Unlabelled Data Improves Bayesian Uncertainty Calibration under Covariate Shift

Alex J. Chan [1] **Ahmed M. Alaa** [2] **Zhaozhi Qian** [1] **Mihaela van der Schaar** [1 2 3]

Abstract

Modern neural networks have proven to be powerful function approximators, providing state-of-the-art performance in a multitude of applications. They however fall short in their ability to quantify confidence in their predictions — this is crucial in high-stakes applications that involve critical decision-making. Bayesian neural networks (BNNs) aim at solving this problem by placing a prior distribution over the network's parameters, thereby inducing a posterior distribution that encapsulates predictive uncertainty. While existing variants of BNNs based on Monte Carlo dropout produce reliable (albeit approximate) uncertainty estimates over in-distribution data, they tend to exhibit over-confidence in predictions made on target data whose feature distribution differs from the training data, i.e., the *covariate shift* setup. In this paper, we develop an approximate Bayesian inference scheme based on *posterior regularisation*, wherein unlabelled target data are used as "pseudo-labels" of model confidence that are used to regularise the model's loss on labelled source data. We show that this approach significantly improves the accuracy of uncertainty quantification on covariate-shifted data sets, with minimal modification to the underlying model architecture. We demonstrate the utility of our method in the context of transferring prognostic models of prostate cancer across globally diverse populations.

1. Introduction

Modern neural networks have achieved the state of the art predictive performance in a wide variety of applications. They are especially useful in areas where a large quantity of labelled i.i.d data are available (Krizhevsky et al., 2012). However, neural networks fall short in their ability to quantify their confidence in the predictions, which leads to difficulties to apply them to mission critical domains. The immediate problem is that neural networks may issue *erroneous predictions with high confidence* (Ovadia et al., 2019). These over-confident predictions would mislead rather than inform human experts' decisions and can lead to severe consequences in high-stakes applications.

In practice, the task of quantifying predictive uncertainty is even more challenging because the training and testing data are typically not i.i.d due to the impact of exogenous factors over time or the inconsistency in data collection. This situation is known as *covariate shift* (Shimodaira, 2000) and various research has indicated that this may cause neural networks to display unexpected behaviour (Ovadia et al., 2019). In the extreme case, they may confidently produce nonsensical predictions for out-of-distribution adversarial examples (Madry et al., 2017). While in this work we do not consider the scenario of a targeted adversarial attack, we would like the network to return a high uncertainty prediction if a test point falls far from any of the training data. We motivate the need for work in the area with a concrete example; consider the case of trying to predict the mortality rate for a group of patients with prostate cancer in a country where we have no labels due to tight privacy regulations on medical data. We have access though to labelled examples from another country which we could use to train the model, the problem being that the populations of each country may differ in their distribution so a model purely trained on the labelled data may perform poorly on the unlabelled data both in terms of accuracy and uncertainty estimation. This type of problem is common in the medical setting and errors here are especially damaging since model predictions may have a direct impact on the treatment a patient receives.

Bayesian neural networks (BNNs) (Neal, 2012) aim to solve the uncertainty quantification problem by learning neural networks via Bayesian inference, a principled way to reason under uncertainty. BNNs encapsulate the prediction uncertainty in the posterior predictive distribution, which is typically intractable and has to be approximated (Graves, 2011; Blundell et al., 2015). While existing approximation methods are able to produce reliable uncertainty estimates

[1] University of Cambridge, Cambridge, UK [2] University of California, Los Angeles, USA [3] Alan Turing Institute, London, UK. Correspondence to: Alex J. Chan <ajc340@cam.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

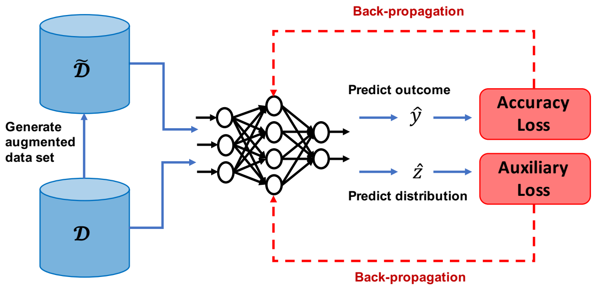

Figure 1. **High-level depiction of our approach.** We first generate our augmented data set with pseudo-labels before feeding forward to make predictions and then back-propagating both errors through the network.

over in-distribution data, it has been shown that they tend to be over-confident under covariate shift.

In this paper, we propose Transductive Dropout, a method leveraging information from the unlabelled target data to find a better approximation to the posterior. We make the following observation: a point being in the target data is an indication that the model should output higher uncertainty because the target distribution is not well-represented by training data due to covariate shift. Therefore, we use whether the data come from training or target set as a "pseudo-label" of model confidence. This naturally leads to a posterior regularisation term which we incorporate into the variational approximation objective. Our regulariser can be easily applied to many of the current network architectures and inference schemes — here, we demonstrate its usefulness in Monte Carlo Dropout, showing that it much more appropriately quantifies its uncertainty under covariate shift. Empirical evaluations demonstrate that our method performs competitively compared to Bayesian and frequentist approaches in the task of prostate cancer mortality prediction across globally diverse populations.

2. Related Work

2.1. Overview of Related Methods

Utilising unlabelled data to improve uncertainty estimate under covariate shift is a previously less explored area in the literature. Here we highlight some of the key methods in the surrounding fields to contextualise our work.

Bayesian Uncertainty Estimate for Neural Networks Bayesian methodology has been applied to quantify the predictive uncertainty of neural networks leading to a large family of methods known as Bayesian Neural Networks (BNN). BNN learns a posterior distribution over parameters that encapsulates the model uncertainty. Due the complexity of deep neural networks, the exact posterior is usually intractable. Hence, much of the research in BNN literature is devoted to finding better approximate inference algorithms for the posterior. Popular approximate Bayesian approaches include dropout-based variational inference (Gal & Ghahramani, 2016; Kingma et al., 2015) and Stochastic Variational Bayesian Inference (Blundell et al., 2015; Graves, 2011; Louizos & Welling, 2017). These methods are known to achieve reliable uncertainty estimate in i.i.d scenario. However, recent research has cast doubt about the validity of these uncertainty estimates under covariate shift (Ovadia et al., 2019). Moreover, the above methods do not make use of any unlabelled data for training or inference.

