The Transfer Performance of Economic Models*

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

July 28, 2022

Abstract

Economists often estimate models using data from a particular setting, e.g. estimating risk preferences in a specific subject pool. Whether a model's predictions extrapolate well across settings depends on whether the estimated model has captured generalizable structure. We provide a tractable formulation for this "out-of-domain" prediction problem, and define the *transfer error* of a model to be its performance on data from a new domain. We derive finite-sample *forecast intervals* that are guaranteed to cover realized transfer errors with a user-selected probability when domains are iid, and use these intervals to compare the transferability of economic models and black box algorithms for predicting certainty equivalents. We find that in this application, black box algorithms outperform the economic models when estimated and tested on different data from the same domain, but models motivated by economic theory generalize across domains better than the black-box algorithms do.

*We thank Andrei Iakovlev, Jiafeng Chen, Nic Fishman, Brandon Kaplowitz, Kyohei Okumura, Michael Ostrovsky, Ashesh Rambachan, Bas Sanders, Jesse Shapiro, and Davide Viviano for helpful comments. 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

1 Introduction

When we estimate models on data, we often hope that the estimated model will be useful for making predictions in domains beyond the specific context 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 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 domains depends on whether the estimated model captures generalizable structure. Understanding out-of-domain predictive performance is especially important now that black-box machine learning methods (such as neural net and random forest algorithms) are increasingly popular in many fields. 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 domains, but so far there has been little empirical validation of that belief: It is an open question whether economic models indeed generalize across domains better than black boxes do.

This paper provides a tractable framework for evaluating cross-domain transfer performance, defines measures for model transferability, and provides *forecast intervals* for transferability that cover realized transfer errors with prescribed probability in finite samples.[1] We apply our proposed framework and methods to evaluate the transferability of models of risk preferences, and find that in this application economic models transfer more reliably than two popular black box algorithms.

Our conceptual framework, described in Section 2, is an extension of the usual "out-of-sample" evaluation to "out-of-domain" evaluation. In the standard out-of-sample test, a model's free parameters are estimated on a training sample, and the predictions of the estimated model are evaluated on a test sample, where the observations in the training and test samples are drawn from the same distribution. We depart from this framework by supposing that the distribution of the data varies across a set of "domains," but that these domain-specific distributions are themselves drawn iid. While this assumption is restrictive, and rules out some interesting prediction problems, we view it as a useful first step which yields easy-to-apply procedures.

[1] We use the term "forecast interval," rather than "confidence interval," to reflect the random nature of target, namely the *realized* (rather than expected, median, etc.) transfer error.

We consider several questions about how well models transfer across domains, and define corresponding measures for each question. First, we ask how well the model will predict in a sample from an as-yet unobserved target domain. We call this the model's *transfer error*. Since the size of the raw transfer error can be difficult to interpret, we define a model's *normalized transfer error* as the ratio of its transfer error to a proxy for the best achievable error on the sample from the target domain. Finally, we ask how much is lost by transferring a model across domains instead of re-estimating the model's parameters on the new domain of interest. 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 new data.

In Section 3, we demonstrate how to construct forecast intervals with guaranteed coverage probability for these measures, using a meta-data set of samples from already observed domains. Specifically, we suppose that a model is estimated on samples from a random set of domains, and tested on a new sample from a new domain. To construct a forecast interval for the estimated model's transfer error (for instance) on the new domain, we split the observed domains in the meta-data set into training and test domains. We estimate the parameters of the model on the samples from the training domains and evaluate its transfer error on each of the test domains. Pooling these transfer errors across different choices of training and test domains gives an empirical distribution of transfer errors. We show that for any quantile τ, the interval bounded by the τ-th and $(1-\tau)$-th quantiles of the pooled transfer error is a valid forecast interval for the transfer error on a new, unseen domain. The same method yields forecast intervals for our other two measures as well.

Section 4 uses these results to evaluate the transferability of predictions of certainty equivalents for binary lotteries. The samples correspond to observations from different subject pools, so a model's transfer error is how well it predicts outcomes in one subject pool when estimated on data from another. We evaluate two models of risk preferences, expected utility and cumulative prospect theory, and two popular black box machine learning algorithms, random forest and kernel regression. We first consider a standard out-of-sample test, where the training and test data are drawn from the same subject pool. We show that the black box algorithms slightly outperform the economic models out-of-sample for most of the subject pools. This could be because the black box learns general properties of the map from lotteries to certainty equivalents that the economic models miss. Alternatively, it could be that the gains of the black boxes are specific to the within-domain prediction task, and do not correspond to improved generalizability. Our evaluation of transfer performance points to the latter. Our main finding, which is robust across all three of our measures, is that while the forecast intervals for the black box algorithms and economic models overlap, the forecast intervals for the black box methods are wider, and their upper bounds are substantially higher. For example the 5th and 95th percentiles of the pooled CPT normalized transfer errors (which constitute a 71% forecast interval) are 1.02 and 2.62, while the same percentiles for a random forest algorithm are 1.02 and 6.42. Thus, even though the economic models perform slightly less well on most of the subject pools, they generate predictions that generalize more reliably across subject pools.

Why do the black boxes perform worse at transfer prediction in this setting? A natural explanation, based on intuition from conventional out-of-sample testing, is that black boxes are very flexible and hence learn idiosyncratic details that do not generalize across subject pools. But when we restrict the analysis to a subset of our samples involving the same set of lotteries, the resulting forecast intervals are nearly identical across all of the prediction methods. Our findings suggest instead that the relatively worse transfer performance of the black boxes is because the black box algorithms extrapolate less effectively from data at one set of lotteries to data at another. Specifically, black boxes transfer worse when the primary source of variation across samples is a shift in the marginal distribution over features (i.e. which lotteries appear in the sample), rather than a shift in the distribution of outcomes conditional on features (the distribution of certainty equivalents given fixed lotteries).

