arXiv:2202.04796v1 [econ.TH] 10 Feb 2022

The Transfer Performance of Economic Models [*]

Isaiah Andrews[†] Drew Fudenberg[‡] Annie Liang[§] Chaofeng Wu[¶]

February 11, 2022

Abstract

Whether a model's performance on a given domain can be extrapolated to other settings depends on whether it has learned generalizable structure. We formulate this as the problem of *theory transfer*, and provide a tractable way to measure a theory's transferability. We derive confidence intervals for transferability that ensure coverage in finite samples, and apply our approach to evaluate the transferability of predictions of certainty equivalents across different subject pools. We find that models motivated by economic theory perform more reliably than black-box machine learning methods at this transfer prediction task.

1 Introduction

When economists estimate models on data, they often hope that the estimated model will be useful for making predictions in settings beyond the narrow setting from which the data were drawn. For example, a pricing model estimated on purchase data from one population of consumers may be used to predict demand in new populations with different demographics

[*] We thank Andrei Iakovlev for helpful comments on this draft. We also thank NSF grants 1851629 and 1951056 for financial support, and the Quest high performance computing facility at Northwestern University for computational resources and staff assistance.

[†] Department of Economics, Harvard

[‡] Department of Economics, MIT

[§] Department of Economics and Department of Computer Science, Northwestern University

[¶] Department of Computer Science, Northwestern University

and tastes. Whether a model's predictive performance on the original data is a good indication of how it will perform on data from new settings depends on whether the estimated model captures generalizable structure.

Understanding how well a model transfers is especially important now that black box methods have found increased popularity within economics. One reason some economists prefer structured economic models is the belief that such models are more likely to capture regularities that are fundamental and apply in a wide variety of settings, but so far there has been little empirical validation of that belief. It is an open question whether economic models indeed transfer to other data sets better than black boxes do.

Our contributions in this paper are to provide a tractable framework for the transfer prediction problem, to define measures for model transferability, and to provide confidence intervals for transferability that (i) ensure coverage in finite samples and (ii) can be used to evaluate economic models and black boxes alike. We apply our proposed framework and methods to evaluate the transferability of models of risk preferences, and demonstrate that in this application the predictions of economic models perform transfer substantially more reliably than those of two popular black box algorithms.

Our conceptual framework is an extension of the usual "out-of-sample" evaluation problem to "out-of-domain" evaluation. In the standard out-of-sample approach, a model's free parameters are estimated on training data, and the estimated model is evaluated for prediction of new test data, where the training and testing data are drawn from the same distribution. We depart from this framework by supposing that the distribution governing the data varies over a set of domains. The analyst has access to a meta dataset consisting of samples from some of these domains. We assume that the domains are drawn iid from a population of candidate domains. While the iid assumption rules out some interesting transfer prediction problems,[1] some structure on the relationship across domains is necessary to draw conclusions about transferability. We view the assumption that the observed domains represent an iid sample as a useful first step which yields easy-to-apply procedures.

We use this framework to formalize several questions central to the problem of model transferability, and propose measures for each question (see Section 3.2). First, we ask how

[1] Examples include time-series problems, where the domains correspond to different time-periods and are statistically dependent, and cross-sectional settings where observed domains are unrepresentative (for instance due to site-selection bias as discussed in Allcott 2015).

well the model will predict on an unseen target domain. We call this the model's *transfer error*. Second, we ask how many times larger this transfer error than the best achievable error on the target domain. We call this the model's *normalized transfer error*. Finally, we ask how much is lost by transferring a model across domains instead of re-estimating the model's parameters on the sample to be predicted. We call this the model's *transfer deterioration*. The first two measures can be used to help us select between models for making predictions in new domains, and the final measure is useful for understanding the value of re-estimating the parameters of a given model on the new data.

In Section 4, we demonstrate how to construct confidence intervals for each of these measures. Given the meta dataset of samples across different domains, we select a subset of domains to use as training domains, and estimate the parameters of the model on this data. We then evaluate the measure of interest, for example the transfer error, of this estimated model on each of the remaining samples in the metadata. Since by assumption all samples are iid, the ranks of the transfer errors are uniformly distributed up to ties. We use this observation to bound the probability that the transfer error on the target domain falls between the τ-th and $(1-\tau)$-th percentile of the observed transfer errors (for any τ).

The above approach is effective for evaluating the transfer error of a model trained on a specific selected set of training domains. But since the samples in the meta dataset are ex-ante symmetric, there is no reason to prefer any choice of samples to use for training over another. We thus develop confidence intervals for transfer error when a model is trained on a random set of training domains. We show that the τ-th and $(1-\tau)$-th percentile transfer errors (across realizations of the training and test domains) can be used to form upper and lower bounds for a confidence interval for the transfer error. The same approach yields confidence intervals for our other two measures as well.

In Section 5, we evaluate the transferability of predictions of certainty equivalents for binary lotteries, where the domains correspond to different subject pools. We evaluate the transferability of a well-known model of risk preferences, Cumulative Prospect Theory, as well as the transferability of two popular black box algorithms.

Our main finding, which is robust across all three of our measures, is that the while the confidence intervals for the black box algorithms and CPT model overlap, the confidence intervals for the black box methods are substantially wider, and their upper bounds are substantially higher. For example, our 71% confidence interval for the normalized transfer error for CPT ranges from 1.05 (i.e., a transfer error that is 1.05 times larger than the best possible error) to 1.88, while the same confidence interval for a neural net algorithm ranges from 1.01 to 13.85. Further supporting these findings, we find that a 71% confidence interval for the ratio between the black box transfer error and the CPT error ranges from 0.8 to 5. Taken together, these results suggest that CPT generalizes more reliably to new domains than do the black box algorithms.

While the transfer errors of these prediction methods differ substantially, the within domain performance of CPT and the black box methods turn out to be very similar. That is, if we train these models on data from one domain and test it on data from that domain, all of the approaches perform similarly well. This suggests that even when CPT and the black box models learn representations that are close for the purposes of prediction in a specific domain, these representations often have different implications in new domains, and the representation learned by CPT generalizes better.

2 Related Literature

Our paper contributes to a recent literature regarding the role of black box algorithms for predicting economic behaviors. Several papers demonstrate that we can use black box algorithms to make predictions that improve upon those of the best existing models (Noti et al., 2016; Plonsky et al., 2017, 2019; Ke et al., 2020; Camerer et al., 2019; Fudenberg and Liang, 2019). While black box methods can be effective when the analyst has a large quantity of data from the setting of interest, our results suggest that black box methods may be less effective for transfer prediction.

Our paper is related to the literature on meta-analyses, which seeks to draw reliable conclusions from disparate empirical studies (see e.g., Benartzi et al. (2017), DellaVigna and Pope (2019), Hummel and Maedche (2019), and Meager (2019)). A common finding in meta-analyses is that estimates across studies differ more than expected due solely to chance, suggesting that the effects being estimated are heterogeneous.[2] Discussions of heterogeneity are sometimes framed in terms of external validity, since a high degree of heterogeneity implies that the results obtained in any particular study are unlikely to be valid in other

[2] See e.g. Naaktgeboren et al. (2016) for a discussion of variability measures used in medical meta-analyses.

settings.[3] While the primary focus in this literature is on coefficient variability rather than transfer performance as we define it, the two are related and a high degree of variability in the parameters of the best-fitting model across domains suggests that transfer performance is likely to be poor.

Our paper is also connected to several areas of computer science. The "transfer prediction problem" that we formulate in Section 3.1 is closely related to *domain generalization* (as introduced in Blanchard et al. (2011) and Muandet et al. (2013)), namely the problem of learning a model that generalizes well to new unseen domains (see Zhou et al. (2021) for a recent survey).[4] Our work differs from this literature in two primary ways. First, while the domain generalization literature develops algorithms for achieving good out of domain generalization in practice, we develop confidence intervals for the out of domain generalization error of a fixed learning procedure. Second, a key part of our motivation is to compare the transferability of atheoretical machine learning models, which look for regularities in the data in a way that doesn't depend on the prediction task considered, to models grounded in economic theory, which are chosen based on the task considered (such as models of risk preferences for predicting certainty equivalents). In contrast, the domain generalization literature has focused almost exclusively on evaluating and improving neural network architectures.[5]

Our paper also connects to the large literature on measures for the "flexibility" or "complexity" of a model class, such as VC dimension and Rademacher complexity, or the restrictiveness measure proposed in Fudenberg et al. (2021). Specifically, one might conjecture that models that are more flexible run a greater risk of overfitting to the domain, and thus have a worse transfer performance. In Appendix C we adapt the well-known Rademacher complexity measure to our transfer setting and develop analogs of standard results, demonstrating that transfer error can be bounded using the (transfer) Rademacher complexity of the model. These bounds do not, however, allow us to rank the expected transfer errors of

[3] Meager (2019) discusses this heterogeneity in the effect of microcredit expansions.

[4] Our problem corresponds to *homogeneous* domain generalization, where the set of outcomes $\mathcal{Y}$ is constant across domains, in contrast to *heterogenous* domain generalization, where the outcome set potentially varies across domains as well.

[5] Zhou et al. (2021) highlights improvements in model architecture as an important avenue for future research: "...most existing [domain generalization (DG)] methods have been focused on the learning part...while paying less attention to designing effective model architectures for DG."

models based on their (transfer) Rademacher complexity. Whether a more complex model transfers more poorly depends on details of the application. In the specific application we consider, we do find that the black box algorithms that we consider are both more flexible (in the sense of higher transfer Rademacher complexity and lower restrictiveness) and are also less reliable for transfer prediction than the more structured economic models.

Finally, the finite-sample validity of our confidence sets rests on exchangeability arguments similar to the recent literature on conformal inference (see e.g. Lei et al., 2018; Chernozhukov et al., 2021), while the arguments we use to eliminate the random choice of test domain builds on Rüschendorf (1982) and Meng (1994).

3 Framework

3.1 Setup

In the standard statistical learning framework, there is a set of observable features $\mathcal{X}$, an outcome set $\mathcal{Y}$, and a fixed distribution $\mathcal{P}$ on $\mathcal{Z} \equiv \mathcal{X} \times \mathcal{Y}$. An analyst has an action set $\mathcal{A}$ and a loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}_+$, which the analyst would like to minimize. The analyst chooses a strategy $\sigma : \mathcal{X} \to \mathcal{A}$ that maps observed features into actions, where the *expected error* (also known as the risk) of this strategy is

$$e_P(\sigma) = \mathbb{E}_{(x,y)\sim P}\left[\ell(\sigma(x), y)\right].$$

The *empirical error* of the strategy σ on a sample S of observations from $\mathcal{Z}$ is

$$e(\sigma, S) = \frac{1}{|S|} \sum_{(x,y)\in S} \ell(\sigma(x), y).$$

We use Σ to denote the set of all possible strategies.

We maintain the definitions above, but depart from the standard framework by introducing domains indexed by $d \in \mathbb{N}$, where the probability distribution on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ varies across domains. We suppose that the analyst has access to samples from a set of n economic domains, which we index by $[n] \equiv \{1, \ldots, n\}$. The distribution of samples corresponds to first drawing a distribution $\mathcal{P}_d$ and sample size m_d for each domain $d \in [n]$ from an unknown distribution $\mu \in \Delta(\Delta(\mathcal{Z}) \times \mathbb{N})$, and then drawing m_d observations independently from $\mathcal{P}_d$. (Throughout, we use $\Delta(X)$ to mean the set of distributions on the set X.) In an abuse of notation, we will simply write $S \sim \mu$ for a sample generated in this way. The analyst's *metadata*

$$\mathbf{M} = (S_1, \ldots, S_n)$$

is a vector of n samples drawn independently from μ.[6] We use $\mathcal{M}$ to denote the space of all metadata of finite length.

