predictor $\mathsf{F_S}$. This is not necessarily the only reasonable choice in describing this setting. Another assertion which should be noted is the choice to accept the source's distribution with the target's explanatory vector i.e. $\mathsf{F_S}(z_\mathsf{n}|\ \boldsymbol{\psi}_{\mathsf{S,n}})\Big|_{\boldsymbol{\psi}_{\mathsf{S,n}}=\boldsymbol{\psi}_\mathsf{n}}$.

$$\begin{aligned} \mathsf{M} \in \mathbf{M} \equiv \{&\text{models (7) with } \mathsf{F_S}(z_\mathsf{n}|\ \boldsymbol{\psi}_{\mathsf{S,n}})\Big|_{\boldsymbol{\psi}_{\mathsf{S,n}}=\boldsymbol{\psi}_\mathsf{n}} \\ &\text{fixed and } \mathsf{M}\,(\mathbf{a}, \mathsf{r}\mid \mathsf{F_S}, \mathbf{z}_{\mathsf{n}-1}) \text{ variational}\} \end{aligned} \tag{8}$$

**Proposition 1.** *The unknown model belongs to the knowledge constrained set,* $\mathsf{M} \in \mathbf{M}$ *(8), and the ideal model* $\mathsf{M_I}$ *is (6). Then the FPD-optimal model is*

$$\mathsf{M}^\circ(\mathbf{a}, \mathsf{r}, z_\mathsf{n}|\mathsf{F}_{\mathsf{S,n}-1}) \propto \mathsf{F_S}(z_\mathsf{n}|\boldsymbol{\psi}_{\mathsf{S,n}})\mathsf{M}^\circ(\mathbf{a}, \mathsf{r}|\mathsf{F}_{\mathsf{S,n}-1}, \mathbf{z}_{\mathsf{n}-1}) \tag{9}$$

*where*

$$\mathsf{M}^\circ(\mathbf{a}, \mathsf{r}\mid \mathsf{F}_{\mathsf{S,n}-1}, \mathbf{z}_{\mathsf{n}-1}) \propto \mathsf{F}(\mathbf{a}, \mathsf{r}|\mathbf{z}_{\mathsf{n}-1}) \exp\left[\int_{z_\mathsf{n}} \ln(\mathsf{F}(z_\mathsf{n}|\mathbf{a}, \mathsf{r}))\mathsf{F_S}(z_\mathsf{n}|\boldsymbol{\psi}_{\mathsf{S,n}})\mathsf{d}z_\mathsf{n}\right] \tag{10}$$

*Proof.* See [11] for similar form proof. □

**Proposition 2.** *Given the optimal source knowledge conditioned model (10), the appropriate model update procedure at each time step n is*

$$\begin{aligned} \mathbf{V}_\mathsf{n}^\circ &= \mathbf{V}_\mathsf{n} + \begin{bmatrix} \hat{z}_{\mathsf{S,n}} \\ \psi_\mathsf{n} \end{bmatrix} \begin{bmatrix} \hat{z}_{\mathsf{S,n}} & \psi_\mathsf{n}^\mathsf{T} \end{bmatrix} + \begin{bmatrix} \mathsf{r}_{\mathsf{S},z_n} & 0 & \cdots & 0 \\ 0 & 0 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \\ \nu_\mathsf{n}^\circ &= \nu_\mathsf{n} + 1 \end{aligned} \tag{11}$$

*where the sources one-step-ahead data predictors first and second moments are*

$$\hat{z}_{\mathsf{S,n}} = \boldsymbol{\psi}_\mathsf{n}^\mathsf{T}\hat{\mathbf{a}}_{\mathsf{S,n}} = \boldsymbol{\psi}_\mathsf{n}^\mathsf{T}\mathbf{V}_{\mathsf{aa,S,n}}^{-1}\mathbf{v}_{\mathsf{a1,S,n}},\ \nu_{\mathsf{S,n}} > 1 \tag{12}$$

$$\mathsf{r}_{\mathsf{S},z_n} = \lambda_{\mathsf{S,n}} \frac{1 + \boldsymbol{\psi}_{\mathsf{n}}^{\mathsf{T}} \mathbf{V}_{\mathsf{aa,S,n}}^{-1} \boldsymbol{\psi}_{\mathsf{n}}}{\nu_{\mathrm{S,n}} - 2}, \;\; \nu_{\mathsf{S,n}} > 2 \tag{13}$$