Semi-Supervised Learning Semi-supervised learning (SSL) covers the broad field of learning from both labelled and unlabelled data (Zhu & Goldberg, 2009). It s generally separated into two with most of the work covering *inductive* SSL which aims to use the unlabelled data to learn a general mapping from the features to the outcome. Many recent works encourage the model to generalise better by using a *regularisation* term computed on the unlabelled data Berthelot et al. (2019). This includes *entropy minimisation* which encourages the model to produce confident predictions on unlabelled data (Grandvalet & Bengio, 2005; Lee, 2013; Jean et al., 2018) and *consistency regularisation* which ensures the predictions for slightly perturbed data stay similar (Sajjadi et al., 2016). The other split covers *transductive* SSL where the aim is to make predictions over only the unlabelled points given with no need to generalise further. As we will show later, the proposed Transductive Dropout fits more into this framework, using the unlabelled data as a regulariser in order to induce a better variational approximation to the intractable posterior distribution.

However, our work is significantly different from traditional SSL in several ways. First, we note that most existing works in SSL focus entirely on using unlabelled data to improve predictive performance (e.g. accuracy), but much less thoughts have been given to improving the uncertainty estimate for those predictions, which is the focus of this paper. Furthermore, our work explicitly addresses the issue of covariate shift between source and target data whereas traditional SSL often assumes that they are i.i.d. In addition, most of the recent work in SSL considers problems like image classification and natural language processing where the methods can leverage the complicated dependencies in the features - we don't consider this a focus and develop a method that works appropriately for tabular data as well.

Unsupervised Domain Adaptation Unsupervised domain adaptation (UDA) is the task of training models to achieve better performance on a target domain, with access to only unlabelled data in the target domain and labelled data from a (different) source domain. Kouw & Loog (2019) contains a detailed review of popular UDA methods. As with SSL, existing works on UDA centre around improving predictive performance rather than producing well-calibrated uncertainty estimates. Our work contributes to the UDA literature by proposing a method to improve the uncertainty estimates on the predictions made in the target domain.

Transfer Learning In the setting of transfer learning (Torrey & Shavlik, 2010) the task does involve a change in distribution over features but typically also involves some amount of labels on the target set (known as one-shot or few-shot learning). This has led to a lot of work that uses the training set to learn a useful prior for a second model that can be trained on the labelled data in the target set (Raina et al., 2006; Karbalayghareh et al., 2018). Given the complete lack of labels in our target data set this is inapplicable for our problem.

3. Preliminaries

3.1. Notation and Problem Setup

Let $\mathbf{x} \in \mathbb{R}^d$ be a d-dimensional feature vector, and $y \in \mathcal{Y}$ be the prediction target; where $\mathcal{Y} = \mathbb{R}$ for regression targets, and $\mathcal{Y} = \{1, \ldots, K\}$ for K-class classification targets. We are presented with *two* sources of training data: a labelled data set $\mathcal{D}_L$, and an unlabelled data set $\mathcal{D}_U$. The labelled data set comprises a collection of n feature-label pairs, i.e., $\mathcal{D}_L = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, whereas the unlabelled set comprises a collection of m feature instances $\mathcal{D}_U = \{\mathbf{x}_j\}_{j=1}^m$.

We assume that $\mathcal{D}_L = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consists of i.i.d samples of features and labels drawn from the distribution

$$(\mathbf{x}_i, y_i) \sim p(\mathbf{x}) \times p(y|\mathbf{x}),\ \forall i \in \{1, \ldots, n\},$$

where both $p(\mathbf{x})$ and $p(y|\mathbf{x})$ are unknown, and could only be accessed empirically through $\mathcal{D}_L$. Throughout the paper, we will refer to $p(\mathbf{x})$ as the *feature distribution* — feature instances in the unlabelled data set are assumed to be drawn from a shifted feature distribution as follows:

$$\mathbf{x}_j \sim p'(\mathbf{x}),\ \forall j \in \{1, \ldots, m\},$$

where $p'(\mathbf{x}) \neq p(\mathbf{x})$, whereas the unobserved labels in the data set $\mathcal{D}_U$, i.e., the blue dots in Figure 2 corresponding to $\{y_j\}_{j=1}^m$, are generated from the same conditional distribution $y_j \sim p(y|\mathbf{x}_j)$. Note that even though the feature distributions $p'(\mathbf{x})$ and $p(\mathbf{x})$ differ, the conditional $p(y|\mathbf{x})$ is invariant across the two data sets. This situation is commonly known as *covariate shift* (Shimodaira, 2000). We denote the entirety of observed data $\mathcal{D} = \{\mathcal{D}_L \cup \mathcal{D}_U\}$.

3.2. Learning from (and for) unlabelled data

Our key objective is to use the (source) labelled data set $\mathcal{D}_L$ to train a model that would be applied to the (target) unlabelled data set $\mathcal{D}_U$. However, since the feature distributions in $\mathcal{D}_L$ and $\mathcal{D}_U$ mismatch, we cannot expect a model trained on $\mathcal{D}_L$ to perfectly generalise to $\mathcal{D}_U$. Thus, we aim at training the model to *learn* which prediction instances can be *confidently* transferred from $\mathcal{D}_L$ to $\mathcal{D}_U$, and which cannot be confidently generalised across the two distributions. To this end, we train the model to score its uncertainty on predictions issued for all feature instances in $\mathcal{D} = \{\mathcal{D}_L \cup \mathcal{D}_U\}$.

We take a *Bayesian* approach to uncertainty estimation. That is, for a model with parameter $\boldsymbol{\theta}$ and a test point $\mathbf{x}^* \sim p'(\mathbf{x})$, the Bayesian posterior distribution over y^* is

$$\underbrace{p(y^*|\mathbf{x}^*, \mathcal{D})}_{\textbf{Total uncertainty}} = \int \underbrace{p(y^*|\mathbf{x}^*, \boldsymbol{\theta})}_{\substack{\textbf{Data}\\\textbf{uncertainty}}} \underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\substack{\textbf{Model}\\\textbf{uncertainty}}} d\boldsymbol{\theta}. \quad (1)$$

The posterior decomposition in (1) comprises two types of uncertainty (Malinin & Gales, 2018): *data uncertainty*, also referred to as aleatoric uncertainty, is the variance of the true conditional distribution $p(y|\mathbf{x})$, reflecting the inherent ambiguity or noise in the true labels y (Gal et al., 2017). The second type of uncertainty, *model uncertainty*, pertains to the model's epistemic uncertainty created by the lack of training examples in the vicinity of the test feature $\mathbf{x}^*$. Since the conditional $p(y|\mathbf{x})$ is invariant across the source and target distributions, it is the model uncertainty that we focus on.