The analyst chooses a *decision rule* $\rho : \mathcal{M} \to \Delta(\Sigma)$ that maps metadata to (potentially randomized) strategies σ. As we discuss below, in some cases it may be useful for the analyst to pick the strategy based on a subset of the metadata. We thus use $\mathbf{M}_T \subseteq \mathbf{M}$, for $T \subseteq [n]$, to denote the *training data*, with $\rho(\mathbf{M}_T)$ the chosen strategy.

Here are two examples for decision rules ρ:

Example 1 (Empirical Risk Minimization). The analyst restricts to a subset of strategies $\Sigma^* \subseteq \Sigma$, where Σ^* is compact in some topology such that $e(\sigma, S)$ is continuous in σ for all S. (In our subsequent application, we will take Σ^* to be those strategies that are justified by a specific economic model.) The training data $\mathbf{M}_T$ is mapped to the strategy in Σ^* that maximizes payoffs on the observations in $\mathbf{M}_T$. This corresponds to "in-sample" minimization of error. To define this formally, for any set of strategies Σ^*, let $\rho_{\Sigma^*} : \mathcal{M} \to \Sigma^*$ to be a decision rule satisfying $\rho_{\Sigma^*}(\mathbf{M}_T) \in \operatorname{argmin}_{\sigma \in \Sigma^*} \frac{1}{n} \sum_{i=1}^n e(\sigma, S_i)$. To save on notation, we define $\sigma^*_{\mathbf{M}_T} \equiv \rho_{\Sigma^*}(\mathbf{M}_T)$ for the strategy induced by decision rule ρ_{Σ^*} and training data $\mathbf{M}_T$.

Example 2 (Black Box Algorithm). The analyst chooses a machine learning algorithm, which takes the training data $\mathbf{M}_T$ as input, and outputs a (potentially randomized) strategy $\sigma : \mathcal{X} \to \mathcal{A}$ obeying a particular architecture (for example, a neural network architecture or an ensemble of decision trees).

3.2 Questions and Proposed Measures

We formalize several questions related to how well a model transfers across domains, and propose measures for each of them.

Transfer Error. Suppose an analyst needs to choose a strategy σ to use in a new domain, but has no prior data from this domain to guide their choice. Instead, the analyst has

[6] The domains are distinguished in this vector. So, for example, the metadata corresponding to observation of z_1 from domain 1 and (z_2, z_3) from domain 2 is $\{(z_1), (z_2, z_3)\}$ as opposed to (z_1, z_2, z_3).

metadata $\mathbf{M}$ consisting of samples from other domains, and uses the decision rule ρ to choose a strategy based on some subset of this metadata, $\mathbf{M}_T$. (In the case of empirical risk minimization, the analyst searches within the restricted class of strategies Σ^* and finds the strategy $\sigma^*_{\mathbf{M}_T}$ that performs best on the training data $\mathbf{M}_T$.) Since the strategy will be used on a new domain, which is governed by a probability distribution different from those governing the samples in $\mathbf{M}_T$, the out-of-domain prediction error of the strategy $\rho(\mathbf{M}_T)$ could be very different from its in-domain error on the training data. The analyst would like to know the error of this strategy on the new domain.

Definition 1. *The* transfer error *of strategy $\rho(\mathbf{M}_T)$ to test sample S is*

$$e(\rho(\mathbf{M}_T), S).$$

One natural use of this measure is for choosing which decision rule to use in a new domain. As we will see in our subsequent application, it is possible for one decision rule to achieve a lower error than another on the training data $\mathbf{M}_T$, but to result in substantially worse errors out-of-domain.

Completeness of Transfer Prediction. The raw transfer error introduced in Definition 1 can be difficult to interpret, since it captures several sources of error. First, the strategy learned from the training data may be a bad strategy for the test domain. This is often the source of error we are most interested in. But additionally, it may be that the sample contains substantial "irreducible noise," so that even the best strategy for minimizing error on that sample does not achieve perfect prediction. In this case, the raw transfer error may be large not because the strategy $\rho(\mathbf{M}_T)$ is a poor choice, but because the sample S is fundamentally unpredictable (given the available features $\mathcal{X}$).[7]

We thus propose a normalization of Definition 1, where we first specify a set of decision rules (including, potentially, various black box methods), and normalize the transfer error by the smallest in-sample error of these rules. We interpret the denominator as a proxy for the best achievable error in the test domain.[8]

[7] Also, "small" errors can be amplified through seemingly inconsequential re-scalings of units. For example, in our application to predicting certainty equivalents for lotteries, measuring payoffs in smaller units (e.g. cents rather than dollars) mechanically results in larger raw transfer errors.

[8] This measure is analogous to the "completeness" measure introduced in Fudenberg et al. (2022), without the use of a baseline model to set a maximal reasonable error, and adapted for the transfer setting by training and testing on samples drawn from different domains.

Definition 2. *Fix a finite set of decision rules R with $\rho \in R$. The R-*normalized transfer error *of strategy $\rho(\mathbf{M}_T)$ to test sample S is*

$$\frac{e(\rho(\mathbf{M}_T), S)}{\min_{\widetilde{\rho} \in R} e(\widetilde{\rho}(S), S)}.$$

This tells us how many times larger the transfer error of the strategy $\rho(\mathbf{M}_T)$ is than the best in-sample error achievable by a decision rule from R. In our subsequent application, we choose R to include decision rules derived from various specifications of an economic model, as well as decision rules derived from two popular black box methods. This ratio is bounded below by 1, and a model achieves this lower bound only if the transfer error is as low as the best in-sample error over R.

Since the denominator is not model-dependent, the within-domain ranking of models using Definitions 1 and 2 is the same. When we look across domains, however, the change in normalization can matter for comparisons of transfer error. The unit-free nature of the normalized error is thus particularly helpful when making comparisons across domains with different units of measurement.

Value of Domain Data. It is also useful to compare the performance of a decision rule across different choices of training data. An analyst interested in making predictions in a given domain would prefer to train their model on a sample from that domain, but such data may be costly to acquire, so if the existing data $\mathbf{M}_T$ can serve as a good proxy, the analyst may prefer to transfer model parameters estimated from other domains. We propose next a measure for the "transfer deterioration" of a model, which captures how much we lose by transferring a given model across domains, rather than re-training it on the domain of interest.

Definition 3. *The* transfer deterioration *of strategy $\rho(\mathbf{M}_T)$ to test sample S is*

$$\frac{e(\rho(\mathbf{M}_T), S)}{e(\rho(S), S)}.$$

This ratio tells us how many times larger the transfer error of $\rho(\mathbf{M}_T)$ is than the in-sample error of strategy $\rho(S)$. Larger quantities suggest that the model is less transferrable across domains. This ratio is bounded below by 1, and a model achieves this lower bound only if the transfer error of the model is the same as its in-sample error.[9] This does not require $\mathbf{M}_T$ and S to be identical: it is sufficient for the training data $\mathbf{M}_T$ and the sample S to lead to the same estimates for model parameters. The larger the transfer deterioration of the decision rule, the more valuable it is to re-train the model on the new data rather than transferring parameters estimated from other domains. (Note that as with Definition 2 we are "stacking the deck" against transfer prediction by comparing using an in-sample error in the denominator.)

A ranking of models by transfer deterioration need not coincide with a ranking of models using either Definitions 1 or 2. For example, a model that achieves approximately constant but large errors across domains would have low transfer deterioration but high transfer error. We thus view this measure as useful primarily for determining the value of re-estimating a given model's parameters, rather than for deciding between models.

4 Approach and Main Results

Our approach for inference on the quantities defined above is as follows. Given the metadata $\mathbf{M} = (S_1, \ldots, S_n)$, we select a subset of samples indexed to $T \subset [n]$ to be training domains. Let $n_T \equiv |T|$ denote the number of training domains, and $\mathbf{M}_T = (S_{d'})_{d' \in T}$ denote the *training data*, which consists of the samples from domains in T. The remaining domains $[n] \backslash T$ are used as test domains, and we use S_{n+1} to denote an (unobserved) target sample where we want to predict performance. For each domain $d \in \{1, \ldots, n+1\}$, define

$$e_{T,d} \equiv e(\rho(\mathbf{M}_T), S_d)$$

to be the (random) transfer error of strategy $\rho(\mathbf{M}_T)$ on the sample S_d.

It will be useful to define $e_{T,(r)}$ to be the r-th largest value in $\{e_{T,d} : d \in [n] \setminus T\}$. Since by assumption the samples $S_1, \ldots, S_{n+1}$ are independently and identically distributed according to μ, the ranks of the transfer errors from T to the domains $d \in \{[n+1] \setminus T\}$ (i.e., the test domains and the target domain) are uniformly distributed, up to ties. Thus we have:

[9] Note that we are comparing an out-of-sample object (in the numerator) to an in-sample object (in the denominator) and so "stack the deck" against the strategy $\rho(\mathbf{M}_T)$. An alternative measure in the same spirit would divide through by a cross-validated error, where the training and test samples drawn from the same domain. Note that in this case transfer deterioration need not admit an interpretable lower bound.

Claim 1 (One-Sided Confidence Interval)**.** *For any quantile $\tau \in (0,1)$, let $r \equiv \lceil \tau(n-n_T) \rceil$. Then*

$$\mathbb{P}\left(\left\{e_{T,n+1} > e_{T,(r)}\right\}\right) \leq 1 - \frac{r}{n-n_T+1}.$$

This implies that $(-\infty, e_{T,(r)}]$ is a level-$\left(\frac{r}{n-n_T+1}\right)$ confidence interval for the transfer error of $\rho(\mathbf{M}_T)$ on the target sample S_{n+1}. By similar reasoning, we obtain:

Claim 2 (Two-Sided Confidence Interval)**.** *For any quantile $\tau \in (0.5,1)$, let $r_L \equiv \lfloor (1-\tau)(n-n_T) \rfloor$ and $r_U \equiv \lceil \tau(n-n_T) \rceil$. Then*

$$\mathbb{P}\left(\left\{e_{T,n+1} \notin \left[e_{T,(r_L)}, e_{T,(r_U)}\right]\right\}\right) \leq 1 - \frac{r_U - r_L}{n-n_T+1}.$$

This implies that $\left[e_{T,(r_L)}, e_{T,(r_U)}\right]$ is a level-$\left(\frac{r_U-r_L}{n-n_T+1}\right)$ confidence set for the transfer error of $\rho(\mathbf{M}_T)$ on the target sample S_{n+1}.

Claims 1 and 2 provide confidence intervals for the performance of the particular strategy $\rho(\mathbf{M}_T)$, where the set of training domains T is fixed. But since the training domains are ex-ante symmetric, there is no reason to prefer any choice of training domains over another. We may thus wish to move from the evaluation of the transfer error of a particular $\rho(\mathbf{M}_T)$ to the evaluation of the transfer error of the decision rule ρ trained on a random set of training domains.

To that end, let $\phi(k, A)$ be the distribution which draws k elements (without replacement) from the set A. We draw the training domains T from this distribution, with $A = [n]$, where each realization of the training data T implies a choice of $n - n_T$ test domains. Below, we pool the transfer errors (across choices of training and test domains) and use the τ-th and $(1-\tau)$-th percentiles of the pooled transfer errors to form a confidence interval for the transfer error given a randomly selected set of training domains.

For a random variable B let

$$\overline{Q}_\tau(B) = \sup\{b : Pr\{B \geq b\} \geq 1-\tau\}$$

and

$$\underline{Q}_\tau(B) = \inf\{b : Pr\{B \leq b\} \geq 1-\tau\}$$

denote the upper and lower τth quantiles, respectively. The upper and lower quantiles coincide for continuously distributed variables.