*Proof.*

$$\mathsf{M}^{\circ}(\mathbf{a}, \mathsf{r} | \mathsf{F}_{\mathsf{S}}, \mathbf{z}_{\mathsf{n}-1}) \propto \mathcal{N}\mathrm{iG}_{\mathsf{a,r}}(\mathbf{V}_{\mathsf{n}-1}, \nu_{\mathsf{n}-1}) \exp \left[ \int_{\mathsf{z_n}} \ln \mathcal{N}_{\mathsf{z_n}}(\psi_{\mathsf{n}}^{\mathsf{T}} \mathbf{a}, \mathsf{r}) \mathsf{St}_{\mathsf{z_n}}(\cdot) \mathsf{d}z_{\mathsf{n}} \right] \tag{14}$$

Taking exponent component.

$$\int_{z_{\mathsf{n}}} \left( \ln \frac{1}{\sqrt{2\pi \mathsf{r}}} - \frac{1}{2\mathsf{r}} (z_{\mathsf{n}} - \psi_{\mathsf{n}}^{\mathsf{T}} \mathbf{a})^2 \right) \mathsf{St}_{z_{\mathsf{n}}}(\cdot) \mathsf{d}z_{\mathsf{n}} \tag{15}$$

$$= \ln \frac{1}{\sqrt{2\pi \mathsf{r}}} - \frac{1}{2\mathsf{r}} \left( \psi_{\mathsf{n}}^{\mathsf{T}} \mathbf{a}\mathbf{a}^{\mathsf{T}} \psi_{\mathsf{n}} - 2\psi_{\mathsf{n}}^{\mathsf{T}} \mathbf{a} \hat{z}_{\mathsf{S,n}} + \int_{z_n} z_{\mathsf{n}}{}^2 \mathsf{St}_{z_{\mathsf{n}}}(\cdot) \mathsf{d}z_n \right) \tag{16}$$

$$= \ln \frac{1}{\sqrt{2\pi \mathsf{r}}} - \frac{1}{2\mathsf{r}} \left( (\psi_{\mathsf{n}}^{\mathsf{T}} \mathbf{a} - \hat{z}_{\mathsf{S,n}})^2 + r_{z_n,\mathsf{S}} \right) \tag{17}$$

Taking (14) and (17)

$$\mathsf{M}^{\circ}(\mathbf{a}, \mathsf{r} | \mathsf{F}_{\mathsf{S}}, \mathbf{z}_{\mathsf{n}-1}) \propto \mathsf{r}^{-\frac{\nu_{\mathsf{n}-1}+1}{2}} \exp \left( -\frac{1}{2\mathsf{r}} \begin{bmatrix} -1 & \mathbf{a}^{\mathsf{T}} \end{bmatrix} \mathbf{V}_{\mathsf{n}-1} \begin{bmatrix} -1 \mathbf{a} \end{bmatrix} + \left( \psi_{\mathsf{n}}^{\mathsf{T}} \mathbf{a} - \hat{\mathsf{z}}_{\mathsf{S,n}} \right)^2 + \mathsf{r}_{\mathsf{z_n},\mathsf{S}} \right) \tag{18}$$

where the sources one-step-ahead data predictors first and second moments are

$$\hat{z}_{\mathsf{S,n}} = \boldsymbol{\psi}_{\mathsf{n}}^{\mathsf{T}} \hat{\mathbf{a}}_{\mathsf{S,n}} = \boldsymbol{\psi}_{\mathsf{n}}^{\mathsf{T}} \mathbf{V}_{\mathsf{aa,S,n}}^{-1} \mathbf{v}_{\mathsf{a1,S,n}}, \;\; \nu_{\mathsf{S,n}} > 1 \tag{19}$$