3.3. Standard approximate Bayesian is not enough...

A true Bayesian model (with appropriate priors) would completely capture model uncertainty in $\mathcal{D}_U$ by simply training the model on $\mathcal{D}_L$ in a supervised fashion, while completely ignoring the unlabelled data in $\mathcal{D}_U$ (Sugiyama & Storkey, 2007). However, exact Bayesian inference in neural networks is generally intractable (and computationally expensive), hence existing practical solutions to Bayesian modelling rely on approximate inference schemes, for example based on Monte Carlo dropout (MCDP) (Gal & Ghahramani, 2016).

While approximate inference via MCDP — with appropriate hyper-parameter tuning — provides reliable uncertainty estimates for in-distribution data (i.e., feature instances in $\mathcal{D}_L$), it has been shown in Ovadia et al. (2019) that these methods lead to miscalibrated estimates of uncertainty for out-of-distribution data. In the next Section, we develop an approximate Bayesian scheme that makes use of the unlabelled data in $\mathcal{D}_U$ to provide more accurate uncertainty estimates on the predictions made for features instances drawn from the shifted distribution $p'(\mathbf{x})$.

4. Transductive Regularisation

How can we use our knowledge of the unlabelled data in $\mathcal{D}_U$ to improve the uncertainty estimates on predictions made for the target distribution $p'(\mathbf{x})$? In this Section, we develop an approximate Bayesian method tailored to this task. Here, we regard a neural network (NN) as a distribution $p(y|\mathbf{x},\theta)$ that assigns a probability to each possible output y.

4.1. Variational inference with posterior regularisation

In a Bayesian framework, we specify a prior distribution $p(\theta)$ on the NN parameters, and obtain the posterior $p(\theta|\mathcal{D})$ via Bayes rule. In practice, the posteriors $p(\theta|\mathcal{D})$ and $p(y|\mathbf{x},\mathcal{D})$ in (1) are both intractable. To address this issue, we use variational inference, whereby we use a surrogate distribution $q_\phi(\theta)$ parameterised by ϕ to approximate $p(\theta|\mathcal{D})$. The parameter ϕ is obtained by minimising the KL-divergence between p and q as follows (Graves, 2011):

$$\phi^* = \arg\min_\phi \mathrm{KL}\big[\, q_\phi(\theta) \,||\, p(\theta|\mathcal{D}) \,\big]. \tag{2}$$

In practice the KL divergence is not minimised directly, rather it is achieved my maximising the Evidence Lower BOund (ELBO), which can be written as:

$$\mathcal{F}(\mathcal{D},\phi) = \mathbb{E}_{q_\phi}\big[\log p(\mathcal{D}|\theta)\big] - \mathrm{KL}\big[\, q_\phi(\theta) \,||\, p(\theta) \,\big], \tag{3}$$

being seen as the balance of two terms. The objective being to maximise the log-likelihood under the surrogate distribution (first term) while regularising the approximation to not be too far from the prior (second term). Variational inference also leads to an approximate posterior predictive distribution $q_\phi(y|\mathbf{x},\mathcal{D})$, obtained by replacing $p(\theta|\mathcal{D})$ in (1) with its variational counterpart $q_\phi(\theta)$. Note that the unlabelled data in $\mathcal{D}_U$ is ancillary to the optimisation problem in (2), since mere evidence maximisation would render $p(\theta|\mathcal{D}_L)$ as the only relevant conditional for finding the variational parameter ϕ. Hence, the vanilla variational Bayes is insufficient in our setup as it cannot capitalise on our knowledge of the unlabelled data in $\mathcal{D}_U$.

To incorporate the unlabelled data in $\mathcal{D}_U$ into our inference machine, we resort to *posterior regularisation* (Zhu et al., 2014). That is, instead of computing the variational posterior that best matches the true posterior in KL distance, we add a regulariser Ω to the objective in (2), i.e.,

$$\phi^* = \arg\min_\phi \mathrm{KL}\big[\, q_\phi(\theta) \,||\, p(\theta|\mathcal{D}) \,\big] + \Omega(q_\phi(\theta|\mathcal{D})), \tag{4}$$

in order to explicitly influence the learned variational posterior so that it produces the desired uncertainty profiles, i.e., posterior variance, over the target feature distribution $p'(\mathbf{x})$.

What do our sought-after uncertainty profiles look like? In order to design the regulariser Ω, we first need to specify the influences it needs to exert on the learned variational posterior q_ϕ. Let $\mathbb{E}[\,q\,]$ and $\mathbb{V}[\,q\,]$ denote the mean and variance of a given distribution q, respectively. A "good" variational posterior is one that matches the true posterior $p(\theta\,|\,\mathcal{D})$, and induces the following uncertainty profile: for any pair of features $\mathbf{x},\mathbf{x}' \sim p'(\mathbf{x})$ drawn from the target distribution, the variational posterior satisfies the following condition:

$$\mathbb{V}[\, q_\phi(y|\mathbf{x},\mathcal{D}) \,] \geq \mathbb{V}[\, q_\phi(y|\mathbf{x}',\mathcal{D}) \,] \Leftrightarrow p(\mathbf{x}') \geq p(\mathbf{x}). \tag{5}$$

That is, the variance of the variational posterior, which quantifies the model's uncertainty, should be smaller for target test points that are close (in distribution) to the labelled data in $\mathcal{D}_L$, and vice versa. The key idea behind our posterior regularisation approach is that the augmentation of labelled and unlabelled data serve as "pseudo-labels" of model confidence — by regarding the condition in (5) as an auxiliary classification task wherein q_ϕ predicts whether a feature $\mathbf{x}$ is drawn from the source or target distributions, we can "train" q_ϕ to make this binary prediction via its variance. Building on this insight, the rest of this Section builds a regulariser Ω that enables q_ϕ to discriminate source and target features.