Proposition 1 (Confidence Intervals). *For any $\tau \in (0,1)$, define*

$$\overline{e}_\tau \equiv \overline{Q}_{\tau,(\widetilde{T},t)\sim\phi(n_T+1,[n])}\left(e_{\widetilde{T},t}\right)$$

as the τth upper quantile of the out-of sample error $e_{\widetilde{T},t}$ when $\widetilde{T}$ consists of n_T domains drawn uniformly at random from $[n]$, and t is drawn uniformly at random from $[n]\setminus\tilde{T}$. Then

$$\mathbb{P}\left(\left\{e_{\widetilde{T},n+1} > \overline{e}_\tau\right\}\right) \leq 2\left(1-\frac{n-n_T}{n-n_T+1}\tau\right).$$

If we further define

$$\underline{e}_\tau \equiv \underline{Q}_{(1-\tau),(\widetilde{T},t)\sim\phi(n_T+1,[n])}\left(e_{\widetilde{T},t}\right),$$

then

$$\mathbb{P}\left(\left\{e_{\widetilde{T},n+1} \notin [\underline{e}_\tau,\overline{e}_\tau]\right\}\right) \leq 4\left(1-\frac{n-n_T}{n-n_T+1}\tau\right)$$

Thus, $[\underline{e}_\tau,\overline{e}_\tau]$ is a level-$\left(4\left(\frac{n-n_T}{n-n_T+1}\tau\right)-3\right)$ confidence interval for the transfer error on the target sample.

In Appendix A.2, we show that this result can be used to construct confidence intervals for transfer errors under "partial transfer," where some parameters of the model are estimated on the metadata $\mathbf{M}_T$ but other parameters are re-estimated on the target sample S_{n+1}. All of the results in this section hold by identical arguments if we instead choose a set of decision rules R and define

$$e_{T,d} \equiv \frac{e(\rho(\mathbf{M}_T), S_d)}{\min_{\rho\in R} e(\rho(S_d), S_d)}$$

to be the R-normalized transfer error of strategy $\rho(\mathbf{M}_T)$ on the target sample S_d, or define

$$e_{T,d} \equiv \frac{e(\rho(\mathbf{M}_T), S_d)}{e(\rho(S_d), S_d)}$$

to be the transfer deterioration of strategy $\rho(\mathbf{M}_T)$ on the target sample S_d. Thus we can construct confidence sets for normalized transfer error and transfer deterioration in exactly the same way.

5 Application

As an illustration of our methodology, we evaluate the transferability of predictions of certainty equivalents for binary lotteries, where the domains correspond to different subject pools. To this end, we have constructed a meta dataset consisting of samples from 44 domains (see Section 5.1). We expect that models of risk preferences trained on data from certain subject pools will be predictive of data elicited from other subject pools, but imperfectly so. Our measures shed light on how reliable these models are for out-of-domain prediction, and which approaches for transfer prediction we should prefer.

We consider prediction methods based on economic models (different specifications of Cumulative Prospect Theory), and prediction methods based on flexible black box algorithms. In principle, given sufficient data from any domain, black box methods can flexibly learn the (potentially complex) structure in that domain, and thus should outperform economic models for predicting new data drawn from that domain. In Section 5.3, we show that the errors of the black box methods are lower than those of the CPT models when trained and tested on data from the same domain.

The black box methods may achieve better within-domain fit by learning idiosyncratic details of the domain, which do not transfer across settings. The question of whether black box methods learn additional structure that is general, or additional structure that is idiosyncratic, is critical to the decision of whether we should prefer black box methods or economic models for transfer prediction. It is also an empirical question: while we prove some results in Appendix C relating the complexity of a model class to its transfer predictiveness, in general the comparison of black box methods and economic models will depend on details of how the domains differ from one another. Using our approach, we are able to comment on this comparison for the problem of predicting certainty equivalents.

In Section 5.4, we use our results to construct confidence intervals for transfer error, normalized transfer error, and transfer deterioration for all of the prediction methods. We find that the confidence intervals for black box methods are substantially wider and higher relative to the confidence intervals for the economic models. In Section 5.5, we conduct a more detailed comparison between CPT and the random forest algorithm. We find that the random forest algorithm actually outperforms the economic model on a majority of test cases, but CPT rarely performs much worse than the random forest algorithm, and sometimes performs much better. These findings collectively suggest that the economic models do a better job of learning structure that more reliably transfers across domains.

5.1 Data

We have constructed a meta dataset including samples from 44 economic domains. This data is drawn from 14 papers, with one paper (a meta-study of risk preferences across countries) contributing samples from 30 domains. We chose to label economic domains based on subject pools, but allow for experimental treatments to vary across problems within domains.[10] Our samples range in size from 72 observations to 8906 observations, with an average of 2752.7 observations per sample. See Appendix B.1 for a more detailed description of our data sources.

Within each sample, observations take the form $(\overline{z}, \underline{z}, p; y)$, where $\overline{z}$ and $\underline{z}$ denote the possible prizes of the lottery, p is the probability of the larger prize, and y is the reported certainty equivalent by a given subject. The analyst's action a is a prediction of the reported certainty equivalent given the description of the lottery $(\overline{z}, \underline{z}, p)$. We consider squared-error loss

$$\ell(a, y) = (a - y)^2,$$

but for ease of interpretation, we report results in terms of root mean squared error, which puts the errors in the same units as the prizes. [11] Since different subjects report different certainty equivalents for the same lottery, the best achievable error on a sample is generally bounded away from zero.

5.2 Models and Decision Rules

We suppose that the analyst uses empirical risk minimization (see Example 1 in Section 3.1), and consider several possible choices for the restricted set of strategies Σ^*.

Cumulative Prospect Theory. First we consider the set of strategies Σ^* derived from the model Cumulative Prospect Theory (CPT). Fixing values for the model's parameters $(\alpha, \beta, \gamma, \zeta)$, each lottery $(\overline{z}, \underline{z}, p)$ is assigned a utility

$$w(p)v(\overline{z}) + (1 - w(p))v(\underline{z}) \tag{1}$$

[10] For example, in Etchart-Vincent and l'Haridon (2011), we pool reported certainty equivalents across three payment conditions: a real-loss condition, a hypothetical-loss condition, and a loss-from-initial-endowment condition.

[11] This transformation is possible because none of the results in this paper change if we redefine $e_P(\sigma) = g(\mathbb{E}_{(x,y)\sim P}\left[\ell(\sigma(x), y)\right])$ and $e(\sigma, S) = g(\frac{1}{|S|}\sum_{(x,y)\in S} \ell(\sigma(x), y)$ for any function g.

where

$$v(z) = \begin{cases} z^{\alpha} & \text{if } z \geq 0 \\ -(-z)^{\beta} & \text{if } z < 0 \end{cases} \tag{2}$$

is a value function for money, and

$$w(p) = \frac{\zeta p^{\gamma}}{\zeta p^{\gamma} + (1-p)^{\gamma}} \tag{3}$$

is a probability weighting function.

For each $\alpha, \beta, \gamma, \delta$, the strategy $\sigma_{(\alpha,\beta,\gamma,\delta)}$ is defined to satisfy

$$\sigma_{(\alpha,\beta,\gamma,\delta)}(\overline{z}, \underline{z}, p) = v^{-1}\big(w(p)v(\overline{z}) + (1-w(p))v(\underline{z})\big).$$

That is, the strategy maps each lottery into the predicted certainty equivalent under CPT with parameters $(\alpha, \beta, \gamma, \delta)$. Following the literature, we impose the restriction that the parameters belong to the set $\Theta = \{(\alpha, \beta, \gamma, \delta) : \alpha, \beta, \gamma \in [0,1], \delta \geq 0\}$ and define the set of CPT strategies to be $\Sigma^* \equiv \{\sigma_\theta\}_{\theta \in \Theta}$.

Besides this choice of Σ^*, we also evaluate strategies corresponding to restricted specifications of CPT that have appeared elsewhere in the literature: CPT with free parameters α and β (setting $\delta = \gamma = 1$) describes an expected utility decision-maker whose utility function is as given in (2); CPT with free parameters α, β and γ (setting $\delta = 1$) is the specification used in Karmarkar (1978); and CPT with free parameters δ and γ (setting $\alpha = \beta = 1$) describes a risk-neutral CPT agent whose utility function over money is $u(z) = z$ but exhibits nonlinear probability weighting.[12] Additionally, we include CPT with the single free parameter γ (setting $\alpha = \beta = \delta = 1$), which Fudenberg et al. (2021) finds to be an effective one-parameter specification.

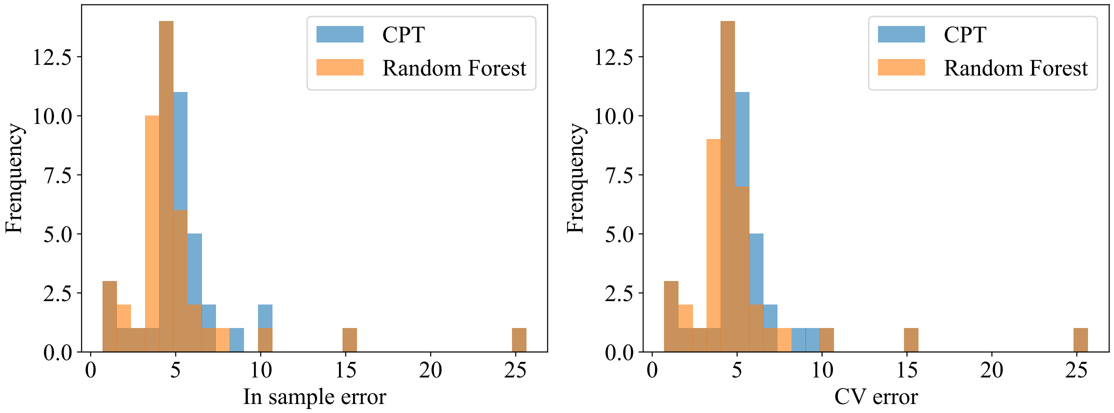
Black Boxes. We also consider decision rules ρ corresponding to two machine learning algorithms. First, we estimate a *random forest*, which is an ensemble learning method consisting of a collection of decision trees. A decision tree is a tree structure that recursively partitions the input space, and learns a constant prediction for each partition element. The random forest algorithm collects the output of the individual decision trees, and returns their average as the prediction. Second, we estimate a *neural net*, which is a network architecture that repeatedly applies nonlinear transformations to primitive features, and aggregates these

[12] See the survey Fehr-Duda and Epper (2012) for further discussion of these different parametric forms, and others which have been proposed in the literature.

transformations into new layers of features. Both approaches are capable of learning complex nonlinear relationships between the features and the outcome to be predicted.

5.3 Within Domain Performance

We first consider how well each of these models performs when trained and evaluated on data from the same domain. Fixing a decision rule ρ, we define the rule's in-sample error in domain d to be $e_d \equiv e(\rho(S_d), S_d)$. We also calculate an out-of-sample error using the standard tenfold cross-validated procedure.[13] Figure 1 reports histograms of in-sample and tenfold cross-validated errors (across the 44 domains) for CPT with 4 parameters and for the random forest algorithm. As expected, since the black box methods can better capture details of the distributions in each of these domains, the distribution of black box errors is first-order stochastically dominated by the distribution of CPT errors. Interestingly, except for a single outlier domain, the distribution of cross-validated black box errors is also first-order stochastically dominated by the distribution of cross-validated CPT errors.



(a) *In sample transfer error*

(b) *CV error*

Figure 1: *Within Domain Prediction Errors*

To obtain a single summary statistic for the comparison between the CPT and random forest errors, we normalize the CPT errors by the random forest error domain-by-domain, and report the average of this ratio (across domains) in Table 1. The average ratio turns out

[13] We split the sample into ten subsets at random, choose nine of the ten subsets for training, and evaluate the estimated model's error on the final subset. The tenfold cross-validated error is the average of the out-of-sample errors on the ten possible choices of test set.