$$\mathsf{r}_{\mathsf{z_n},\mathsf{S}} = \lambda_{\mathsf{S,n}} \frac{1 + \boldsymbol{\psi}_{\mathsf{n}}^{\mathsf{T}} \mathbf{V}_{\mathsf{aa,S,n}}^{-1} \boldsymbol{\psi}_{\mathsf{n}}}{\nu_{\mathsf{S,n}} - 2}, \;\; \nu_{\mathsf{S,n}} > 2 \tag{20}$$

□

The update model outlined in proposition 2 for accepting source knowledge is of similar form to the typical Bayesian update procedure in (5) where the sources expected datum is used in place of an observed target datum and a new additive value in the $(1, 1)$ element of the EIM. Note that the addition of the sources' second moment $\mathsf{r}_{\mathsf{S},z_n}$ has no effect on the expected value of the targets parameters and can only increase target parameter uncertainty. Also note how the targets $\nu$ hyperparameter is updated identically to the typical update procedure (5), inferring that the target accepts this new knowledge as if it were a full datum and then after increases uncertainty accordingly. An important consequence of this characteristic of the transfer is that accepting knowledge from the a source inference node may not always be in the targets best interest.

### 3.3 Knowledge transfer with obstinate acceptance criterion

With regards to robustness to real world data and where assumptions outlined in the preceding sections do not hold or is not a reasonable assertion, a naive source knowledge acceptance criterion is proposed which we have informally called the obstinate acceptance criterion. The obstinate acceptance criterion is characterised as the following:

> *If after accepting and processing of knowledge from a source inference task by the target inference task results in the target becoming less confident in its local parameters, the target should reject the source knowledge. Conversely, if after accepting knowledge from a source inference task results in the target becoming more sure in its parameters, then the knowledge should not be rejected.*

$$\mathsf{M}^{\circ}(\mathbf{a}, \mathsf{r} \mid \mathsf{F}_{\mathsf{S,n}-1}, \mathbf{z}_{\mathsf{n}-1}) \equiv \begin{cases} \mathsf{F}(\mathbf{a}, \mathsf{r} | \mathbf{z}_{\mathsf{n}-1}) & , \sigma_{\boldsymbol{a}} | \mathsf{F}_{\mathsf{S}} > \sigma_{\boldsymbol{a}} \\ \mathsf{F}(\mathbf{a}, \mathsf{r} | \mathbf{z}_{\mathsf{n}-1}) \exp \left[ \int_{z_{\mathsf{n}}} \ln(\mathsf{F}(z_{\mathsf{n}} | \mathbf{a}, \mathsf{r})) \mathsf{F}_{\mathsf{S}}(z_{\mathsf{n}} | \boldsymbol{\psi}_{\mathsf{S,n}}) \mathsf{d}z_{\mathsf{n}} \right] & , \text{otherwise} \end{cases} \tag{21}$$

Equation (21) shows a definition of the obstinate acceptance criterion where $\sigma_{\boldsymbol{a}} | \mathsf{F}_{\mathsf{S}} \cdot \mathsf{I}_{\mathsf{p}}$ is the second moment of $\mathsf{F}(\mathbf{a} | \mathsf{F}_{\mathsf{S}}, \cdot)$ and $\sigma_{\boldsymbol{a}} \cdot \mathsf{I}_{\mathsf{p}}$ is the second moment of $\mathsf{F}(\mathbf{a} | \cdot)$. The obstinate acceptance criterion is a naive and blunt method to decide what external knowledge should be accepted by the target. Future work should consider the use of a 'soft' boundary where source knowledge is weighted according to how relevant the target believes it to be.