4.2. Posterior regularisation via transductive dropout

As discussed above, we seek a variational posterior that best fits the labelled data in $\mathcal{D}_L$, and discriminates source and target data. Before proceeding, we first define an augmented data set $\widetilde{\mathcal{D}} = \{(\mathbf{x}_k, y_k, z_k)\}_{k=1}^{n+m}$, where

$$(\mathbf{x}_k, y_k, z_k) = \begin{cases} (\mathbf{x}_k, y_k, 0), & \forall\ (\mathbf{x}_k, y_k) \in \mathcal{D}_L, \\ (\mathbf{x}_{k-n+1}, *, 1), & \forall\ \mathbf{x}_{k-n+1} \in \mathcal{D}_U, \end{cases}$$

where $*$ corresponds to a missing value for the label y. In addition, we define the monotonic function $g : \mathbb{R}^+ \to [0,1]$ as a map from positive real values to the unit interval. Given the variational distribution q_ϕ, our prediction of whether the feature $\mathbf{x}$ comes from the source or target distributions is

$$\hat{z}_\phi(\mathbf{x}) \triangleq g\Big(\mathbb{V}[\, q_\phi(y|\mathbf{x},\mathcal{D}) \,]\Big), \tag{6}$$

which follows directly from the condition in (5). Given (6), we define the regulariser Ω in (4) as the cross-entropy loss between predicted and true auxiliary variables, $\hat{z}$ and z, i.e.,

$$\begin{aligned} &\Omega(q_\phi(\theta\,|\,\widetilde{\mathcal{D}})) \\ &\quad = \sum_{k=1}^{n} \log\big(1 - \hat{z}_\phi(\mathbf{x}_k)\big) + \sum_{k=n+1}^{n+m} \log\big(\hat{z}_\phi(\mathbf{x}_k)\big). \end{aligned} \tag{7}$$

Thus, our variational posterior is obtained by plugging the regulariser $\Omega(q_\phi(\theta\,|\,\widetilde{\mathcal{D}}))$ in (4) and solving for ϕ, with the optional inclusion of a hyperparamter λ to control the level of regularisation. The exact choice of g can as well be controlled although from our experiments it made little difference, and we settled on $g(x) = 1 - \frac{1}{1+x}$. We note that this regularisation scheme addresses the issue of over-confident predictions on the target set without taking the naive approach of just increasing the variance everywhere — it is

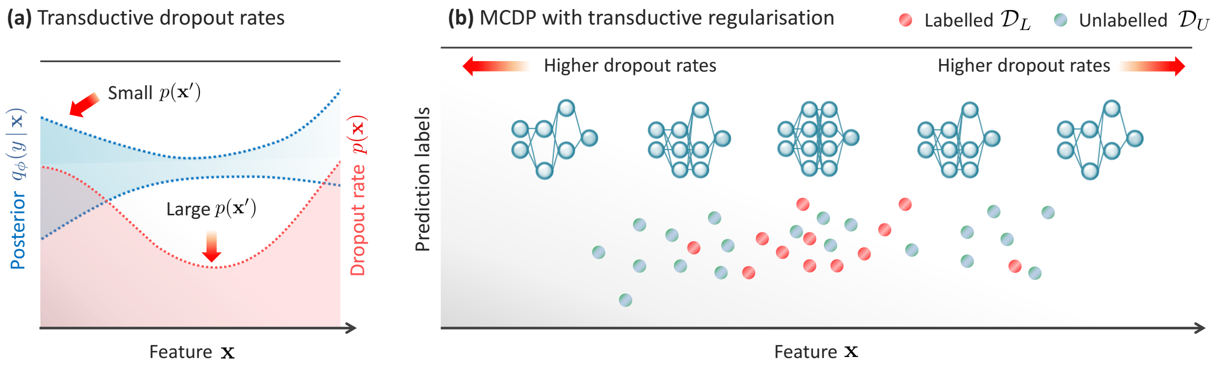

Figure 2. **Pictorial depiction of transductive dropout inference.** (a) Here, we depict an exemplary one-dimensional feature space, along with the corresponding variational posterior $q_\phi(y|\mathbf{x})$ and feature-dependent dropout rate $p(\mathbf{x})$. Transductive dropout inference operates by adapting the dropout rate so that it induces larger posterior variance for regions with dense concentration of unlabelled data, but low density for labelled data (small $p(\mathbf{x}')$ for some $\mathbf{x}' \sim p'(\mathbf{x})$). (b) This panel shows an exemplary realisation of labelled and unlabelled data sets for the same example in panel (a). Red dots are fully observed, whereas for blue ones, we only observe the locations but not the outputs on the y-axis. The typical behaviour of the transductive dropout is to increase the dropout rates in regions where unlabelled data are denser than labelled data, creating more variability in the Monte carlo samples of the network outputs. Here, exemplary instances of test-time dropout applied to the network architecture for different values of the feature $\mathbf{x}$ are depicted.

balanced by the location of the source data set that will lower the variance in our appropriately confident locations. Since the regulariser above solves the *transductive* learning problem of classifying source and target data in a way that resembles semi-supervised learning (Rohrbach et al., 2013), we call Ω a transductive regulariser. In what follows, we propose a practical way to implement transductive regularisation within the MCDP approximate inference framework.

Transductive Dropout. We extend the MCDP approximate inference scheme in (Gal & Ghahramani, 2016) by applying our posterior regularisation penalty, and allowing the dropout rates to vary per data point, dependent on the feature values. By enabling the dropout rates to be a function of $\mathbf{x}$, we provide more degrees-of-freedom to flexibly craft the posterior variance $\mathbb{V}[q_\phi]$ so that it accurately discriminates source and target data points.