Related Literature. Our paper is related to the literature on meta-analyses (Benartzi et al., 2017; DellaVigna and Pope, 2019; Hummel and Maedche, 2019; Meager, 2019), although the primary focus in that literature is on how much the estimated coefficients vary across samples, rather than on variation in how well the models predict. The two concerns are related but not the same: A high degree of variability in the parameters of a given model across domains suggests that the transfer deterioration of the model will be high (since we obtain very different estimates by re-fitting the model), but transfer errors and normalized transfer errors for a given model can be large even if the parameters are stable across domains (for instance because the model is unable to capture structure present in some or all domains).

Our paper is also connected to *domain generalization* in computer science (as introduced in Blanchard et al. (2011) and Muandet et al. (2013)), much of which aims to find models that generalize well to new unseen domains (see Zhou et al. (2021) for a recent survey).[2] This domain generalization literature develops algorithms for achieving good out-of-domain generalization; in contrast we develop forecast intervals for the out-of-domain generalization error of a given algorithm, allowing cross-model comparisons as in our application.

Our paper contributes to a recent literature comparing the predictiveness of black box algorithms with that of more structured economic models, e.g. Noti et al. (2016), Plonsky et al. (2017), Plonsky et al. (2019), Camerer et al. (2019), Fudenberg and Liang (2019), and Ke et al. (2020). 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 they may be less effective for transfer prediction.[3]

Finally, our forecast intervals are closely related to recent work in statistics and machine learning on conformal inference, where we replace that literature's assumption of exchangeability across observations with one of exchangeability across domains. A version of our forecast intervals which uses a random choice of training domain can be viewed as a split conformal interval (see Lei et al., 2018), while our preferred forecast intervals eliminate the random choice of training domain through an averaging argument as in Rüschendorf (1982), Meng (1994), and more recently Vovk and Wang (2020).

2 Framework

2.1 Statistical model

There is a (random) feature or covariate vector X taking values in the set $\mathcal{X}$, and a (random) outcome Y taking values in the set $\mathcal{Y}$. The analyst predicts Y given X using a prediction rule $\sigma : \mathcal{X} \to \mathcal{Y}$. (Below we discuss how this rule is chosen.) The prediction rule is evaluated

[2] 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. There is also a related but distinct literature on *domain adaptation*, which aims to improve predictions when some data from the target domain is available – see Zhou et al. (2021).

[3] A recent paper examining the out-of-domain performance of economic models is Külpmann and Kuzmics (2022), which estimates various game-theoretic models on 2×2 normal-form games and evaluates their performance on 3×3 normal-form games.

on a sample S of observations (x, y) using a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

Definition 1. *The error of prediction rule σ on sample S is*

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

i.e., the average loss when using σ to predict y given x.

In a conventional out-of-sample test, the analyst's choice of prediction rule σ is based on a training sample S_{train} of observations (x, y) drawn iid from a distribution $P \in \Delta(\mathcal{X} \times \mathcal{Y})$, and the chosen prediction rule is then evaluated on a test sample S_{test} of observations drawn iid from the same distribution P.

We extend the standard framework by supposing that the analyst observes samples S_d from multiple domains $d \in \{1, ..., n\}$. For each observed domain d, a sample of m_d observations (x, y) is drawn iid from a distribution P_d, where P_d and m_d are themselves drawn from an unknown *meta-distribution* $\mu \in \Delta(\mathcal{P} \times \mathbb{N})$ over the set of distributions $\mathcal{P} \equiv \Delta(\mathcal{X} \times \mathcal{Y})$ and the set of sample sizes $\mathbb{N}$, where each sample S_d is generated by first drawing a distribution and sample size $(P_d, m_d) \sim \mu$, and then drawing m_d observations $(x, y) \sim_{iid} P$.[4] In an abuse of notation, we will simply write $S_d \sim \mu$ for a sample generated in this way. This reduces to iid sampling of observations if μ assigns probability one to a single distribution $\mathcal{P}$, or assigns probability one to the sample size $m = 1$, but we are primarily interested in settings where neither of these is the case.

The analyst has access to *meta-data* consisting of n samples drawn independently from μ

$$\mathbf{M} = (S_1, \ldots, S_n).^5$$

The analyst chooses a prediction rule σ using a *decision rule*, which is a (potentially randomized) map ρ from $\mathcal{M}$, the set of all meta-data of finite length, to Σ, the set of all prediction rules. As we discuss below, in some cases it may be useful for the analyst to base this choice on a strict subset of the observed samples. We thus use $\mathbf{M}_T \subseteq \mathbf{M}$, for $T \subseteq [n] \equiv \{1, \ldots, n\}$, to denote the *training meta-data*, and use $\rho(\mathbf{M}_T) \in \Sigma$ for the chosen prediction rule.

[4] Section 2.3 discusses the iid assumption. We don't require that P_d have support $\mathcal{X} \times \mathcal{Y}$, and in particular allow the support of X to vary across domains.

[5] The domains are distinguished in this vector. So, for example, the meta-data 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).