# 4 Experiments

## 4.1 Experimental setup

In this section, the previously described transfer learning algorithm is tested on both artificially generated and also real data. In each case, the performance of the target model one-step-ahead predictor, $\mathsf{F}(z_{n+1}|\mathsf{z}_n, \cdots)$, is the principal focus.

## 4.2 Monte-Carlo Simulations

The following subsection outlines a set of Monte Carlo simulations within a context of AR-4 processes. In each experiment, 2 sets of randomly generated AR-4 parameters are sampled and used to generate 2 series of 130 data each. In each instance, the first 100 data are discarded as part of an initialisation procedure. The first set is referred to as the target data $\mathsf{z_n}$, while the second data set is referred to as the source's dataset $\mathsf{z_{S,n}}$. In each experiment all inference tasks are performed synchronously (e.g. at time step $T = 10$ every task processes their $10^{\text{th}}$ respective datum. If relevant the transfer step occurs after processing of the local datum).

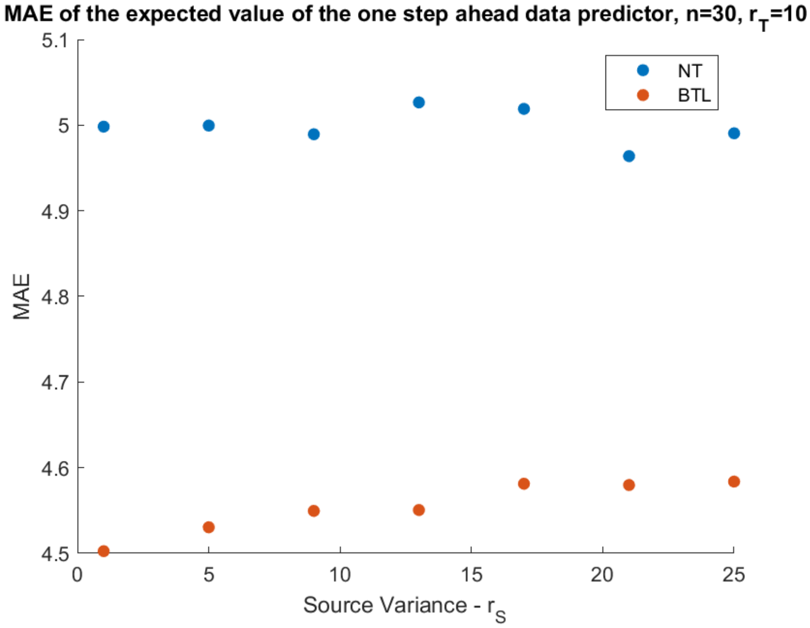

Figure 2: Mean absolute error between the expected value of the target's one step ahead data predictor $\mathsf{F}(z_{n+1}|\mathsf{z}_n)$ and the true realisation of the datum at every time step with varying levels of source task variance. Experiment was conducted over 200'000 Monte Carlo iterations.

Figure 2 show the results of an experiment where the target's generative model variance is held constant at a value of 10 and with a varying source's generative model variance as the operating condition. The metric adopted is the Mean Absolute Error (MAE) between the expected value of the target's one step ahead data predictor $\mathsf{F}(z_{n+1}|\mathsf{z}_n)$ and the true realisation of the next datum. Results show that the BTL approach results in a modest increase in performance with MAE values staying around the 4.5 point for each experiment in comparison to a steady 5 result for the no transfer setting. When using this MAE metric, the source variance operating condition appears to be independent to the mean prediction accuracy of the target.

$$\mathsf{D_{KL}}\Big(\mathsf{F}(z_n|\mathsf{z}_{n-1}, \mathsf{F_S})\Big|\Big|\mathsf{F}(z_n|\boldsymbol{a}, r)\Big) \tag{22}$$

Figure 3 show the results of the same experiment as described in the paragraph above with a change in metric to the average KLD between the target's one step ahead data predictor $\mathsf{F}(z_n|\mathsf{z}_{n-1}, \mathsf{F_S})$ and the true generative model $\mathsf{F}(z_n|\boldsymbol{a}, r)$ and is shown in equation (22). This atypical metric was used as it describes the relative entropy between the distribution possessed by the target, describing its belief in its next datum and the true generative model which describes the underlying distribution which the datum is