Let p be the dropout rate of the underlying NN model. We parameterise p to be dependent on the feature value $\mathbf{x}$ as follows. Let $v_\beta(.)$ be a neural network with a sigmoid output layer and parameters β, i.e., $v_\beta : \mathbb{R}^d \to [0,1]$ maps feature values to dropout rates so that $p = v_\beta(\mathbf{x})$. This equates to an approximate posterior distribution over the NN weights:

$$q_\phi(\mathbf{w}) = \prod_{i=1}^{N} (1 - v_\beta(\mathbf{x}))^{\frac{w_i}{m_i}}\, v_\beta(\mathbf{x})^{\frac{m_i - w_i}{m_i}} \qquad (8)$$

for $w_i \in \{m_i, 0\}$, 0 otherwise, where $\mathbf{w} = \{w_i\}_i$ is the set of weights for the NN modelling the conditional distribution $q_\phi(y|\mathbf{x}, \mathcal{D})$. This leaves an optimisation objective (of the form in (4)) over the variational parameters $\phi = \{\beta, \mathbf{m}\}$. Using the equivalence between KL minimisation and squared loss minimisation under dropout regularisation, we can write the objective function in (4) as

$$R(\phi) = \sum_{\mathbf{x}_i \in \mathcal{D}_L} \|\mathbb{E}[q_\phi(y_i\,|\,\mathbf{x}_i)]\|_2^2 + \Omega(q_\phi(\theta\,|\,\widetilde{\mathcal{D}})), \qquad (9)$$

with the possibility of adding an $L2$ regulariser $\|\phi\|_2^2$ as well. As we can see, this objective incorporates both labelled and unlabelled data: the data set $\mathcal{D}_L$ contributes to the first term, which is concerned with fitting the observed labels drawn from the source distribution, whereas the second term, which depends on the entire augmented data set $\widetilde{D}$, makes sure that the induced variational posterior is aware of the mismatch between source and target feature distributions. We can see that this scheme, as depicted in figure 1, acts in a similar way to (3), primarily optimising the likelihood of the data under the approximation while constrained by a requlariser on the form of the distribution, only now the regulariser induces more specific behaviour and makes use of $\mathcal{D}_U$.

The regulariser in (9) can be computed in backpropagation using sample estimates of the posterior variance as follows. Let $\tilde{\phi}$ be the current estimate of the variational parameters at a given iteration of the gradient descent procedure. To evaluate the model loss and gradients, we use the MCDP forward pass to sample M outputs $\{\hat{y}_k^1, \ldots, \hat{y}_k^M\}$ for every $\mathbf{x}_k$ in $\widetilde{\mathcal{D}}$, and compute a Monte Carlo sample estimate of the transductive regularisation term as follows:

$$\widehat{\Omega}(q_{\tilde{\phi}}(\theta\,|\,\widetilde{\mathcal{D}})) = g\left(\frac{1}{M}\sum_{m=1}^{M}(\hat{y}_k^m - \bar{y}_k)^2\right). \qquad (10)$$

Computations of the estimator in (10) only involve the forward pass, and evaluating its gradients is straightforward.

Key insights Figure 2 provides a pictorial depiction of our transductive dropout inference procedure applied to an exemplary, one-dimensional feature space. A key insight is that transductive dropout inference learns to adapt the dropout rate so that it induces larger posterior uncertainty for regions with dense concentration of unlabelled data, but low density for labelled data.

4.3. Limitations

With the target data included in the training regime, it would seem that this method does not lend itself to online deployment where new test points come in over time. We note though that retraining is not practically necessary every time a new prediction needs to be made, indeed given an initial collection of test points it would only be useful when we encounter a significant amount of new data that is covariate shifted further even than the original targets.

5. Experiments

5.1. Toy Dataset

In this section, we consider a toy 1-d regression example to show how standard BNNs produce overconfident uncertainty estimates under covariate-shift.

The features in the source and target data sets are generated i.i.d. from $p(x) = N(7, 4)$ and $p'(x) = N(11, 4)$ respectively. The labels are generated as $y = f(x) + \epsilon$ where $f(x) = \frac{\sin x}{2} + \frac{x}{4} + \frac{x^2}{100}$ and ϵ is zero-centred Gaussian noise with standard deviation 0.1. The source and target data sets contain 50 data points each.

Figure 3 compares the fit of MCDP (left) and transductive dropout (right). Both networks consist of two fully connected hidden layers with 32 and 64 neurons per respective layer with tanh activations. We see they produce similar mean predictions near the labelled training points. However, MCDP starts to issue over-confident predictions as the feature distribution shifts away from the training data. On the other hand, transductive dropout learns to output larger uncertainty in the areas of low density under $p(x)$. In figure 4 we plot smoothed density estimates of the learnt dropout rates for both source and target distribution points. In the source distribution the learnt rates are all quite tight around 0.18, while in the target distribution there is a much bigger spread reflective of the points' distances from labelled data.

5.2. Prostate Cancer Mortality Prediction

Background Prostate cancer is the third most common cancer in men, with half a million new cases each year around the world (Quinn & Babb, 2002). It is far more common among the elderly with around 75% of cases occur in men aged over 65 years. Therefore, prostate cancer is expected to bring increasing healthcare burden to countries with ageing population (Hsing et al., 2000). The latest clinical guideline for prostate cancer treatment recommends watchful waiting or non-invasive treatment for early-stage patients who have *low mortality rate* (Heidenreich et al., 2011). Surgery (Radical Prostatectomy) is recommended instead for high-risk patients whose health condition deteriorates rapidly. The patient's survival outlook therefore plays an important role in the treatment decisions. Hence, improved accuracy and uncertainty quantification for mortality prediction will help clinicians to design effective treatment plans and improve patients' life expectancy.

Dataset We consider the problem of predicting and estimating the uncertainty of the mortality rate for patients with prostate cancer. Our training data consists of 240,486 patients enrolled in the American SEER program (SEER, 2019), while for our target data we consider a group of 10,086 patients enrolled in the British Prostate Cancer UK program (UK, 2019). For both sets of patients we have identical covariate data with information concerning the age, PSA, and Gleason scores as well as what clinical stage they're at and which, if any, treatment they are receiving. Note that while we have the same features for both sets this is an area where we expect a level of covariate shift given the different programs and the transition from American to British patients. Indeed we do see this, without giving a full break down of the summary statistics, patients in the Prostate Cancer UK are in general older with higher Gleason scores though not as far along in the clinical stages.