We focus on two classes of decision rules ρ:

Example 1 (Estimation of an Economic Model). Consider a set of prediction rules Σ^* derived from an economic model (see examples in Section 4.2).[6] Let ρ_{Σ^*} be a decision rule satisfying

$$\rho_{\Sigma^*}(\mathbf{M}_T) \in \operatorname*{argmin}_{\sigma \in \Sigma^*} \frac{1}{|\mathbf{M}_T|} \sum_{S \in \mathbf{M}_T} e(\sigma, S) \quad \forall \mathbf{M}_T \in \mathcal{M}. \tag{1}$$

That is, the training data $\mathbf{M}_T$ is mapped into a prediction rule in Σ^* that minimizes the average error across the training samples. This procedure is known as empirical risk minimization, and when Σ^* is parameterized, it corresponds to estimating the model's free parameters on the training data.

Example 2 (Training a Black Box Algorithm). The analyst chooses a machine learning algorithm, which takes the meta-data as input, and outputs a prediction rule σ. We use "black box" as a catchall term for flexible prediction methods which are not based on explicit economic structure; the specific black box methods we consider in our application are random forests and kernel regression.

2.2 Measures of transferability

We are interested in the performance of the prediction rule $\rho(\mathbf{M}_T)$ on a new target sample $S_{n+1} \sim \mu$. Below we propose several measures for the prediction rule's transferability.

First, the analyst may be interested in the raw error of $\rho(\mathbf{M}_T)$ on the new sample.

Definition 2 (Transfer Error). *The* transfer error *of prediction rule $\rho(\mathbf{M}_T)$ on target sample S_{n+1} is $e(\rho(\mathbf{M}_T), S_{n+1})$.*

This raw transfer error depends on the predictability of the target sample: If the outcomes y in the target sample can only be poorly predicted using the features x, the lowest achievable error may be large even given perfect knowledge of the distribution in that domain. This lowest achievable error may also differ across domains, so it can potentially mask important differences across models. We thus propose normalizing the transfer error by a proxy for the best achievable error in the test domain. Specifically, we normalize transfer errors by the smallest in-sample error achieved by decision rules within a pre-specified set R.[7]

[6] We assume that Σ^* is compact in a topology that makes $e(\sigma, S)$ continuous in σ for all S.

[7] This measure is similar to the "completeness" measure introduced in Fudenberg et al. (2022), without

Definition 3 (Normalized Transfer Error)**.** *Fix a finite set of decision rules R with $\rho \in R$. The* normalized transfer error *of prediction rule $\rho(\mathbf{M}_T)$ on test sample S_{n+1} is*

$$\frac{e(\rho(\mathbf{M}_T), S_{n+1})}{\min_{\widetilde{\rho} \in R} e(\widetilde{\rho}(S_{n+1}), S_{n+1})}.$$

This quantity tells us how many times larger the transfer error of the prediction rule $\rho(\mathbf{M}_T)$ is than the best in-sample error achievable by a decision rule from R. For rules used in practice, we would expect the normalized transfer error to be bounded below by 1.[8,9]

Our third and final measure evaluates a different kind of transferability, namely how much is lost by estimating a model on other samples instead of re-training it on the sample of interest.

Definition 4 (Transfer Deterioration)**.** *The* transfer deterioration *of prediction rule $\rho(\mathbf{M}_T)$ on target sample S_{n+1} is*

$$\frac{e(\rho(\mathbf{M}_T), S_{n+1})}{e(\rho(S_{n+1}), S_{n+1})}.$$

This quantity tells us how many times larger a decision rule's transfer error is than its in-sample error in the test sample, i.e., $e(\rho(S_{n+1}), S_{n+1})$. We again expect this ratio to be bounded below by 1 for rules used in practice (see footnote 8); it is equal to 1 only if the transfer error of the model is the same as its in-sample error.[10] The larger the transfer deterioration of the decision rule, the more valuable it is to re-train the model on the new domain instead of transferring parameters estimated from other domains.

2.3 Discussion

IID sampling. Our approach assumes that the distributions governing the different samples S_i are themselves independent and identically distributed. This assumption is not

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.

[8] When R consists of empirical risk minimization rules (see Example 1), the normalized transfer error is bounded below by 1 by construction, and a model achieves this lower bound only if the transfer error is as low as the best in-sample error over R. Normalized transfer error is not bounded below by 1 for all decision rules, for example a decision rule that maps all meta-data to a uniform draw from Σ. We do not expect such rules to be used in practice.

[9] 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 $\rho(\mathbf{M}_T)$. An alternative measure in the same spirit would divide through by a cross-validated error, where the training and test data are drawn from the same domain.

[10] This does not require $\mathbf{M}_T$ and S_{n+1} to be identical: it is sufficient for the training data $\mathbf{M}_T$ and the sample S_{n+1} to lead to the same estimates for model parameters.

always appropriate: For instance, it would fail in time-series settings (where each domain corresponds to one time period) if there is serial correlation across periods. Similarly, if we consider an experiment conducted at multiple sites, an assumption that the distributions governing these different sites are iid may fail if there is site selection bias.[11]

At the same time, some assumption on how the unobserved domains relate to the observed domains is necessary to make progress on this problem.[12] The assumption that domain-specific distributions are themselves drawn iid lets us apply all the implications of random sampling, particularly joint exchangeability of the data from the observed and unobserved domains. As we show in the subsequent section, this assumption allows us to construct valid forecast intervals for transfer performance, without need of any further restrictions on (or knowledge about) the meta-distribution μ, which makes the assumption an effective starting point.

What is a domain? The choice of how to define the set of domains determines the transfer question the analyst is interested in, and also the content of the iid sampling assumption. Suppose, for instance, that our meta-data consists of experimental results from multiple papers, where each paper reports results from experiments at multiple sites. If we specify that each site is a separate domain, iid sampling corresponds to drawing further sites, some from the papers already observed and others potentially from as-yet-unobserved papers. Alternatively, if we define each paper as a separate domain, iid sampling corresponds to drawing new papers, each with its own sites.