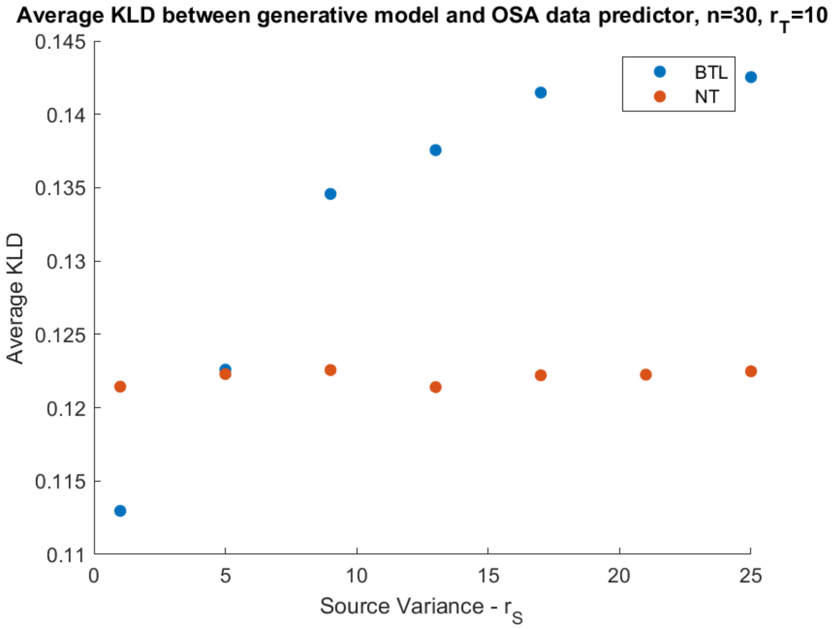

Figure 3: Mean Kullback-Liebler Divergence from the true generative model $\mathsf{F}(z_\mathsf{n}|\mathbf{a},\mathsf{r},\psi_\mathsf{n}) = \mathcal{N}(\psi_\mathsf{n}^\mathsf{T}\mathbf{a},\mathsf{r})$ to the one step ahead data predictor $\mathsf{F}(z_\mathsf{n}|\mathbf{z}_{\mathsf{n}-1}) = \mathsf{St}-\mathsf{t}(.)$ at every time step with varying levels of source task variance.

sampled. The true generative model can be considered the best the target could hope to achieve in terms of predictive power. The average KLD metric has the further advantage of comparing the "*accuracy in distribution*" in contrast to the more typical comparison of a moment of the targets distribution (e.g. first moment) to the realisation of the relevant distribution (i.e. the true generative model).

Results from this second experiment are somewhat different to the first experiment where the metric used was the MAE. Once again the BLT method performs well, but only for lower values of source variance. For source variance values above $\sim 5$, the transfer in knowledge leads to a higher mean KLD value between the targets data predictor and the true generative model when compared to the no transfer scenario. This illustrates the effect of the negative transfer from the source inference task reducing the effective predicative power of the target.

## 4.3 Real-Data Experimentation

Motivated to experiment with the effectiveness of the algorithm in a real world environment, testing was conducted on price data of various common food stuffs since 1850. The data set *long-term-price-index-in-food-commodities-1850-2015*[1] was used which provides data on the price of common food stuffs yearly since 1850. In our test, we adopt AR models of varying order to model prices of lamb, pork and beef. We then use the models for both lamb and pork to help increase the learning rate of beef in an online manner. Diffuse (i.e. flat) priors were used for each of the 3 models.

Figure 4 show the MAE values between the targets one step ahead data prediction and the true realisation of the next datum at every time step. Results show that when using lamb knowledge, beef price predictions were more accurate and when using both lamb and pork knowledge, there was an even larger reduction in MAE (although with more diminishing returns).