Benchmarks We compare our method against competitive methods from the probabilistic deep learning literature based on their prevalence and applicability. While we consider this work quite different to semi-supervised learning, which do not usually consider improving uncertainty estimates, we also include MixMatch as a benchmark (Berthelot et al., 2019). The methods we consider are:

1. *MLP* - Standard feed forward neural network to benchmark accuracy.
2. *Dropout* - Monte Carlo dropout with rate 0.5 (Gal & Ghahramani, 2016; Srivastava et al., 2014)
3. *Concrete Dropout* - Dropout with the rate treated as an additional variational parameter and is optimised with respect to the ELBO (Gal et al., 2017).
4. *Ensemble* - Ensemble of feed forward MLPs (Lakshminarayanan et al., 2017) with $K = 10$ the number of models in the ensemble.
5. *MixMatch* - We implement a version of the MixMatch algorithm (Berthelot et al., 2019) where we perform one round of label guessing and mixup and without sharpening. As the base predictive model we use a MC Dropout network.

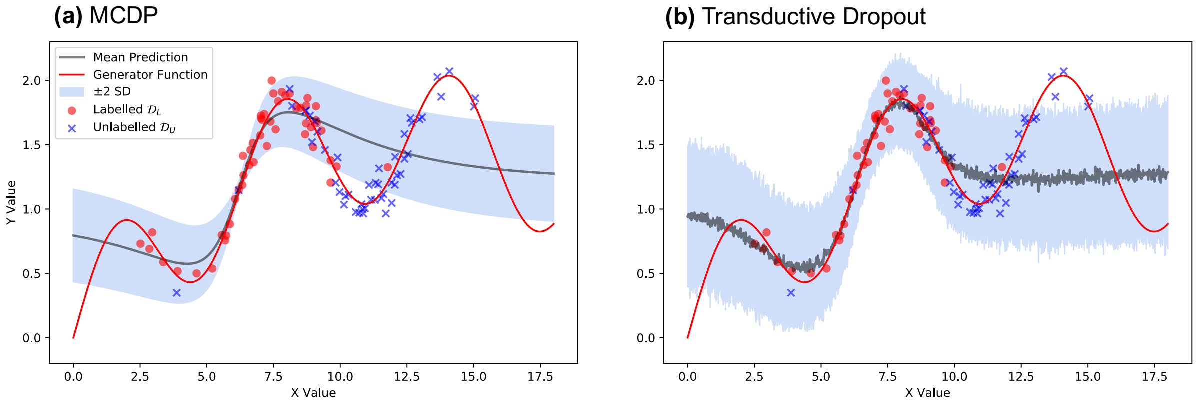

Figure 3. **Comparison of uncertainty predictions** (a) Here we show the 95% confidence intervals for MCDP, demonstrating that while appropriate over the labelled data, they remain overconfident at the unlabelled data. (b) This panel shows the predictions for transductive dropout - the mean prediction remaining equally as accurate while producing more uncertainty at the unlabelled locations

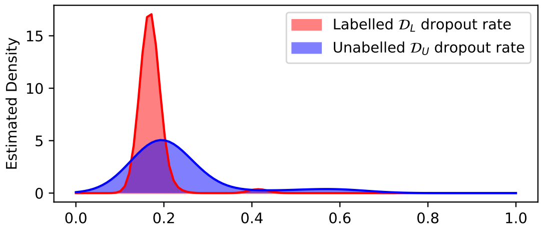

Figure 4. **Dropout rate distribution** Smoothed density estimates of the dropout rate distribution over the source and target sets.

6. *Last Layer Approximations (LL)* - Approximate inference for only the parameters of the last layer of the network (Riquelme et al., 2018), using Dropout.
7. *Transductive Dropout - No Regularisation (TDNR)* - We implement transductive dropout as described above but without the addition of our variance regulariser to show that the gains are not just down to the ability to adapt the dropout rate to the input.

For all of the neural networks we consider the same architecture of two fully connected hidden layers of 128 units each and tanh activation function. The initial weights are randomly drawn from N(0, 0.1) and all networks are trained using Adam (Kingma & Ba, 2015). Hyperparameter optimisation remains an open problem under covariate shift - we used a validation set consisting of 10% of the labelled data selected, not entirely randomly, but based on propensity score matching in order to obtain a set more reflective of the target data. With this, hyperparemeters were selected for all model through grid search.

Evaluation metrics We consider five evaluation metrics for a comprehensive understanding of the model performance. First, we consider the prediction accuracy as measured by AUROC shown as “TEST PERF.” in table 1 (Bewick et al., 2004). Second, we consider the standard deviation of the posterior predictive distribution as a (unnormalised) predictor for whether or not the model will make an error on a given input. The corresponding AUROC score (“ERROR PRED”) measures the agreement between model uncertainty and the chance to predict wrongly, and hence reflects whether the model is well-calibrated. Third, we present the average width of the 95% predictive interval as a measure of general model confidence on unlabelled data (“CI WIDTH”). Next, we show the standard deviation of the predictive distribution on misclassified data (“MISCLASSIFIED SD”). Finally, we show the increased number of patients receiving treatment (INPT) using the associated uncertainty in the model and a risk level of 15%. All quantities related to the posterior distribution are estimated by MC sampling.

Main results First, we note that transductive dropout yields an improvement in the AUROC on the mortality prediction against the other benchmarks, demonstrating that our improved uncertainty calibration does not come at the cost of mean accuracy. Our focus though is on the calibration of our uncertainty estimates. While ultimately it is impossible to properly test how close uncertainty predictions are to what would be the *true* uncertainty, we test by using the posterior predictive variance to classify whether or not the model will make a mistake. The intuition here is that if the model is appropriately uncertain the variance will be high when a mistake is likely and low when not, thus a high performance on using variance as a predictor for when the model will make a mistake should demonstrate appropriate uncertainty estimates. Here we see that transductive dropout significantly outperforms the other benchmarks, suggesting that in general the high variance predictions are indeed associated with those that are more likely to be wrong. We

Table 1. Area under the ROC curve for two tasks, first correctly predicting the mortality rate of patients in the test set and secondly predicting whether for a given patient the model will make an error. We also report the average confidence interval (CI) length over test predictions, the average standard deviation (SD) at miss-classified points, and the increased number of patients receiving treatment (INPT) using the associated uncertainty in the model and a risk level of 15%.