In principle, the "more different" the distributions are that govern the different samples, the worse performance will be by all of our measures. In the edge case where each sample is governed by the same distribution P on $\mathcal{X} \times \mathcal{Y}$, transfer prediction reduces to out-of-sample prediction for iid data. Consequently, performance for out-of-domain prediction will mirror performance for out-of-sample, but in-domain, prediction, up to differences in sample size. We confirm this in Online Appendix C.2, where we construct artificial domains by randomly assigning observations to samples, using data from our subsequent application.

[11] One way this selection bias can arise is if the observed sites were chosen based on characteristics which are correlated with effect sizes, as Allcott (2015) found in the Opower energy conservation experiments.

[12] Certain generalizations to non-iid cases are immediate from our results. Suppose, for instance, that the iid assumption is "nearly true" in a given application, in the sense that the joint distribution across the observed and unobserved domains is close (in total variation distance) to a distribution arising from iid sampling. In this case the coverage probabilities of our forecast intervals are reduced by at most the total variation distance between the true distribution and the "closest" iid distribution.

Relationship between measures. On any fixed target sample, transfer error and normalized transfer error generate the same ranking of models, since the denominator of normalized transfer error is model-independent. In contrast, when performance is aggregated across multiple target samples, transfer error and normalized transfer error will typically lead to different rankings, since normalized transfer error penalizes a model more for performing worse on predictable samples (i.e., samples where the best achievable error is lower) than on samples that are hard to predict, while transfer error does not.

Transfer deterioration is designed to address a different question than the other measures, and a ranking of models by their transfer deterioration need not coincide with a ranking of models using either transfer error or normalized transfer error. For example, a model that achieves approximately constant but large errors across samples would have low transfer deterioration (showing that retraining would not be worthwhile) but high transfer error.

Evaluating transfer performance versus learning a "best" cross-domain prediction rule. Economists often use estimated parameter values from one domain to make predictions and inform policy decisions in another. For example, we might have data on prices, demand, demographics, and tax revenue under a particular tax rate, and use a structural model to extrapolate to revenue under another tax rate.[13] If, moreover, we had data on how well such extrapolation exercises had performed in the past, we could under suitable assumptions use this past performance to predict how well they are likely to perform in the future. The next section formalizes forecast intervals for this future performance.

A complementary question is how to estimate a "best" prediction rule for transfer. One might consider, for example, using cross-validation across domains to tune parameters, building a Bayesian hierarchical model that explicitly models a distribution over domains, or choosing parameters with worst-case guarantees as in the literature on distributionally robust optimization (Rahimian and Mehrotra, 2019). The development of estimation procedures that lead to better transfer performance is an interesting direction, but one that we view as distinct from our goal of providing forecast intervals for transfer performance. For this reason, in our subsequent application we examine the transfer performance of estimation procedures that have previously been used in the literature. We note, however, that each of the approaches mentioned above can be formalized as a decision rule (see Appendix A for

[13]For example, the model may impose structural assumptions about how consumer responses to (as yet unobserved) tax changes relate to their responses to (observed) price changes due to other factors.

details), and hence our subsequent results will also imply forecast intervals for their transfer performance.

3 Main results

Recall that we select a subset of domains indexed by $T \subset [n]$ to use as 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*, i.e., the samples from domains in T. The remaining domains $[n] \backslash T$ are used as test domains, and we again use S_{n+1} to denote the (unobserved) target sample.

Let $e_{T,d}$ be any random variable that can be expressed as a function of the random meta-data $\mathbf{M}_T$ (where $T \subseteq \{1, \ldots, n\}$), the sample S_d (where $d \in \{1, \ldots, n+1\}$), and (possibly) an independent random variable ξ. This includes our three measures as special cases:

Example 3 (Transfer Error). For any decision rule ρ, set $e_{T,d} \equiv e(\rho(\mathbf{M}_T), S_d)$ to be the transfer error of prediction rule $\rho(\mathbf{M}_T)$ on the sample S_d.

Example 4 (Normalized Transfer Error). For any decision rule ρ and set of decision rules R, set $e_{T,d} \equiv \frac{e(\rho(\mathbf{M}_T), S_d)}{\min_{\rho \in R} e(\rho(S_d), S_d)}$ to be the normalized transfer error of prediction rule $\rho(\mathbf{M}_T)$ on the sample S_d.

Example 5 (Transfer Deterioration). For any decision rule ρ, set $e_{T,d} \equiv \frac{e(\rho(\mathbf{M}_T), S_d)}{e(\rho(S_d), S_d)}$ to be the transfer deterioration of prediction rule $\rho(\mathbf{M}_T)$ on the sample S_d.

We will focus on these three measures, but our subsequent results also apply to other specifications for $e_{T,d}$, such as the following.

Example 6 (Ratios of Model Errors). For any two decision rules ρ and ρ', set

$$e_{T,d} \equiv \frac{e(\rho(\mathbf{M}_T), S_d)}{e(\rho'(\mathbf{M}_T), S_d)}$$

to be the ratio of the transfer errors of the two decision rules. (Or similarly, set $e_{T,d}$ to be the ratio of the transfer deterioration of the two decision rules.)

Example 7 (Partial Transfer). Choose an economic model that can be represented as $\Sigma^* = \{\sigma_{\theta,\lambda}\}_{\theta \in \Theta, \lambda \in \Lambda}$ for compact sets Θ and Λ. The parameters in Λ are domain-specific parameters that will be re-estimated on the target sample, while the parameters in Θ are transferred across domains. 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 $e(\theta_{\mathbf{M}_T}, S)$, where

$$\theta_{\mathbf{M}_T} \in \operatorname*{argmin}_{\theta \in \Theta} \sum_{S \in \mathbf{M}_T} \min_{\lambda \in \Lambda} e(\sigma_{\theta,\lambda}, S)$$

is a best-fitting value of θ on the training data $\mathbf{M}_T$.[14] Set $e_{T,d} \equiv e(\theta_{\mathbf{M}_T}, S_d)$ to be the partial transfer error on sample S_d.