Model	In-Sample	Cross-Validated
CPT variants		
γ	1.31	1.32
α, β	1.25	1.25
δ, γ	1.22	1.23
α, β, γ	1.19	1.19
$\alpha, \beta, \delta, \gamma$	1.18	1.19

Table 1: *Average Ratio of Errors (CPT/Random Forest)*

to be only slightly larger than 1 for each of the CPT variants, so the CPT error is on average larger than the random forest error but not by much. For example, the out-of-sample error of the 4-parameter variant of CPT is on average only 1.18 times as large as the random forest error. This is consistent with Fudenberg et al. (2021)'s finding that CPT is not a particularly restrictive model on the domain of binary lotteries (i.e., most conditional mean functions can be well approximated by CPT with some choice of parameter values).

From these results alone, we cannot distinguish whether the better within-domain performance of the random forest algorithm is achieved by learning generalizable structure that CPT misses, or if the random forest simply learns idiosyncratic details of behavior in each domain that do not transfer across domains. We can distinguish between these two explanations by evaluating the transfer performance of these models.

5.4 Transfer Performance

We next apply our approach to construct confidence intervals for transfer error, normalized transfer error, and transfer deterioration for each of the decision rules described above. In our metadata, there are $m = 44$ domains, and we choose $k = 1$ of these to use as the training domain. This choice of k corresponds to the question, "If I draw one domain at random, and then try to generalize to another domain, how well do I do?," which we view as a natural option to think about. (In Appendix B.3 we report results for a different choice of k.)

Table 2 reports two-sided confidence intervals constructed using Proposition 1 with $\tau = 0.95$. That is, the upper bound of the confidence interval is the 95% percentile of the pooled transfer errors (across realizations of the training and test domains), and the lower

Model	Transfer Error	Normalized Error	Deterioration
CPT variants			
γ	[1.89,22.98]	[1.07,1.57]	[1.00,1.16]
α, β	[1.63,21.30]	[1.13,1.89]	[1.00,1.56]
δ, γ	[1.59,19.56]	[1.04,1.89]	[1.00,1.42]
α, β, γ	[1.51,21.03]	[1.05,1.76]	[1.00,1.47]
$\alpha, \beta, \delta, \gamma$	[1.50,20.77]	[1.05,1.88]	[1.00,1.59]
ML algorithms			
Random Forest	[3.54,58.62]	[1.01,6.51]	[1.01,6.51]
Neural Net	[3.82,132.43]	[1.01,13.85]	[1.01,13.85]

Table 2: *71% Confidence Intervals*

bound of the confidence interval is the 5% percentile of the pooled transfer errors. Applying Proposition 1, these are 71% confidence intervals for the parameters of interest. Choosing larger τ results in wider confidence intervals that have higher confidence levels. In Appendix B.2, we report some of these alternative confidence intervals, including a more traditional 90% confidence interval.

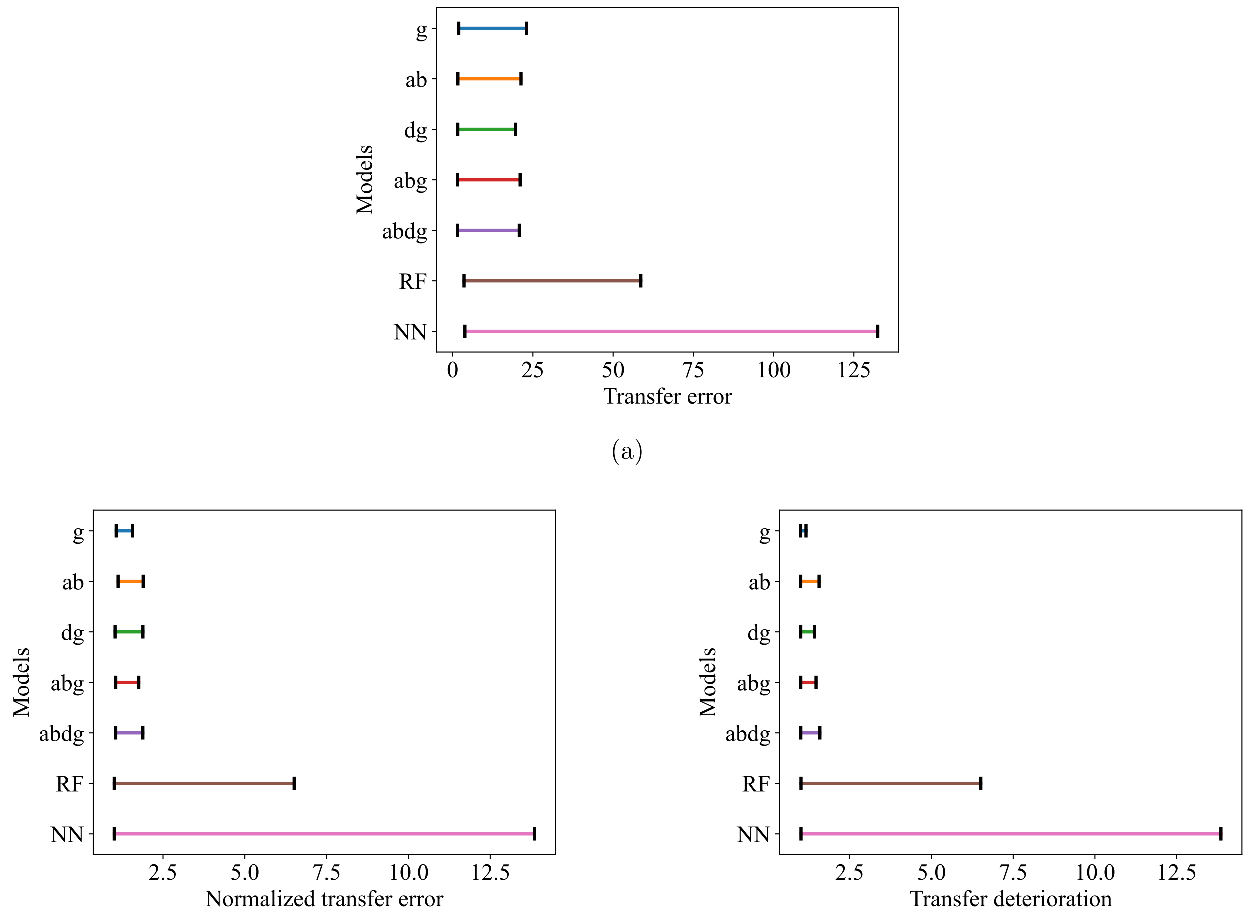
Panel (a) of Figure 2 plots these 71% confidence intervals for transfer error. We find that while the confidence intervals overlap, the midpoints and upper bounds of the black box confidence intervals are substantially higher than the corresponding quantities for the CPT confidence intervals, which suggests that the economic models generalize better in this setting. We find that the confidence intervals for black box methods are significantly wider than the confidence intervals for the economic models. This suggests that we should have less certainty in what the transfer error of a black box method will be.

Relative to the clear differences between the black box methods and economic models, the variation in these confidence intervals for the different CPT specifications is quite small. There is however an interesting comparison between the two-parameter specifications of CPT: CPT(δ, γ), which assumes risk-neturality but permits nonlinear probability weighting, and CPT(α, β), which assumes that the agent is an expected utility maximizer, but permits the agent to be risk averse. Although the difference is not large, we find that both confidence bounds are shifted higher for CPT(α, β) than for CPT(δ, γ), and that this is true not only for the transfer error confidence intervals, but also for our other two measures. This suggests that allowing for probability weighting parameters may be more important for capturing common structure across subject pools than allowing for risk aversion.

Panel (b) of Figure 2 plots the 71% confidence intervals for R-normalized transfer error, where the set R of decision rules includes the rules described in Section 5.2. Here we again find that the confidence intervals for black box transfer errors are significantly wider and also higher. Across the CPT models, the largest upper bound for the confidence interval is 1.89, corresponding to a transfer error that is two times larger than the best in-sample error. In contrast, the upper bound for the confidence interval for the neural net is 13.85, suggesting that the transfer error of a black box model trained on a different domain can be much worse than the best in-sample error. This contrast is especially interesting in light of our earlier finding in Section 5.3 that CPT's error is not very different from the random forest error in the within-domain prediction problem. Thus, although CPT and the black box models learn representations that are close for the purposes of prediction in one domain, these representations can lead to rather different predictions in new domains, and the representation learned by CPT generalizaes better.

Panel (c) of Figure 2 plots 71% confidence intervals for transfer deterioration. Here we observe a clear monotonicity: both the lower and upper bound of the confidence intervals are weakly increasing in the number of CPT parameters (and larger still for the black box methods). This monotonicity is not guaranteed by definition of transfer deterioration, since although in-sample error must weakly decrease when a new parameter is added, this decrease in in-sample error could in principle be compensated for by an even larger decrease in transfer error. Panel (c) tells us that this is not the case.

Finally, we note that while the confidence intervals in Figure 2 are constructed using the 5% and 95% percentiles of pooled transfer errors, these percentile choices are not special. Figure 3 reports for each of our three measures a plot of the τ-th percentile of the pooled transfer errors as τ varies. To improve readability, we remove extreme numbers by truncating $\tau \in [5, 95]$. Any paired choice of τ and $(1-\tau)$-th percentiles from these figures can be used to construct a confidence interval, again applying Proposition 1.



(a)

(b)

(c)

Figure 2: *71% Confidence Intervals for (a) Transfer Error, (b) Normalized Transfer Error, and (c) Transfer Deterioration.*

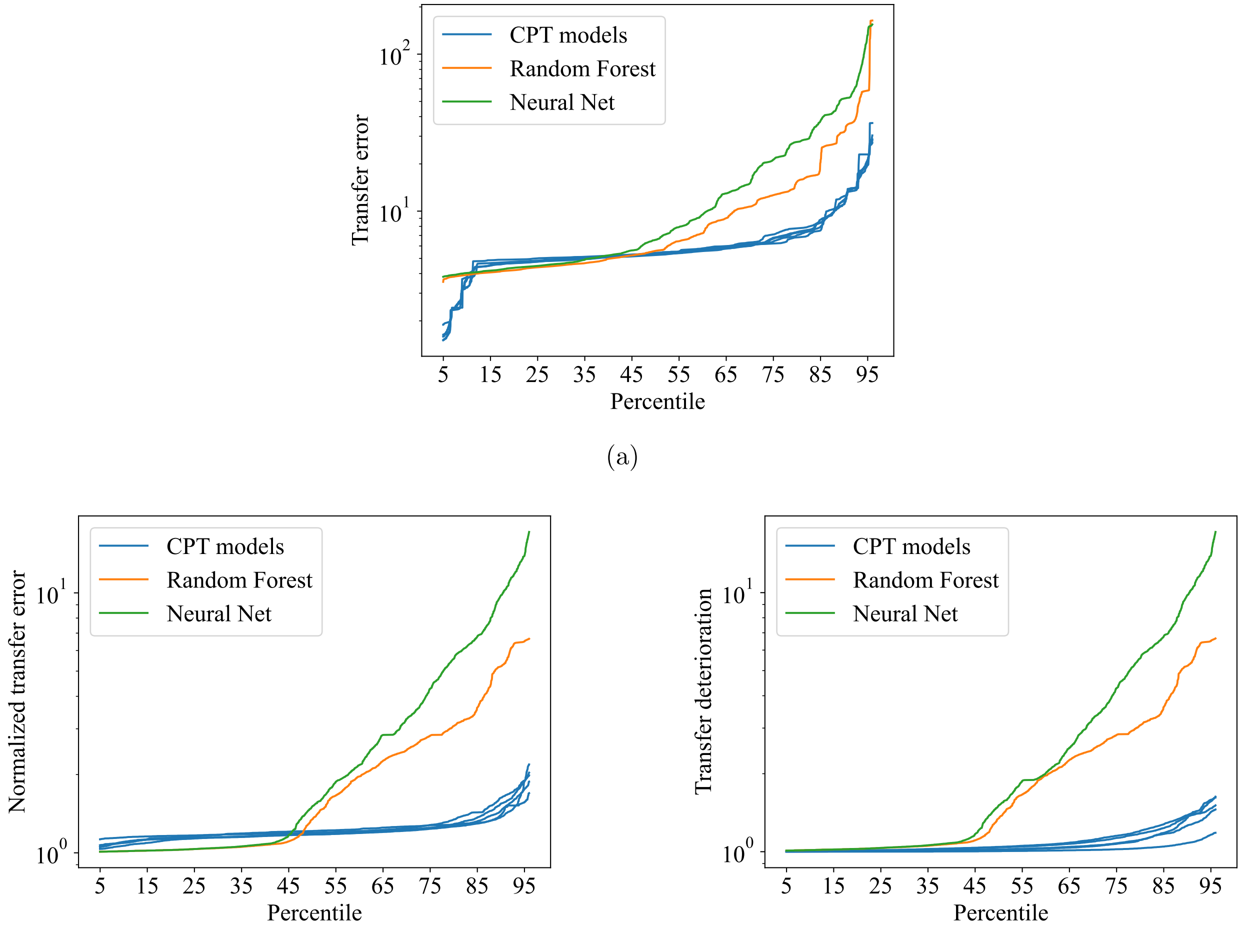
5.5 Comparing CPT with the Random Forest Algorithm

We now conduct a more detailed study of the comparison between the transfer error of CPT and the transfer error of the random forest algorithm. Let ρ^{CPT} denote the decision rule corresponding to CPT, and ρ^{RF} denote the decision rule corresponding to the random forest algorithm. Define

$$r_{T,d} = \frac{e(\rho^{RF}(\mathbf{M}_T), S_d)}{e(\rho^{CPT}(\mathbf{M}_T), S_d)}$$

to be the ratio of the random forest transfer error to the CPT transfer error, henceforth the *transfer error ratio*.