Method	Test Perf.	Error Pred.	CI Width	Misclassified SD	INPT
MLP	0.720 ± 0.012	N/A	N/A	N/A	N/A
MC Dropout	0.729 ± 0.016	0.730 ± 0.016	0.093	0.025	8
Concrete Dropout	0.791 ± 0.012	0.794 ± 0.012	0.151	0.066	76
Ensemble	0.761 ± 0.014	0.782 ± 0.014	0.037	0.018	8
MixMatch	0.728 ± 0.016	0.726 ± 0.016	0.082	0.021	0
LL	0.723 ± 0.014	0.696 ± 0.014	0.073	0.028	22
TDNR	0.836 ± 0.010	0.808 ± 0.011	0.197	0.068	18
Transductive Dropout	0.861 ± 0.009	0.857 ± 0.009	0.130	0.110	189

additionally focus on these predictions that each method gets wrong and look at the average standard deviation at each of these points. Here transductive dropout shows on average it's much less confident about its incorrect predictions than the the other benchmarks, which is the preferred behaviour. It is important to note that this is not at the expense of confidence over all predictions as we show that both concrete dropout and TDNR both have on average larger confidence intervals than transductive dropout.

Impact on patients Given our motivations we also ground the performance of our method in how it could be used in real world decision making on the treatments offered to patients. There are many reasons treatment options may not be offered to patients including cost and potential side effects, as such there will usually be an associated risk level which a patient must be above in order to receive treatment. It's thus very damaging to patients for a model to confidently predict them to be low risk when they are indeed not. In Table 1 we set a 15% threshold, and show how many more patient would receive treatment if we consider coverage of the 95% confidence interval on the patients risk, with the assumption that these cases can be handed off to a human expert who will correctly classify them. We see that transductive dropout results in a large increase in previously patients misclassified as low risk receiving treatment and we develop the impact on treatment options further in Figure 5. Here we set a treatment risk threshold at 50% and show how the size of any predicted confidence intervals over a patients risk impacts the increased number of patients correctly receiving treatment. Naturally for all methods as the confidence interval grows the number of now correctly treated patients increases but transductive dropout consistently outperforms the other benchmarks as it is less often confidently incorrect in its risk prediction.

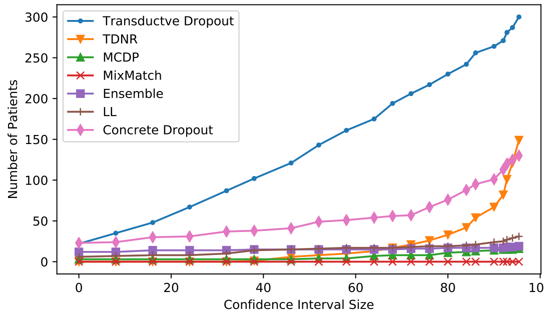

Figure 5. **Improved patient outcome** We show how many more patients, for a risk threshold of 50%, correctly receive treatment as the size of the confidence interval on the prediction of risk changes.

How does the covariate-shift affect uncertainty? Of interest is to consider how the covariate shift has actually impacted our models performance. To that end we consider the feature distribution of those points misclassified by the model to see how it compares to both source and target sets. One of the most important factors affecting both the treatment and survival of prostate cancer patient is the age at diagnosis (Bechis et al., 2011). Studies have shown that older patients tend to have worse survival outlook and are more likely to receive surgery (Hall et al., 2005). In our source data, the average age at diagnosis is 66 years old moving up to 70 in the target set. Comparing to the distribution over ages for incorrectly predicted cases, where the average is 74, we see that it is for the patients who are considerably older than those usually seen in the training data that the model is less sure about. We see a similar story in their PSA scores (measurements of *prostate specific antigen* in the patients blood). PSA score is known to be a highly sensitive indicator for the risk level and severity of prostate cancer, and it is widely adopted in cancer screening and monitoring (Grimm et al., 2012). Again we see an increase in the average from 14.8 to 18.4 from source to target set but for those that are incorrectly classified the mean is much higher at 28.6. The percentage of patients receiving surgery in the incorrectly classified group is twice that of those correctly classified, suggesting that our models are least confident in areas which we might think are the most at risk given domain knowledge - the more elderly with high levels of PSA. The model struggles with them (is much less confident) though as they are values which don't have high density in the training data, demonstrating that blind application of a model to a covariate shifted data set may easily yield surprisingly incorrect predictions. Fortunately transductive dropout tends to return high uncertainty over its predictions on this covariate shifted data such that the practitioner can suitably inform any decisions to be taken as a result of these predictions.

5.3. Additional Results

We focused here on an important real-world example problem, but with the aim of demonstrating generalisation, we provide further benchmark results on some publicly available data sets from the UCI machine learning repository (Dua & Graff, 2017) in appendix A.

6. Conclusions

In this paper we introduced transductive regularisation, a method for using unlabelled data to calibrate the variance of Bayesian neural networks by introducing the auxiliary task of using the posterior predicted variance to discriminate between source and target distributions. We showed that this amounts to performing posterior regularisation in approximate Bayesian inference and results in more useful uncertainty predictions. We examined an instantiation of this framework within MCDP, transductive dropout, and demonstrate its applicability in the real task of predicting prostate cancer mortality, where it outperforms the tested benchmarks and demonstrates a higher level of appropriate uncertainty calibration.

Future Work The question of perfect calibration is clearly not solved and there are several immediate directions for further work that present themselves. First is an extension to frequentist probabilistic ensembles, this is not entirely trivial since the capturing of model uncertainty comes from averaging across elements of the ensemble making estimates during training more complicated to obtain. Second would be an application in the active learning setting and using targeted labelled data acquisition in order to appropriately reduce the model uncertainty - the sub-network used to predict the rate in transductive dropout could inform an important part of an acquisition function.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. Research in this paper was supported by the National Science Foundation (NSF grants 1524417 and 1722516), and the US Office of Naval Research (ONR).

References

Bechis, S. K., Carroll, P. R., and Cooperberg, M. R. Impact of age at diagnosis on prostate cancer treatment and survival. *Journal of Clinical Oncology*, 29(2):235, 2011.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5050–5060, 2019.