For ease of exposition we will simply refer to $e_{T,d}$ as transfer error in what follows, understanding that the subsequent results apply to all of the above measures.

Our goal in this section is to develop forecast intervals for the transfer error $e_{T,n+1}$ on the target sample, that is, interval-valued functions of the (observed) meta-data $(S_1, \dots, S_n)$ which cover $e_{T,n+1}$ with prescribed probability, regardless of the unknown distribution μ across domains. Since we expect that in many settings of interest only a limited number of domains will be observed, our focus is on finite-sample results.

It will be useful to define $e_{T,(r)}$ to be the r-th smallest value in $\{e_{T,d} : d \in [n] \setminus T\}$. Since by assumption $S_1, \dots, S_{n+1}$ 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 samples and the target sample) are uniformly distributed, up to ties. Thus we have:

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

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

This implies that $(-\infty, e_{T,(r)}]$ is a level-$\left(\frac{n-n_T}{n-n_T+1}\tau\right)$ forecast 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 Forecast Interval)**.** *For any quantile* $\tau \in (0.5, 1)$*, let* $r_U \equiv \lceil \tau(n - n_T) \rceil$ *and* $r_L \equiv n - n_T + 1 - r_U$*. Then*

$$\mathbb{P}\left(e_{T,n+1} \notin \left[e_{T,(r_L)}, e_{T,(r_U)}\right]\right) \leq 2 - \frac{2\tau(n - n_T)}{n - n_T + 1}.$$

[14] The decision rule ρ should specify a way to break ties.

This implies that $\left[e_{T,(r_L)}, e_{T,(r_U)}\right]$ is a level-$\left(\frac{2\tau(n-n_T)}{n-n_T+1}\right)$ forecast interval for the transfer error of $\rho(\mathbf{M}_T)$ on the target sample S_{n+1}.

Claims 1 and 2 provide forecast intervals for the performance of the particular prediction rule $\rho(\mathbf{M}_T)$, where the set of training domains T is fixed. The resulting intervals are versions of the split conformal intervals considered in the conformal inference literature (see e.g. Lei et al., 2018), with the difference that samples in our setting play the role that observations do in the usual single-domain setup.

These results provide forecast intervals for $e_{T,n+1}$ where T is an arbitrarily chosen set of training domains. But since the training domains are ex-ante symmetric there is no reason to prefer any choice of training domains over another. We thus move from evaluation of the transfer error of a particular prediction rule $\rho(\mathbf{M}_T)$ to 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 uniformly (without replacement) from the set A. We draw the set of training domains T from this distribution, with $k = n_T$ and $A = [n]$, where each realization of T implies a choice of $n - n_T$ distinct test domains. Below, we pool the transfer errors across choices of training and test domains and use the τ-th and $(1-\tau)$-th quantiles of the pooled transfer errors to form a forecast interval for the transfer error given a randomly selected set of training domains.

For any random variable B let

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

and

$$\underline{Q}_\tau(B) = \inf\{b : \mathbb{P}(B \leq b) \geq \tau\}$$

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

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

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

to be the τth upper quantile of the transfer 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* $[n] \setminus \tilde{T}$.[15] *Then*

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

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

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

Thus, $[\underline{e}_\tau, \overline{e}_\tau]$ is a level-$\left(4\left(\frac{n-n_T}{n-n_T+1}\tau\right) - 3\right)$ forecast interval for the transfer error on the target sample. The proof of this result relies on Meng (1994), which shows how to construct valid (if potentially conservative) tests based on random variables which are second-order stochastically dominated by a $U[0,1]$ distribution. To prove Proposition 1 we show how to construct a $U[0,1]$ test statistic for each choice of training domain, and then use Meng (1994)'s result to establish validity for a test based on the average.

In this result $e_{\widetilde{T},n+1}$ is a random variable whose value depends on the realizations of training samples $\{S_t : t \in \widetilde{T}\}$ and the target sample $n+1$.

The meta-data involve both a finite number of observed domains and a finite number of observations per domain. These two sources of finiteness enter into our results in different ways. Increasing the number of observations per domain changes the distribution of $e_{\widetilde{T},n+1}$: In the limit of infinitely many observations per domain, the error $e_{\widetilde{T},n+1}$ measures how well the best predictor from the model class in the training domains transfers across domains, while if the number of observations is small, the error $e_{\widetilde{T},n+1}$ measures how well an imperfectly estimated model transfers. In contrast, increasing the number of observed domains (holding fixed the distribution over sample sizes within each domain) does not change the distribution of $e_{\widetilde{T},n+1}$, but instead allows this distribution to be estimated more precisely.

The parameters n_T and τ are choice variables for the analyst. The size of τ influences the width of the forecast interval, where larger choices of τ lead to wider forecast intervals with higher confidence guarantees. The choice of n_T determines how many samples in the meta-data are used for training versus testing. Larger choices of n_T mean that the model

[15] We write $\overline{Q}_{\tau,(\widetilde{T},t)\sim\phi(n_T+1,[n])}\left(e_{\widetilde{T},t}\right)$, rather than $\overline{Q}_\tau\left(e_{\widetilde{T},t}\right)$, to emphasize that the meta-data $\mathbf{M}$ are held fixed, so the only source of randomness is drawing $(\widetilde{T},t) \sim \phi(n_T+1,[n])$.

will be estimated on a larger quantity of data, but we will have fewer samples on which to evaluate the performance of the estimated model.