Figure 5 shows the same experiment as described previously. The obstinate acceptance criterion is not adopted in this experiment. Here we can see the results are similar to the case when the obstinate acceptance criterion is adopted with the exception of some new extreme outliers. In this experiment it is clear to see that the obstinate acceptance criterion has made the algorithm more robust to negative transfer.

[1] `https://ourworldindata.org/food-prices`

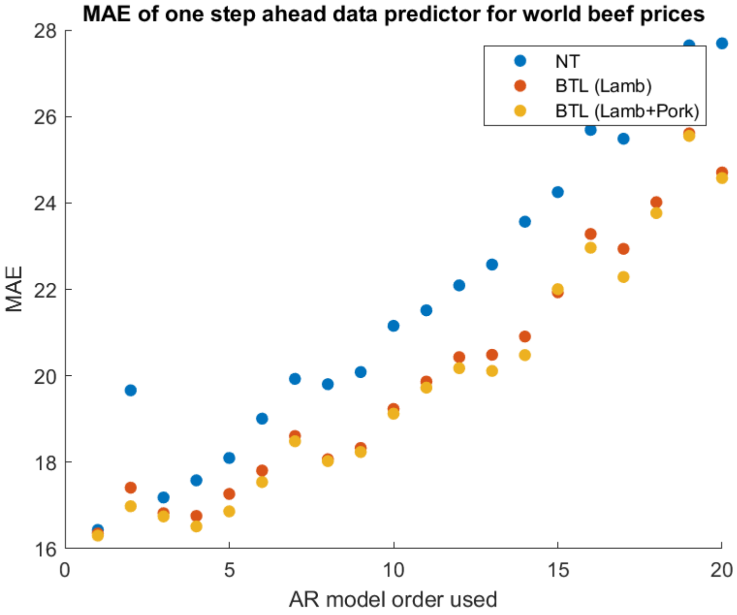

Figure 4: Mean absolute error between the target's one-step ahead data predictor and true realisation at every time step (yearly) for the price of beef with varying AR model orders. Three conditions are investigated, no transfer (NT), Bayesian Transfer Learning (BTL) with lamb model transfer only and BTL with lamb and pork model knowledge transfer. This experiment incorporates the obstinate acceptance criterion.

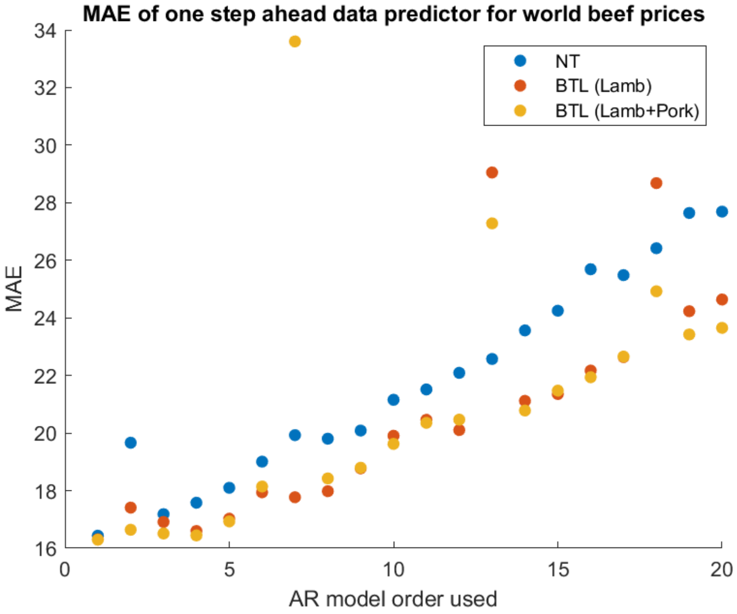

Figure 5: Mean absolute error between the target's one-step ahead data predictor and true realisation at every time step (yearly) for the price of beef with varying AR model orders. Three conditions are again investigated, no transfer (NT), Bayesian Transfer Learning (BTL) with lamb model transfer only and BTL with lamb and pork model knowledge transfer. This experiment does not incorporate the obstinate acceptance criterion. Note the change in range of the y-axis relative to fig 4.