Panel (a) of Figure 4 reports 71% confidence intervals for $r_{T,d}$ for each of the CPT



(a)

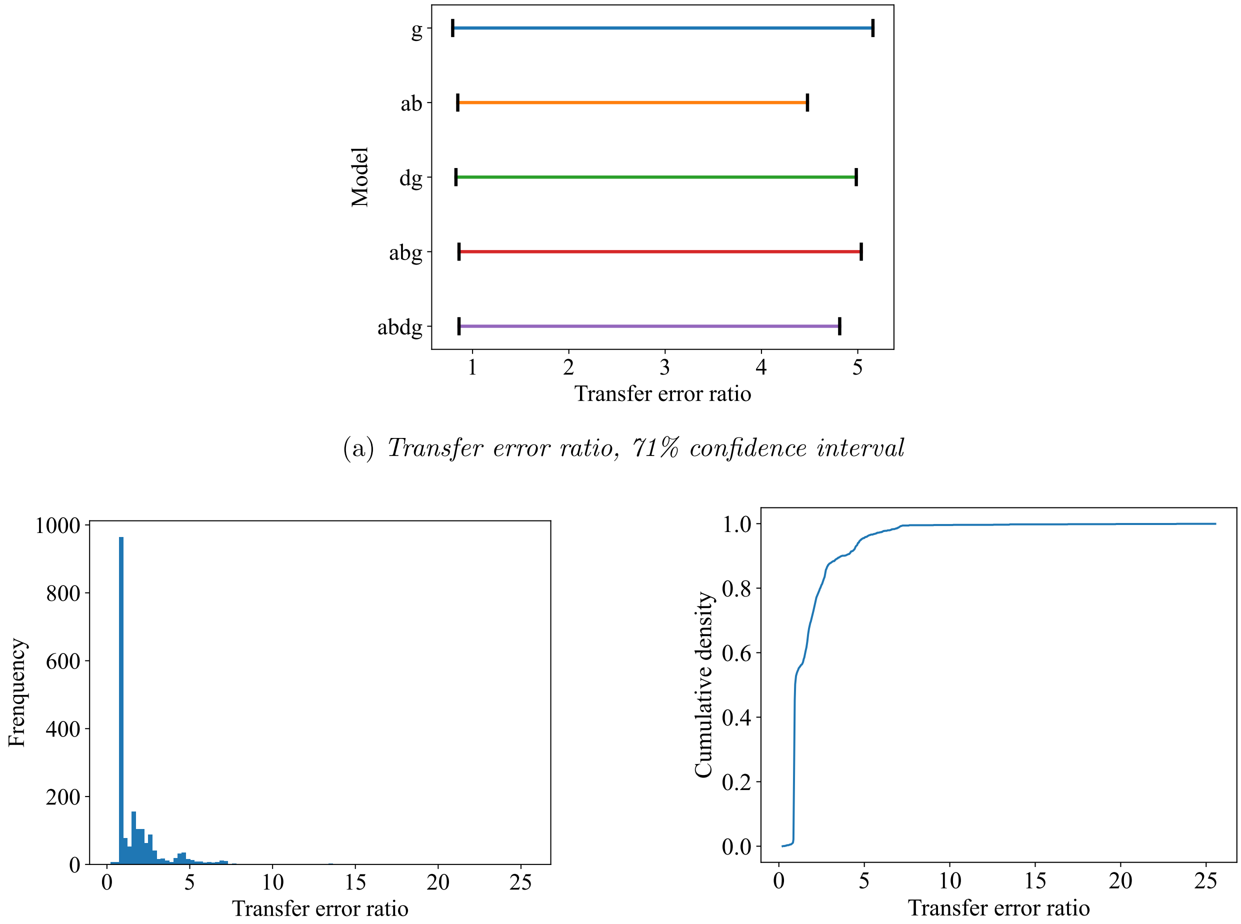
(b)

(c)

Figure 3: *The τ-th percentile of (a) Transfer Error, (b) Normalized Transfer Error, and (c) Transfer Deterioration.*

specifications (applying Proposition 1). The lower bound for each of the CPT models is approximately 0.8, while the upper bound is as large as 5. These confidence intervals reflect both our uncertainty about the size of the transfer errors due to the finiteness of our metadata, and also variation in the transfer error due to the random selection of training and target domains. Thus the width of these intervals does not go to zero even if we have data from many domains, and we expect the finding that these confidence intervals include 1 will be robust.

Panel (b) of Figure 4 reports the histogram of transfer error ratios for the CPT model with 4 parameters when the training domains T and the target domain d are drawn uniformly at random from the set of domains in the metadata. This distribution has a large cluster of ratios just below 1 (i.e., CPT transfers slightly worse than the random forest) and a long right tail of ratios achieving a max value of 25.5 (i.e., the random forest error can be up to 25 times as large as the CPT error). The cumulative distribution function of $r_{T,d}$, reported in Panel (c) of Figure 4, reveals that the random forest algorithm in fact outperforms CPT in a majority (approximately 53%) of (T, d) pairs, but CPT rarely achieves a transfer error that is much worse than the random forest and sometimes achieves a transfer error that is much better. One possible explanation for this distribution of ratios is that behavior is similar across a large set of domains, and quite heterogeneous across the remaining domains. If this is so, the random forest algorithm (which picks up on more fine-tuned details of behavior) may achieve better transfer prediction among the set of similar domains, but transfer very poorly from these domains to the remaining domains. In contrast, if CPT identifies structure commonly shared across all of these domains, then (as we find) its transfer performance would be relatively stable. Together with the previous findings, these results suggest that economic models may not always transfer better than black box methods, but they do transfer more reliably.



(a) *Transfer error ratio, 71% confidence interval*

(b) *Transfer error ratio, histogram*

(c) *Transfer error ratio, CDF*

Figure 4: *Ratio of random forest transfer error to CPT transfer error*

References

Mohammed Abdellaoui, Peter Klibanoff, and Lætitia Placido. 2015. Experiments on compound risk in relation to simple risk and to ambiguity. *Management Science* 61, 6 (2015), 1306–1322.

Hunt Allcott. 2015. Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics* 130, 3 (2015), 1117–1165.

Vital Anderhub, Werner Güth, Gneezy, and Sonsino. 2001. On the interaction of risk and time preferences: An experimental study. *German Economic Review* 2, 3 (2001), 239–253.

Shlomo Benartzi, John Beshears, Katherine L. Milkman, Cass R. Sunstein, Richard H. Thaler, Maya Shankar, Will Tucker-Ray, William J. Congdon, and Steven Galing. 2017. Should Governments Invest More in Nudging? *Psychological Science* 28, 8 (2017), 1041–1055.

B Douglas Bernheim and Charles Sprenger. 2020. On the empirical validity of cumulative prospect theory: Experimental evidence of rank-independent probability weighting. *Econometrica* 88, 4 (2020), 1363–1409.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc.

Ranoua Bouchouicha and Ferdinand M Vieider. 2017. Accommodating stake effects under prospect theory. *Journal of Risk and Uncertainty* 55, 1 (2017), 1–28.

Adrian Bruhin, Helga Fehr-Duda, and Thomas Epper. 2010. Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica* 78, 4 (2010), 1375–1412.

Colin F Camerer, Gideon Nave, and Alec Smith. 2019. Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning. *Management Science* 65, 4 (2019), 1867–1890.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. 2021. An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. *J. Amer. Statist. Assoc.* 116, 536 (2021), 1849–1864.

Mark Dean and Pietro Ortoleva. 2019. The empirical relationship between nonstandard economic behaviors. *Proceedings of the National Academy of Sciences* 116, 33 (2019), 16262–16267.

Stefano DellaVigna and Devin Pope. 2019. Stability of Experimental Results: Forecasts and Evidence. (2019).

Nathalie Etchart-Vincent and Olivier l'Haridon. 2011. Monetary incentives in the loss domain and behavior toward risk: An experimental comparison of three reward schemes including real losses. *Journal of risk and uncertainty* 42, 1 (2011), 61–83.

Yuyu Fan, David V Budescu, and Enrico Diecidue. 2019. Decisions with compound lotteries. *Decision* 6, 2 (2019), 109.

Helga Fehr-Duda, Adrian Bruhin, Thomas Epper, and Renate Schubert. 2010. Rationality on the rise: Why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty* 40, 2 (2010), 147–180.

Helga Fehr-Duda and Thomas Epper. 2012. Probability and Risk: Foundations and Economic Implication of Probability-Dependent Risk Preferences. *Annual Review of Economics* 4 (2012), 567–593.

Drew Fudenberg, Wayne Gao, and Annie Liang. 2021. How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories. (2021). Working Paper.

Drew Fudenberg, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. Measuring the Completeness of Economic Models. (2022). Forthcoming in the Journal of Political Economy.

Drew Fudenberg and Annie Liang. 2019. Predicting and Understanding Initial Play. *American Economic Review* 109, 12 (2019), 4112–4141.

Yoram Halevy. 2007. Ellsberg revisited: An experimental study. *Econometrica* 75, 2 (2007), 503–536.

Dennis Hummel and Alexander Maedche. 2019. How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics* 80 (2019), 47–58.

Uday Karmarkar. 1978. Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior & Human Performance* 21, 1 (1978), 67–72.

Shaowei Ke, Chen Zhao, Zhaoran Wang, and Sung-Lin Hsieh. 2020. Behavioral Neural Networks. (2020). Working Paper.

Mathieu Lefebvre, Ferdinand M Vieider, and Marie Claire Villeval. 2010. Incentive effects on risk attitude in small probability prospects. *Economics Letters* 109, 2 (2010), 115–120.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. Distribution-Free Predictive Inference for Regression. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1094–1111.

Olivier l'Haridon and Ferdinand M Vieider. 2019. All over the map: A worldwide comparison of risk preferences. *Quantitative Economics* 10, 1 (2019), 185–215.

Rachael Meager. 2019. Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics* 11, 1 (2019), 57–91.

Xiao-Li Meng. 1994. Posterior Predictive p-Values. *Annals of Statistics* 22, 3 (1994), 1142–1160.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain Generalization via Invariant Feature Representation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* (Atlanta, GA, USA) *(ICML'13)*. JMLR.org, I–10–I–18.