4 Application

As an illustration of our methods, 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-data set consisting of samples of reported certainty equivalents from 44 subject pools (see Section 4.1). We consider decision rules based on various specifications of expected utility and cumulative prospect theory, as well decision rules based on two black box algorithms.

Section 4.1 describes our data in more detail, and Section 4.2 describes the decision rules. In Section 4.3, we first conduct "within-domain" out-of-sample tests, where the training and test data are drawn from the same domain. The out-of-sample errors of the black box algorithms are lower than those of the economic models in most of our domains. Section 4.4 constructs forecast intervals for transfer error, normalized transfer error, and transfer deterioration. The forecast intervals for black box methods are substantially wider and higher relative to the forecast intervals for the economic models for all three measures. The following subsections examine why it is the case that the black box algorithms predict better than the economic models out-of-sample on most domains, but poorly out-of-domain.

4.1 Data

We have constructed a meta-data set including samples of certainty equivalents from 44 subject pools, which we treat as the domains. These data are drawn from 14 papers in experimental economics, with twelve papers contributing one sample each, one paper contributing two, and a final paper (a study of risk preferences across countries) contributing 30 samples. In Online Appendix C.3, we repeat our analysis with each paper treated as a separate domain, and show that the results are qualitatively similar.[16] Our samples range in size from 72 observations to 8906 observations, with an average of 2752.7 observations per

[16]In both cases, the domains sometime combines observations from different experimental treatments. 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.

sample.[17] We convert all prizes to dollars using purchasing power parity exchange rates in the year of the paper's publication from OECD (2022).

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 (and we use the convention that always $|\overline{z}| > |\underline{z}|$), p is the probability of the larger prize, and y is the reported certainty equivalent by a given subject. Thus our feature space is $\mathcal{X} = \mathbb{R} \times \mathbb{R} \times [0, 1]$, the outcome space is $\mathcal{Y} = \mathbb{R}$, and a prediction rule is any mapping from binary lotteries into predictions of the reported certainty equivalent. We use squared-error loss $\ell(y, y') = (y - y')^2$ to evaluate the error of the prediction, 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.[18] Since different subjects report different certainty equivalents for the same lottery, the best achievable error on a sample is generally bounded away from zero.

4.2 Models and black boxes

We consider several possible decision rules ρ.

Expected Utility. First we consider an expected utility agent with a CRRA utility function parametrized by $\eta \geq 0$. For $\eta \neq 1$, define

$$v_\eta(z) = \begin{cases} \frac{z^{1-\eta}-1}{1-\eta} & \text{if } z \geq 0 \\ -\frac{(-z)^{1-\eta}-1}{1-\eta} & \text{if } z < 0 \end{cases}$$

and for $\eta = 1$, set $v_\eta(z) = \ln(z)$ for positive prizes and $v_\eta(z) = -\ln(-z)$ for negative prizes. For each $\eta \geq 0$, define the prediction rule σ_η to be

$$\sigma_\eta(\overline{z}, \underline{z}, p) = v_\eta^{-1}\big(p \cdot v_\eta(\overline{z}) + (1 - p) \cdot v_\eta(\underline{z})\big).$$

That is, the prediction rule σ_η maps each lottery into the predicted certainty equivalent for an EU agent with utility function v_η. The corresponding decision rule ρ^{EU} maps each meta-data realization $\mathbf{M}_T$ to the prediction rule in $\Sigma_{EU} \equiv \{\sigma_\eta : \eta \geq 0\}$ that minimizes the error on the meta-data, as described in (1).

[17] Online Appendix C.1 describes our data sources in more detail.

[18] This transformation is possible because none of the results in this paper change if we redefine $e(\sigma, S) = g(\frac{1}{|S|} \sum_{(x,y)\in S} \ell(\sigma(x), y))$ for any function g. Root-mean-squared error corresponds to setting $g(x) = \sqrt{x}$.

Cumulative Prospect Theory. Next we consider the set of prediction rules Σ_{CPT} derived from the parametric form of Cumulative Prospect Theory (CPT) first proposed by Goldstein and Einhorn (1987) and Lattimore et al. (1992). Fixing values for the model's parameters $(\alpha, \beta, \delta, \gamma)$, each lottery $(\overline{z}, \underline{z}, p)$ is assigned a utility

$$w(p)v(\overline{z}) + (1 - w(p))v(\underline{z}) \tag{2}$$

where

$$v(z) = \begin{cases} z^\alpha & \text{if } z \geq 0 \\ -(-z)^\beta & \text{if } z < 0 \end{cases} \tag{3}$$

is a value function for money,

$$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma} \tag{4}$$

is a probability weighting function, and by convention $|\overline{z}| > |\underline{z}|$.

For each $\alpha, \beta, \gamma, \delta$, the prediction rule $\sigma_{(\alpha,\beta,\gamma,\delta)}$ is defined as

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

That is, the prediction rule 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 prediction rules to be $\Sigma_{CPT} \equiv \{\sigma_\theta\}_{\theta \in \Theta}$. The corresponding decision rule ρ^{CPT} maps each meta-data realization $\mathbf{M}_T$ to the prediction rule in Σ_{CPT} that minimizes the error on the training data, as described in (1).

We also evaluate prediction rules 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 (3); 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 who exhibits nonlinear probability weighting.[19] Additionally, we include CPT with the single free parameter γ (setting $\alpha = \beta = \delta = 1$), which Fudenberg et al. (2021) found to be an especially effective

[19]See Fehr-Duda and Epper (2012) for further discussion of these different parametric forms, and some non-nested versions that have been used in the literature.

one-parameter specification.