# 5 Discussion and Conclusions

This document approaches transfer learning in a distributed autoregressive setting where multiple source learners provide a target learnerknowledge in the form of a distribution at each time step.

In this document an algorithm for the transfer of knowledge between a source and target autoregressive inference task is proposed. The algorithm is consistent with the axiomatically justified FPD distributional design framework. At every time step the target inference task accepts a distribution from each of the source inference tasks and uses it to increase the predictive power of the targets own model. With interests to making the algorithm more robust to non-ideal settings where design assumptions no longer hold, an obstinate acceptance criterion is proposed. The criterion asserts that the target should only accept external knowledge if it increases its confidence in its parameters.

Experiments adopting the transfer learning algorithm and acceptance criterion were conducted on both artificially generated data and a real dataset provided by `ourworldindata.org`. In both cases results show good performance with significant positive knowledge transfer occurring. The adoptions of the obstinate acceptance criterion was also shown to increase the ability of the method to reject external knowledge which otherwise would have resulted in negative transfer.

Results in this document show promise for the algorithm. Regardless, further work is required to bring the research to publication. Notably more simulations and experimentation is required in relation to ensuring accuracy of the claim of increased performance. Additionally, a full comparison of the algorithm against other state-of-the-art transfer learning algorithms is desirable.

# References

[1] Torrey, L., & Shavlik, J. (2010). Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (pp. 242-264). IGI global.

[2] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[3] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. Journal of Big data, 3(1), 9.

[4] Marx, Z., Rosenstein, M. T., Kaelbling, L. P., & Dietterich, T. G. (2005). Transfer learning with an ensemble of background tasks. Inductive Transfer, 10.

[5] Raina, R., Ng, A. Y., & Koller, D. (2006, June). Constructing informative priors using transfer learning. In Proceedings of the 23rd international conference on Machine learning (pp. 713-720).

[6] Karbalayghareh, A., Qian, X., & Dougherty, E. R. (2018). Optimal Bayesian transfer learning. IEEE Transactions on Signal Processing, 66(14), 3724-3739.

[7] Kárný, M., & Guy, T. V. (2006). Fully probabilistic control design. Systems & Control Letters, 55(4), 259-265.

[8] Kárný, M. (1996). Towards fully probabilistic control design. Automatica, 32(12), 1719-1722.

[9] Kárný, M., & Kroupa, T. (2012). Axiomatisation of fully probabilistic design. Information Sciences, 186(1), 105-113.

[10] Papež, M., & Quinn, A. (2019, December). Bayesian transfer learning between Gaussian process regression tasks. In 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (pp. 1-6). IEEE.

[11] Foley, C., & Quinn, A. (2017). Fully probabilistic design for knowledge transfer in a pair of Kalman filters. IEEE Signal Processing Letters, 25(4), 487-490.

[12] Papež, M., & Quinn, A. (2020). Bayesian transfer learning between Student-t filters. Signal Processing, 107624.

[13] Ephraim, Y., & Roberts, W. J. (2005). Revisiting autoregressive hidden Markov modeling of speech signals. IEEE Signal Processing Letters, 12(2), 166-169.

[14] Juang, B. H., & Rabiner, L. (1985). Mixture autoregressive hidden Markov models for speech signals. IEEE Transactions on Acoustics, Speech, and Signal Processing, 33(6), 1404-1413.

[15] Pacurar, M. (2008). Autoregressive conditional duration models in finance: a survey of the theoretical and empirical literature. Journal of Economic Surveys, 22(4), 711-751.

[16] Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. Journal of Applied Econometrics, 28(5), 777-795.

[17] Šmídl, V., & Quinn, A. (2005). Mixture-based extension of the AR model and its recursive Bayesian identification. IEEE Transactions on Signal Processing, 53(9), 3530-3542.