Zahra Murad, Martin Sefton, and Chris Starmer. 2016. How do risk attitudes affect measured confidence? *Journal of Risk and Uncertainty* 52, 1 (2016), 21–46.

Christiana A. Naaktgeboren, Eleanor A. Ochodo, Wynanda A. Van Enst, Joris A. H. de Groot, Lotty Hooft, Mariska M. G. Leeflang, Patrick M. Bossuyt, Karel G. M. Moons, and Johannes B. Reitsma. 2016. *BMC Medical Research Methodology* 16, 1 (2016), 6.

Gali Noti, Effi Levi, Yoav Kolumbus, and Amit Daniely. 2016. Behavior-Based Machine-Learning: A Hybrid Approach for Predicting Human Decision Making. *CoRR* abs/1611.10228 (2016). arXiv:1611.10228

Ori Plonsky, Reut Apel, Eyal Ert, Moshe Tennenholtz, David Bourgin, Joshua Peterson, Daniel Reichman, Thomas Griffiths, Stuart Russell, Evan Carter, James Cavanagh, and Ido Erev. 2019. Predicting human decisions with behavioral theories and machine learning. *CoRR* abs/1904.06866 (2019). arXiv:1904.06866

Ori Plonsky, Ido Erev, Tamir Hazan, and Moshe Tennenholtz. 2017. Psychological forest: Predicting human behavior. *AAAI Conference on Artificial Intelligence* 31, 1 (2017), 656–662.

Ludger Rüschendorf. 1982. Random Variables with Maximum Sums. *Advances in Applied Probability* 14, 3 (1982), 623–632.

Shai Shalev-Shwartz and Shai Ben-David. 2019. *Understanding machine learning: From* *theory to algorithms*. Cambridge University Press.

Matthias Sutter, Martin G Kocher, Daniela Glätzle-Rützler, and Stefan T Trautmann. 2013. Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review* 103, 1 (2013), 510–31.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2021. Domain Generalization: A Survey. (2021). Working Paper.

A Supplementary Material to Section 4

A.1 Proof of Proposition 1

For a given $T \sim \phi(n_T, [n])$, let $F_T(a) = \sum_{d \in [n] \setminus T} 1\{e_{T,d} < a\}$ count the number of held-out domains with transfer error strictly below a. The assumption that $(\mathcal{P}_d, m_d)$ is drawn iid from μ in each domain implies that if we evaluate $F_T(\cdot)$ at the transfer error of the test domain, $e_{T,n+1}$, the distribution of $F_T(e_{T,n+1})$ is dominated by a discrete uniform distribution over $\{0, ..., n - n_T\}$ in the sense of first-order stochastic dominance. Hence, for $F_T^*(a) = \frac{1}{n-n_T+1} F_T(a)$, the distribution of $F_T^*(e_{T,n+1})$ is dominated by a uniform $[0,1]$ distribution, again in the sense of first-order stochastic dominance. In particular, for each T we can construct a random variable $U_T \sim U[0,1]$ such that $F_T^*(e_{T,n+1}) \leq U_T$.[14] For $T \sim \phi(n_T, [n])$, U_T is a mean-preserving spread of $\mathbb{E}_{T \sim \phi(n_T,[n])}[U_T]$.

We now use Lemma 1 from Meng (1994), which applies to a random variable W with mean 1/2 that is second-order dominated by a uniform random variable. Setting $W = 1 - \mathbb{E}_{T \sim \phi(n_T,[n])}[U_T]$ and $\alpha = 1 - \psi$ in the lemma's conclusion shows that

$$\mathbb{P}\left\{\mathbb{E}_{T \sim \phi(n_T,[n])}[U_T] \geq \psi\right\} \leq 2(1-\psi).^{15}$$

By construction $\mathbb{E}_{T \sim \phi(n_T,[n])}[F_T^*(e_{T,n+1})] \leq \mathbb{E}_{T \sim \phi(n_T,[n])}[U_T]$, so

$$\mathbb{P}\left\{\mathbb{E}_{T \sim \phi(n_T,[n])}[F_T^*(e_{T,n+1})] \geq \psi\right\} \leq 2(1-\psi).$$

Hence, if we define a confidence set by collecting the set of values a where $\mathbb{E}_{T \sim \phi(n_T,[n])}[F_T^*(a)]$ falls below ψ,

$$CS = \{a : \mathbb{E}_{T \sim \phi(n_T,[n])}[F_T^*(a)] < \psi\},$$

we have that

$$\mathbb{P}\{e_{T,n+1} \in CS\} \geq 1 - 2(1-\psi) = 2\psi - 1.$$

To characterize CS, note that by definition

$$\mathbb{E}_{T \sim \phi(n_T,[n])}[F_T^*(a)] = \frac{n - n_T}{n - n_T + 1} \mathbb{E}_{(T,t) \sim \phi(n_T+1,[n])}[1\{e_{T,t} < a\}].$$

[14] In the case where the transfer errors $e_{T,t}$ are continuously distributed, for instance, it suffices to take $U_T = F_T^*(e_{T,n+1}) + u_T$ for $u_T \sim [0, \frac{1}{n-n_T+1}]$ independent of the data.

[15] This conclusion may also be obtained using earlier results from Rüschendorf (1982).

Hence,

$$CS = \left\{a : \mathbb{E}_{(T,t)\sim\phi(n_T+1,[n])}[1\{e_{T,t} < a\}] < \frac{n - n_T + 1}{n - n_T}\psi\right\} =$$
$$\left\{a : \mathbb{P}_{(T,t)\sim\phi(n_T+1,[n])}\left(\{e_{T,t} \geq a\}\right) \geq \frac{n - n_T + 1}{n - n_T}\psi\right\},$$

and $\sup\{a : a \in CS\} = \overline{Q}_{\tau,(T,t)\sim\phi(n_T+1,[n])}(e_{T,t})$ for $\tau = \frac{n-n_T+1}{n-n_T}\psi$ by the definition of $\overline{Q}_\tau$. The result for one-sided confidence intervals is immediate, while the result for two-sided intervals follows by applying the same argument to the lower tail.

A.2 Extension to Partial Transfer

So far we have supposed that the model is trained on the training data and ported fully to the new domain. In some cases, the analyst may wish to transfer certain parameters estimated on the metadata, while re-estimating other parameters on the new domain. Here it is useful to explicitly parametrize $\Sigma^* = \{\sigma_{\theta,\lambda}\}_{\theta\in\Theta,\lambda\in\Lambda}$, where the parameters in the compact set Λ are domain-specific parameters that are re-estimated, while the parameters in the compact set Θ are ported across domains. Our definitions and methodology can be generalized to accommodate this "partial transfer."

For any θ and sample S, define

$$e(\theta, S) = \min_{\lambda\in\Lambda} \frac{1}{|S|} \sum_{(x,y)\in S} \ell(\sigma_{\theta,\lambda}(x), y)$$

to be the minimal achievable error on the sample S when optimizing over λ given the fixed value of θ. The *partial transfer error from* $\mathbf{M}_T$ *to* S is then $e(\theta_{\mathbf{M}_T}, S)$, where

$$\theta_{\mathbf{M}_T} = \operatorname*{argmin}_{\theta\in\Theta} \left(\operatorname*{argmin}_{\lambda\in\Lambda} e(\sigma_{\theta,\lambda}, \mathbf{M}_T)\right)$$

is the best-fit value of θ on the training data $\mathbf{M}_T$. Defining $e_{T,d} \equiv e(\theta_{\mathbf{M}_T}, S_d)$, all of our results in Section 4 hold without modification.

B Supplementary Material to Section 5

B.1 Description of Data

We briefly describe the individual samples in our metadata. There are 44 total domains.

Source of Data	# Obs	# Subj	Country	Gains Only
Abdellaoui et al. (2015)	801	89	France	Y
Fan et al. (2019)	4750	125	US	Y
Bouchouicha and Vieider (2017)	3162	94	UK	N
Sutter et al. (2013)	661	661	Austria	Y
Etchart-Vincent and l'Haridon (2011)	3036	46	France	N
Fehr-Duda et al. (2010)	8560	153	China	Y
Lefebvre et al. (2010)	72	72	France	Y
Halevy (2007)	366	122	Canada	Y
Anderhub et al. (2001)	183	61	Israel	Y
Murad et al. (2016)	2131	86	UK	Y
Dean and Ortoleva (2019)	1032	179	US	Y
Bernheim and Sprenger (2020)	1071	153	US	Y
Bruhin et al. (2010)	8906	179	Switzerland	N
Bruhin et al. (2010)	4669	118	Switzerland	N
l'Haridon and Vieider (2019)	1708	61	Australia	N
l'Haridon and Vieider (2019)	2548	95	Belgium	N
l'Haridon and Vieider (2019)	2350	84	Brazil	N
l'Haridon and Vieider (2019)	2240	80	Cambodia	N
l'Haridon and Vieider (2019)	2687	96	Chile	N
l'Haridon and Vieider (2019)	5711	204	China	N
l'Haridon and Vieider (2019)	3072	128	Colombia	N
l'Haridon and Vieider (2019)	2968	106	Costa Rica	N
l'Haridon and Vieider (2019)	2770	99	Czech Republic	N
l'Haridon and Vieider (2019)	3906	140	Ethiopia	N
l'Haridon and Vieider (2019)	2604	93	France	N
l'Haridon and Vieider (2019)	3639	130	Germany	N
l'Haridon and Vieider (2019)	2352	84	Guatemala	N
l'Haridon and Vieider (2019)	2492	89	India	N
l'Haridon and Vieider (2019)	2352	84	Japan	N
l'Haridon and Vieider (2019)	2716	97	Kyrgyzstan	N

l'Haridon and Vieider (2019)	1791	64	Malaysia	N
l'Haridon and Vieider (2019)	3360	120	Nicaragua	N
l'Haridon and Vieider (2019)	5638	202	Nigeria	N
l'Haridon and Vieider (2019)	2660	95	Peru	N
l'Haridon and Vieider (2019)	2491	89	Poland	N
l'Haridon and Vieider (2019)	1959	70	Russia	N
l'Haridon and Vieider (2019)	1819	65	Saudi Arabia	N
l'Haridon and Vieider (2019)	1988	71	South Africa	N
l'Haridon and Vieider (2019)	2240	80	Spain	N
l'Haridon and Vieider (2019)	2212	79	Thailand	N
l'Haridon and Vieider (2019)	2070	74	Tunisia	N
l'Haridon and Vieider (2019)	2240	80	UK	N
l'Haridon and Vieider (2019)	2701	97	US	N
l'Haridon and Vieider (2019)	2436	87	Vietnam	N

B.2 Alternative Confidence Intervals

We report here alternative confidence intervals for our three measures. Table 4 constructs two-sided confidence intervals whose lower bounds are the minimum transfer error (among the pooled transfer errors) and upper bounds are the maximum transfer error. Applying Proposition 1, these are 90% confidence intervals. Table 5 constructs one-sided confidence intervals whose upper bounds are the 95% transfer error. Applying Proposition 1, these are 86% confidence intervals. These confidence intervals share all the properties described in the main text for the two-sided 71% confidence intervals.

B.3 Alternative Choice of n_T

Here we consider an alternative choice for the number of training domains, setting $n_T = 3$ instead of $n_T = 1$.[16] This corresponds to randomly choosing 3 of the 44 domains to be the training domains, finding the best strategy in Σ^* for this pooled data, and using the estimated strategy to predict the remaining 41 samples.

[16] We chose $n_T = 3$ to preserve a large number of testing domains, and to keep the number of tests manageable. (The choice of $n_T = 3$ already requires computing $\binom{44}{3} * 31 = 543,004$ transfer errors.)