Black Boxes. Finally, we consider decision rules ρ corresponding to two machine learning algorithms. First, we train a *random forest*, which is an ensemble learning method consisting of a collection of decision trees.[20] Second, we train a *kernelized ridge regression* model, which modifies OLS to weight observations at nearby covariate vectors more heavily, and additionally places a penalty term on the size of the coefficients. Specifically, we use the RBF kernel $\kappa(x, \tilde{x}) = e^{-\gamma \|x-\tilde{x}\|_2^2}$ to assess the similarity between covariate vectors x and $\tilde{x}$. Given training data $\{(x_i, y_i)\}_{i=1}^N$, the estimated weight vector is $\mathbf{w} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1}\mathbf{y}$, where $\mathbf{K}$ is the $N \times N$ matrix whose (i, j)-th entry is $\kappa(x_i, x_j)$, $\mathbf{I}_N$ is the $N \times N$ identity matrix, and $\mathbf{y} = (y_1, \ldots, y_N)'$ is the vector of observed outcomes in the training data. (Following Pedregosa et al. (2011), we set $\lambda = 1$ and $\gamma = 1/(\#\text{covariates}) = 1/3$.) The estimated prediction rule is $\sigma(x) = \sum_{i=1}^N w_i \kappa(x, x_i)$.[21]

4.3 Within-domain performance

We first evaluate how these models perform when trained and evaluated on data from the same subject pool. We compute the tenfold cross-validated out-of-sample error for each decision rule in each of the 44 domains.[22] We find that the two black box methods (random forest and kernel regression) each achieve lower cross-validated error than EU and CPT in 38 of the 44 domains, although the improvement is not large. Figure 1 reports the CDF of tenfold cross-validated errors for the random forest, kernel regression, EU, and CPT.[23]

[20]A decision tree 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. Each decision tree is trained with a sample (of equal size to our training data) drawn with replacement from the actual training data. At each decision node, the tree splits the training samples into two groups using a True/False question about the value of some feature, where the decision split is chosen to greedily minimize mean squared error. This process stops when a pre-set criterion, such as maximum depth, is met.

[21]See Chapter 14 of Murphy (2012) for further reference.

[22]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.

[23]As we show in Online Appendix C.4, the CDFs for the in-sample errors for these prediction methods are likewise very close.

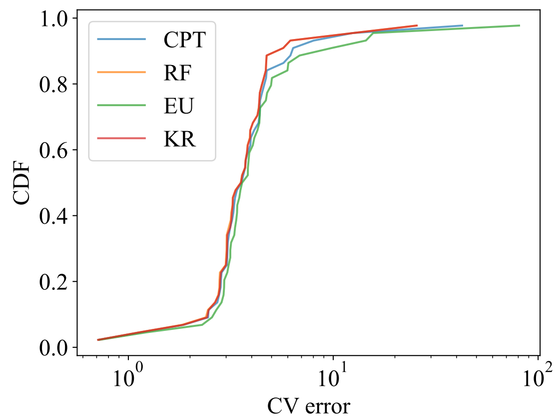


Figure 1: *CDF of Cross-Validated Errors. (The CDFs for kernel regression and random forest overlap.)*

To obtain simpler summary statistics for the comparison between the economic models and black boxes, we normalize each economic model's error (in each domain) by the random forest error. Table 1 averages this ratio across domains and shows that on average, the cross-validated errors of the economic models are slightly larger than the random forest error: the CPT error is on average 1.06 times the random forest error, and the EU error is on average 1.21 times the random forest error.[24]

Model	Normalized Error
EU	1.21
CPT variants	
γ	1.12
α, β	1.22
δ, γ	1.08
α, β, γ	1.07
$\alpha, \beta, \delta, \gamma$	1.06

Table 1: *Average ratio of out-of-sample errors*

These results suggest that from the perspective of within-domain prediction, the different prediction methods we consider are comparable, with the black boxes performing slightly better. But these results do not distinguish whether the economic models and black boxes achieve similar out-of-sample errors by selecting approximately the same prediction rules, or if the rules they select lead to substantially different predictions out-of-domain. We also

[24]The numbers reported in Table 1 would be nearly identical if we normalized by the kernel regression error instead.

cannot determine whether the slightly better within-domain performance of the black box algorithms is achieved by learning generalizable structure that the economic models miss, or if the gains of the black boxes are confined to the domains on which they were trained. We next separate these explanations by evaluating the transfer performance of the models.

4.4 Transfer performance

We use our results in Section 3 to construct forecast intervals for transfer error, normalized transfer error, and transfer deterioration for each of the decision rules described above. In our meta-data there are $n = 44$ domains, and we choose $n_T = 1$ of these to use as the training domain. This choice of n_T corresponds to the question, "If I draw one domain at random, and then try to generalize to another domain, how well do I do?" (In Online Appendix C.8 we report results for a different choice of n_T, with very similar results.)

Figure 2 displays two-sided forecast intervals for transfer error, normalized transfer error (where R includes all decision rules shown in the figure), and transfer deterioration. These forecast intervals use $\tau = 0.95$,[25] so the upper bound of the forecast interval is the 95th percentile of the pooled transfer errors (across choices of the training and test domains), and the lower bound of the forecast interval is the 5th percentile of the pooled transfer errors. Applying Proposition 1, these are 71% forecast intervals. Choosing larger τ results in wider forecast intervals that have higher coverage levels, and we report some of these alternative forecast intervals in Online Appendix C.6, including a more traditional 90% forecast interval.