Bewick, V., Cheek, L., and Ball, J. Statistics review 13: receiver operating characteristic curves. *Critical care*, 8 (6):508, 2004.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, 2015.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Gal, Y., Hron, J., and Kendall, A. Concrete dropout. In *Advances in Neural Information Processing Systems*, pp. 3581–3590, 2017.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pp. 529–536, 2005.

Graves, A. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pp. 2348–2356, 2011.

Grimm, P., Billiet, I., Bostwick, D., Dicker, A. P., Frank, S., Immerzeel, J., Keyes, M., Kupelian, P., Lee, W. R., Machtens, S., et al. Comparative analysis of prostate-specific antigen free survival outcomes for patients with low, intermediate and high risk prostate cancer treatment by radical therapy. results from the prostate cancer results study group. *BJU international*, 109:22–29, 2012.

Hall, W., Jani, A., Ryu, J., Narayan, S., and Vijayakumar, S. The impact of age and comorbidity on survival outcomes and treatment patterns in prostate cancer. *Prostate Cancer and Prostatic Diseases*, 8(1):22–30, 2005.

Heidenreich, A., Bellmunt, J., Bolla, M., Joniau, S., Mason, M., Matveev, V., Mottet, N., Schmid, H.-P., van der Kwast, T., Wiegel, T., et al. Eau guidelines on prostate cancer. part 1: screening, diagnosis, and treatment of clinically localised disease. *European urology*, 59(1):61–71, 2011.

Hsing, A. W., Tsao, L., and Devesa, S. S. International trends and patterns of prostate cancer incidence and mortality. *International journal of cancer*, 85(1):60–67, 2000.

Jean, N., Xie, S. M., and Ermon, S. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *Advances in Neural Information Processing Systems*, pp. 5322–5333, 2018.

Karbalayghareh, A., Qian, X., and Dougherty, E. R. Optimal bayesian transfer learning. *IEEE Transactions on Signal Processing*, 66(14):3724–3739, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *ICLR*, 2015.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pp. 2575–2583, 2015.

Kouw, W. M. and Loog, M. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, pp. 2, 2013.

Louizos, C. and Welling, M. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2218–2227. JMLR. org, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

Quinn, M. and Babb, P. Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. part i: international comparisons. *BJU international*, 90(2): 162–173, 2002.

Raina, R., Ng, A. Y., and Koller, D. Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 713–720. ACM, 2006.

Riquelme, C., Tucker, G., and Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Rohrbach, M., Ebert, S., and Schiele, B. Transfer learning in a transductive setting. In *Advances in neural information processing systems*, pp. 46–54, 2013.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 1163–1171, 2016.

SEER. Surveillance, epidemiology, and end results (seer) program, 2019. URL `www.seer.cancer.gov`.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Sugiyama, M. and Storkey, A. J. Mixture regression for covariate shift. In *Advances in Neural Information Processing Systems*, pp. 1337–1344, 2007.

Torrey, L. and Shavlik, J. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264. IGI Global, 2010.

UK, P. C. Prostate cancer uk, 2019. URL https://prostatecanceruk.org/.

Zhu, J., Chen, N., and Xing, E. P. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1): 1799–1847, 2014.

Zhu, X. and Goldberg, A. B. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

A. Additional Results

We provide some additional results in table 2 on publicly available data sets taken from the UCI machine learning repository. Specifically we take three data sets: *Breast Cancer*, *Iris*, and *Wine* before slightly adapting them to more naturally fit the covariate shifted setting. First we make them a binary classification problem by taking the class with the largest members as positive and all others as negative. We then split the data into training and testing sets by projecting on to the first principal component and sampling a 20% testing set weighted by this value.

For all of the neural networks we consider the same architecture of two fully connected hidden layers of 32 and 64 hidden units each with tanh activation function. The initial weights are randomly drawn from N(0, 0.1) and all networks are trained using Adam. We consider the prediction accuracy as measured by AUROC shown as “TEST PERF.” as well as the standard deviation of the posterior predictive distribution as a (unnormalised) predictor for whether or not the model will make an error on a given input. The corresponding AUROC score (“ERROR PRED”) measures the agreement between model uncertainty and the chance to predict wrongly, and hence reflects whether the model is well-calibrated.

We see tranductive dropout always performs strongly on test performance, and though not always the best is certainly competitive in all cases, demonstrating their doesn’t appear to be a toll on mean predicitive power. Further though we see that transductive dropout does remain the best across the data sets on the task of error prediction, demonstrating better uncertainty calibration, the focus of this work.

Table 2. For the three datasets we present the area under the ROC curve for two tasks, first correctly predicting the classification in the test set and secondly predicting whether for a given test point the model will make an error.

METHOD	BREAST CANCER		IRIS		WINE	
	TEST PERF.	ERROR PRED.	TEST PERF.	ERROR PRED.	TEST PERF.	ERROR PRED.
MC DROPOUT	0.979 ± 0.012	0.662 ± 0.033	0.937 ± 0.044	0.063 ± 0.046	0.972 ± 0.026	0.775 ± 0.155
CONCRETE DROPOUT	0.791 ± 0.006	0.794 ± 0.006	0.952 ± 0.038	0.847 ± 0.055	1.000 ± 0.000	0.915 ± 0.050
ENSEMBLE	0.978 ± 0.011	0.675 ± 0.007	0.960 ± 0.041	0.571 ± 0.115	0.993 ± 0.007	0.939 ± 0.037
MIXMATCH	0.986 ± 0.010	0.529 ± 0.046	0.242 ± 0.069	0.758 ± 0.069	0.889 ± 0.050	0.611 ± 0.105
LL	0.950 ± 0.033	0.329 ± 0.032	0.929 ± 0.064	0.071 ± 0.064	0.986 ± 0.013	0.575 ± 0.230
TDNR	0.979 ± 0.013	0.945 ± 0.026	0.940 ± 0.045	0.657 ± 0.105	1.000 ± 0.000	0.890 ± 0.055
TRANSDUCTIVE DROPOUT	0.968 ± 0.017	0.975 ± 0.015	0.956 ± 0.045	0.877 ± 0.082	1.000 ± 0.000	0.951 ± 0.034