Model	Transfer Error	Normalized Error	Deterioration
CPT main variants			
γ	[0.81,13413.09]	[1.03,6.74]	[1.00,5.59]
α, β	[0.71,18421.52]	[1.00,6.77]	[1.00,5.27]
δ, γ	[0.71,25526.76]	[1.00,7.34]	[1.00,5.95]
α, β, γ	[0.71,18421.52]	[1.00,6.87]	[1.00,5.60]
$\alpha, \beta, \delta, \gamma$	[0.71,21023.20]	[1.00,6.83]	[1.00,5.95]
ML algorithms			
Random Forest	[1.04,29468.03]	[1.00,75.91]	[1.00,75.91]
Neural Net	[0.88,29496.19]	[1.00,113.85]	[1.00,113.85]

Table 4: *90% Confidence Intervals*

Figure 5 is the analog of Figure 2. Again we choose $\tau = 0.95$, thus constructing confidence intervals whose lower bounds are the 5% percentile of pooled transfer errors, and whose upper bounds are the 95% percentile of pooled transfer errors. Applying Proposition 1, these are 54% confidence intervals. The most notable change is that the random forest confidence interval shrinks considerably, which suggests the transfer error of the random forest algorithm becomes less variable when it is trained on more domains. Otherwise, all of the qualitative statements we main in the main text for $k = 1$ continue to hold. In particular, as with $n_T = 1$, we find that the confidence intervals for each of our measures (transfer error, normalized transfer error, and transfer deterioration) have higher lower and upper bounds for the black box algorithms than the CPT specifications.

Figure 6 is the analog of Figure 4. The qualitative findings are again the same as in the main text, but we see further evidence that the random forest algorithm improves when trained on samples from more domains. For example, the confidence intervals for the transfer error ratio (Panel (a)) now range from approximately 0.8 to approximately 4.3, instead of 0.8 to 5. Comparing Panel (b) of Figures 4 and 6, we also see that the distribution of ratios shifts down, so the relative performance of the random forest algorithm improves. This finding that black box methods are more robust when trained on a greater variety of data samples is consistent with earlier findings described in Zhou et al. (2021).

Model	Transfer Error	Normalized Error	Deterioration
CPT main variants			
γ	[0,22.98]	[1,1.57]	[1,1.16]
α, β	[0,21.30]	[1,1.89]	[1,1.56]
δ, γ	[0,19.56]	[1,1.89]	[1,1.42]
α, β, γ	[0,21.03]	[1,1.76]	[1,1.47]
$\alpha, \beta, \delta, \gamma$	[0,20.77]	[1,1.88]	[1,1.59]
ML algorithms			
Random Forest	[0,58.62]	[1,6.51]	[1,6.51]
Neural Net	[0,132.43]	[1,13.85]	[1,13.85]

Table 5: *86% Confidence Intervals*

C (Transfer) Rademacher Complexity

Consider the one-domain statistical learning framework reviewed in Section 3.1, where (X, Y) is governed by the distribution $\mathcal{P}$. The well-known *Rademacher complexity* of a set of strategies Σ^* with respect to the sample S is defined as

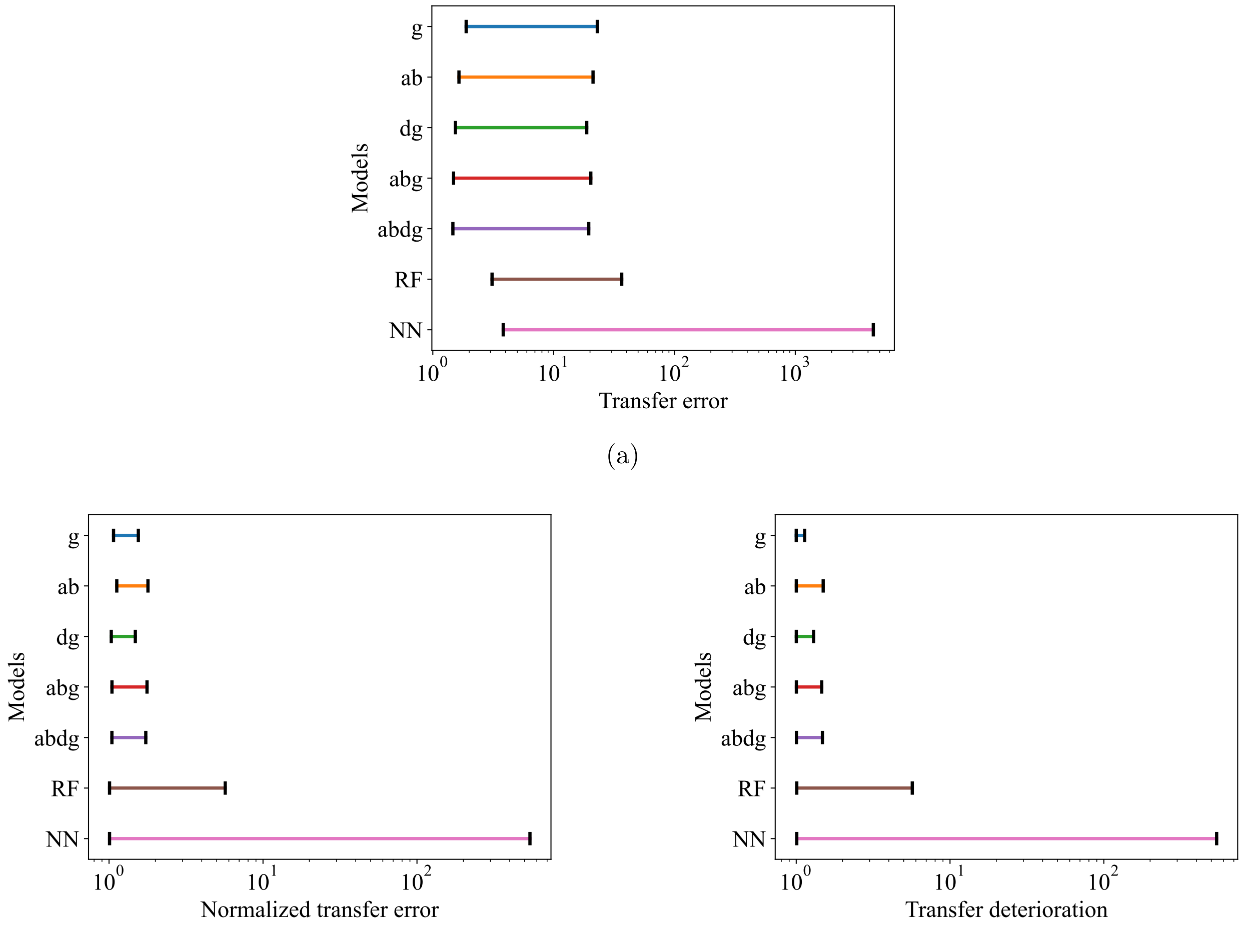
$$R(\Sigma^*, S) := \frac{1}{m}\mathbb{E}_{w\sim\{\pm 1\}^m}\left[\sup_{\sigma\in\Sigma^*}\sum_{(x_i,y_i)\in S} w_i \cdot \ell(\sigma(x_i), y_i)\right]$$

where $w \sim \{\pm 1\}^m$ means that w is drawn uniformly at random from the set of length-m binary vectors. Rademacher complexity is a measure for the expressiveness of Σ^*. We adapt this idea for our multiple-domain setting as follows.

Definition 4. *The* transfer Rademacher complexity *of* Σ^* *given metadata* $\mathbf{M} = (S_1, \ldots, S_m)$ *is*

$$\mathcal{R}(\Sigma^*, \mathbf{M}) = \frac{1}{m}\mathbb{E}_{w\sim\{\pm 1\}^m}\left[\sup_{\sigma\in\Sigma^*}\sum_{i=1}^{m} w_i \cdot \left(\frac{1}{n_i}\sum_{(x^i_j,y^i_j)\in S_i} \ell(\sigma(x^i_j), y^i_j)\right)\right]$$

Transfer Rademacher complexity can be used to upper bound on the expected transfer error of a strategy. The result below complements our results in the main text, which provide confidence intervals for the (random) transfer error on the target sample, but do not have implications for the expected transfer error.



(a)

(b)

(c)

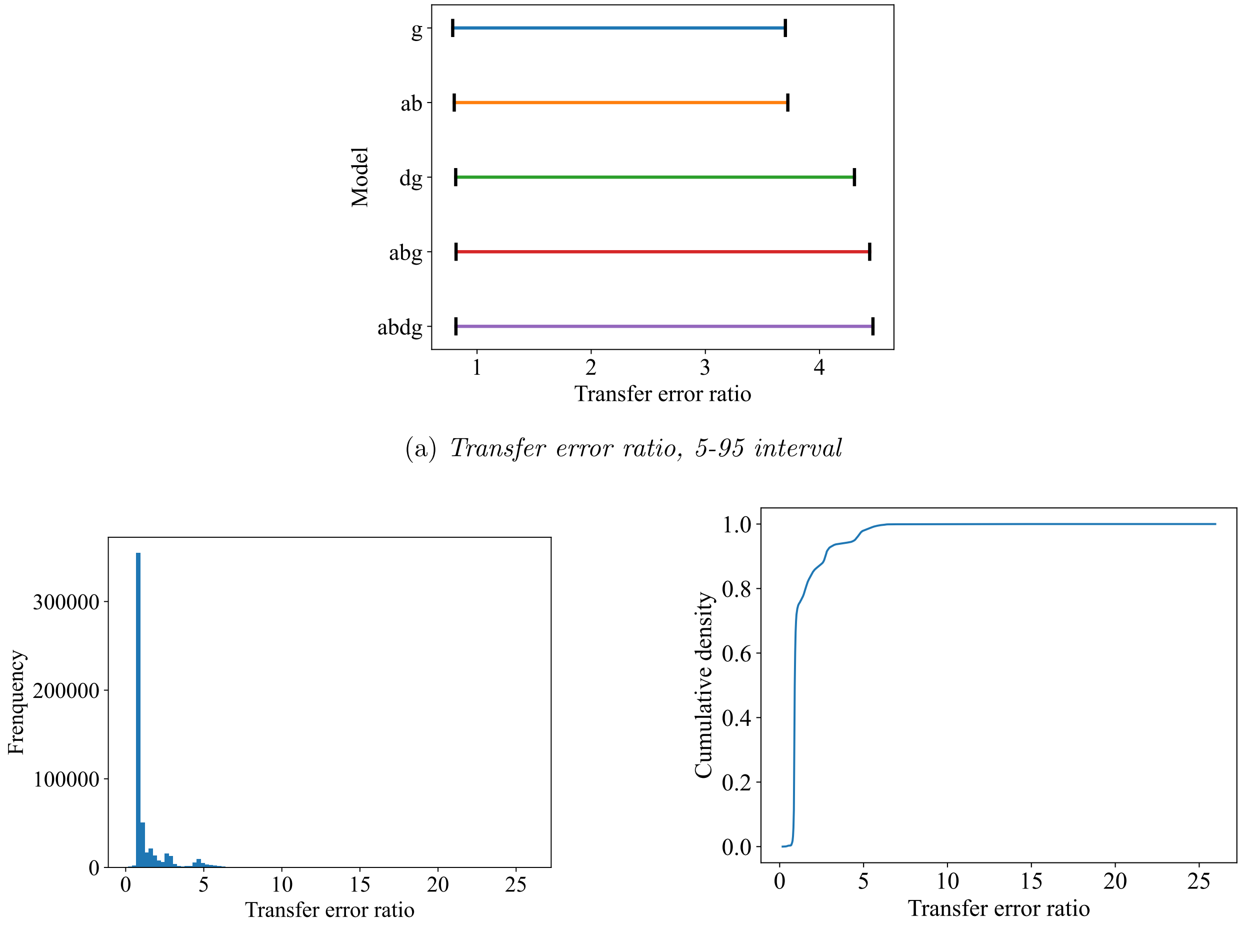
Figure 5: *54% Confidence Intervals for (a) Transfer Error, (b) Normalized Transfer Error, and (c) Transfer Deterioration, with the choice of $k = 3$.*

Proposition 2. *Assume that for all $z \in \mathcal{X} \times \mathcal{Y}$ and $\sigma \in \Sigma^*$ we have that $|\ell(\sigma, z)| \leq c$. Then, with probability at least $1 - \delta$, for all $\sigma \in \Sigma^*$,*

$$E_{S \sim \mu}[e(\sigma, S)] \leq e(\sigma, \mathbf{M}) + 2R(\Sigma^*, \mathbf{M}) + 4c\sqrt{\frac{2\ln(4/\delta)}{n}}$$

where n is the size of $\mathbf{M}$.

We omit the proof since this result follows from a standard result for the one-domain setting, see for example Theorem 26.5 in Shalev-Shwartz and Ben-David (2019). This result says that the more "complex" the set of strategies Σ^* is (in the sense of higher transfer Rademacher complexity), the greater the possible difference between the error of $\sigma^*_{\mathbf{M}}$ on the



(a) *Transfer error ratio, 5-95 interval*

(b) *Transfer error ratio, histogram*

(c) *Transfer error ratio, CDF*

Figure 6: *Ratio of random forest transfer error to CPT transfer error, training on 3 domains*

metadata $\mathbf{M}$ and its error on a new sample from a new domain. In our application, the upper bound on the loss is so large that the bound is not useful.

We prove some properties of transfer Rademacher complexity below. Proposition 3 says that an FOSD shift up in the distribution of the number of observations per domain leads to a decrease in expected transfer Rademacher complexity, while Proposition 4 says that expected transfer Rademacher complexity is decreasing in the total number of domains. Throughout this section, we use μ_d denote the distribution of the domain d, and $\mu_{m|d}$ the conditional distribution of the sample size m given d. As in the main text, samples S are drawn according to $\mu = \mu_{n|d} \circ \mu_d$, which first draws $d \sim \mu_d$ and then draws $n \sim \mu_{n|d}$.

Proposition 3. *Suppose $\mu_{m|d}$ first order stochastically dominates $\mu'_{m|d}$ for all d. Then for* *any μ_d, if we set $\mu = \mu_{m|d} \circ \mu_d$ and $\mu' = \mu'_{m|d} \circ \mu_d$,*

$$\mathbb{E}_{\mathbf{M}\sim\mu^n}\left[R(\Sigma^*, \mathbf{M})\right] \leq \mathbb{E}_{\mathbf{M}\sim(\mu')^n}\left[R(\Sigma^*, \mathbf{M})\right]$$

i.e., a FOSD shift up in the distribution of the number of observations per domain leads to a decrease in expected Rademacher complexity.

Proof. We first establish two lemmas.

Lemma 1. *Suppose $X_1, X_2, \dots$ are independent and identically distributed random variables. Define $\overline{X}^n \equiv \frac{1}{n}\sum_{i=1}^n X_i$ for any $n \in \mathbb{Z}_+$. Then, for any $n > n'$, the distribution of $\overline{X}^{n'}$ is a mean preserving spread of the distribution of $\overline{X}^n$.*

Proof. Fix any $n, n' \in \mathbb{Z}_+$ satisfying $n > n'$. Then

$$\begin{aligned}\overline{X}^{n'} - \overline{X}^n &= \frac{1}{n'}(X_1 + \cdots + X_{n'}) - \frac{1}{n}(X_1 + \cdots + X_n) \\ &= \frac{1}{n \cdot n'}\left[(n - n')(X_1 + \cdots + X_{n'}) - n'(X_{n'+1} + \cdots + X_n)\right] \equiv \epsilon\end{aligned}$$

So $\overline{X}^{n'} = \overline{X}^n + \epsilon$. Moreover,

$$\begin{aligned}\mathbb{E}\left[\epsilon \mid \overline{X}^n\right] &= \frac{1}{n \cdot n'}\mathbb{E}\left[(n - n')(X_1 + \cdots + X_{n'}) - n'(X_{n'+1} + \cdots + X_n) \mid X_1 + \dots X_n\right] \\ &= \frac{(n - n')n'}{n \cdot n'} \cdot \mathbb{E}(X_1 \mid X_1 + \cdots + X_n) - \frac{n'(n - n')}{n - n'} \cdot \mathbb{E}(X_1 \mid X_1 + \cdots + X_n) = 0\end{aligned}$$

by symmetry and linearity of expectation. So the distribution of $\overline{X}^{n'}$ is a mean preserving spread of the distribution of $\overline{X}^n$, as desired. □

Lemma 2. *Fix any strategy σ and domain d. If $\mu_{m|d}$ FOSD $\mu_{m'|d}$, $M \sim \mu_{m|d}$ and $M' \sim \mu_{m'|d}$, and $Z_1, Z_2, \dots \sim_{iid} P_d$, then the distribution of $\frac{1}{M'}\sum_{j=1}^{M'} l(\sigma, Z_j)$ is a mean-preserving spread of the distribution of $\frac{1}{M}\sum_{j=1}^{M} l(\sigma, Z_j)$.*

Proof. Let $X_j \equiv l(\sigma, Z_j)$ for any $j \in \mathbb{Z}_+$, where we suppress the dependence of X_j on σ for brevity, and define the sample average

$$\overline{X}^m(\sigma) \equiv \frac{1}{m}\sum_{j=1}^m X_j \quad \text{for any } m \in \mathbb{Z}_+.$$

The desired conclusion holds if we can show that the distribution of $\overline{X}^{M'}(\sigma)$ is a mean preserving spread (mps) of the distribution of $\overline{X}^M(\sigma)$.

Choose any convex function $f : \mathbb{R} \to \mathbb{R}$ and consider

$$\mathbb{E}\left[f\left(\overline{X}^{M}(\sigma)\right)\right] = \int f(\overline{X}^{m}(\sigma))d\mu_m(m)$$

By Lemma 1, the distribution of $\overline{X}^{m'}(\sigma)$ is a mean preserving spread of the distribution of $\overline{X}^{m}(\sigma)$ for any fixed realizations $m > m'$. Thus $\mathbb{E}\left(f(\overline{X}^{m}(\sigma))\right)$ is decreasing in m, and a FOSD shift up in the distribution of μ_m must result in a lower value for the integral. The distribution of $\overline{X}^{M}(\sigma)$ thus dominates the distribution of $\overline{X}^{M'}(\sigma)$ in the convex order, implying that the latter is a mean-preserving spread of the former. □

Corollary 1. *Fix any $n \in \mathbb{Z}_+$ and any sequence of constants $w_1, \dots, w_n$. Let $\overline{X}_1^{M}(\sigma), \dots, X_n^{M}(\sigma)$ be independent random variables whose distribution is the same as $\overline{X}^{M}(\sigma)$ defined in the previous proof. Likewise, $X_1^{M'}(\sigma), \dots, X_n^{M'}(\sigma)$ are independent random variables identical in distribution to $\overline{X}^{M'}(\sigma)$. Then $\sum_{i=1}^{n} w_i \cdot \overline{X}_i^{M'}(\sigma)$ is a mean-preserving spread of the distribution of $\sum_{i=1}^{n} w_i \cdot \overline{X}_i^{M}(\sigma)$.*

This follows because mps is preserved under summation. Now we will prove the proposition. Fix any n. Then

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}\sim\mu^n}[R(\Sigma^*, \mathbf{M})] &= \frac{1}{n}\mathbb{E}_{\mathbf{M}^n\sim\mu^n}\left[\mathbb{E}_{w\sim\{\pm1\}^n}\left[\sup_{\sigma\in\Sigma^*}\sum_{i=1}^{n} w_i \cdot \left(\frac{1}{M_i}\sum_{j=1}^{M_i} l(\sigma, Z_j))\right)\right]\right] \\
&= \frac{1}{n}\sum_{\sigma\in\{\pm1\}^n}\frac{1}{2^n}\cdot\mathbb{E}_{\mathbf{M}^n\sim\mu^n}\left[\sup_{\sigma\in\Sigma^*}\sum_{i=1}^{n} w_i\left(\frac{1}{M_i}\sum_{j=1}^{M_i} l(\sigma, Z_j)\right)\right] \\
&= \frac{1}{n}\sum_{\sigma\in\{\pm1\}^n}\frac{1}{2^n}\cdot\mathbb{E}_{d^n\sim\mu_d^n}\left[\mathbb{E}_{M^n\sim\mu_{m|d}^n}\left[\sup_{\sigma\in\Sigma^*}\sum_{i=1}^{n} w_i\cdot\overline{X}^{M_i}(\sigma)\right]\right] \\
&\leq \frac{1}{n}\sum_{\sigma\in\{\pm1\}^n}\frac{1}{2^n}\cdot\mathbb{E}_{d^n\sim\mu_d^n}\left[\mathbb{E}_{(M')^n\sim(\mu'_{m|d})^n}\left[\sup_{\sigma\in\Sigma^*}\sum_{i=1}^{n} w_i\cdot\overline{X}^{M_i'}(\sigma)\right]\right] \\
&= \mathbb{E}_{\mathbf{M}\sim(\mu')^n}[R(\Sigma^*, \mathbf{M})]
\end{aligned}$$

where the inequality follows from Corollary 1 and convexity of the supremum. □

Proposition 4. *Let n, n' be positive integers satisfying $n > n'$, Then*

$$\mathbb{E}_{\mathbf{M}^n\sim\mu^n}\left[R(\Sigma^*, \mathbf{M}^n)\right] \leq \mathbb{E}_{\mathbf{M}^{n'}\sim\mu^{n'}}\left[R(\Sigma^{\prime}\mathbf{M}^{n'})\right]$$

i.e., expected Rademacher complexity is decreasing in the total number of domains.

Proof. For each strategy σ, define the random variable

$$X_i(\sigma) \equiv w_i \cdot \left(\frac{1}{M_i} \sum_{j=1}^{M_i} \ell(\sigma, Z_j) \right)$$

where w_i is drawn from a uniform distribution on $\{-1, +1\}$, and independently a domain $d_i \sim \mu_d$ is drawn, $M_i \sim \mu_{m|d}$, and Z_j are drawn iid from the distribution P_{d_i}.

Let $\overline{X}^m(\sigma) \equiv \frac{1}{n} \sum_{i=1}^{n} X_i(\sigma)$ for every $n \in \mathbb{Z}_+$. Then by Lemma 1, $\overline{X}^{n'}(\sigma)$ is a mps of $\overline{X}^n(\sigma)$ for every $n > n'$ and every h. So

$$\begin{aligned}
\mathbb{E}_{\mathbf{M}^n \sim \mu^n}[R(\Sigma^*, \mathbf{M}^n)] &= \frac{1}{n} \mathbb{E}_{\mathbf{M}^n \sim \mu^n} \left[\mathbb{E}_{w \sim \{\pm 1\}^n} \left[\sup_{\sigma \in \Sigma^*} \sum_{i=1}^{n} w_i \cdot \left(\frac{1}{M_i} \sum_{j=1}^{M_i} l(\sigma, Z_j)) \right) \right] \right] \\
&= \mathbb{E}_{\mathbf{M}^n \sim \mu^n} \left[\mathbb{E}_{w \sim \{\pm 1\}^n} \left(\sup_{\sigma \in \Sigma^*} \sum_{i=1}^{n} \overline{X}^n(\sigma) \right) \right] \\
&\leq \mathbb{E}_{\mathbf{M}^{n'} \sim \mu^{n'}} \left[\mathbb{E}_{w \sim \{\pm 1\}^{n'}} \left(\sup_{\sigma \in \Sigma^*} \sum_{i=1}^{n'} \overline{X}^{n'}(\sigma) \right) \right] \\
&= \mathbb{E}_{\mathbf{M}^{n'} \sim \mu^{n'}}[R(\Sigma^*, \mathbf{M}^{n'})]
\end{aligned}$$

where the inequality follows from convexity of the supremum. □