Our main takeaway from Figure 2 is that although the prediction methods we consider are very similar from the perspective of within-domain prediction, they have very different out-of-domain implications. Specifically, the upper bounds for the black box forecast intervals are substantially larger than the upper bounds for the economic model forecast intervals.

Panel (a) of Figure 2 shows that the black box forecast intervals for transfer error have upper bounds that are roughly twice those of the economic models. For the normalized transfer error, which removes the common variation across models that emerges from variation in the predictability of the different target samples, the contrast between the economic models and the black boxes grows larger. Thus, although the economic models and the black box models select prediction rules that are close for the purposes of prediction in the training domain, they sometimes lead to very different predictions in the test domain, and

[25] See Table 4 in Appendix C.5 for the exact numbers.

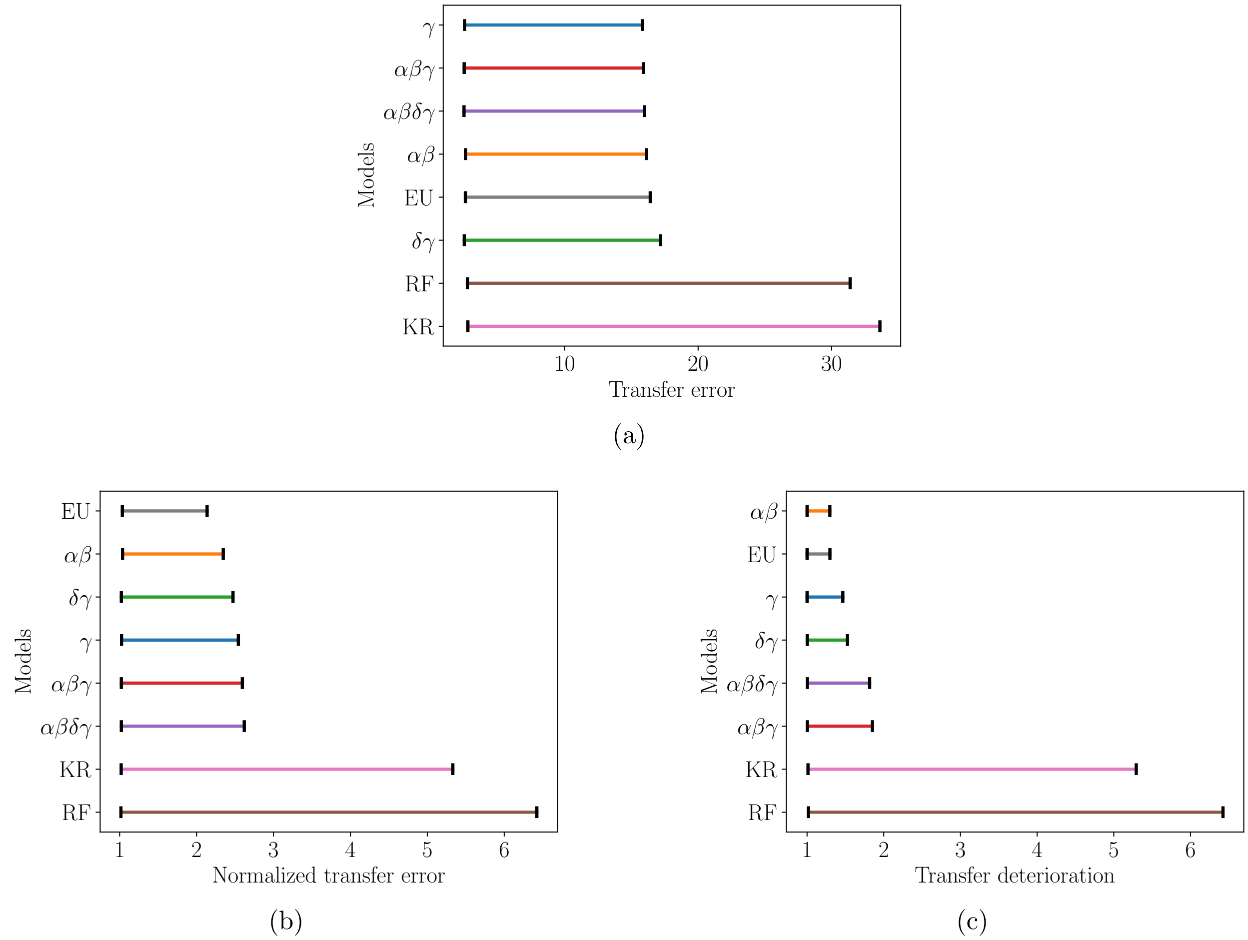


Figure 2: *71% forecast intervals for (a) transfer error, (b) normalized transfer error (with R consisting of the decision rules shown in the figure), and (c) transfer deterioration.*

the prediction rules selected by the economic models generalize substantially better.

Panel (c) of Figure 2 shows an even starker contrast between the black boxes and economic models, which suggests that the value of retraining a black box on the target domain is quite high, while it is less important to re-estimate the economic models. The ordering of the upper bounds of the forecast intervals is roughly consistent with the number of free parameters (with the exception of CPT(α, β), which has lower transfer deterioration than the single-parameter EU model and also CPT(γ)).

All the forecast intervals overlap for each of the three measures. This is not surprising, as variation in the transfer error due to the random selection of training and target domains cannot be eliminated even with data from many domains. We thus expect the black box intervals and the economic model intervals to overlap so long as the economic model errors on “upper tail” training and target domain pairs exceed the black box errors on “lower tail” training and target domain pairs. In Section 5.1 we provide confidence intervals for different quantiles of the transfer error distribution, with similar conclusions.

We provide several robustness checks and complementary analyses in the appendix. Online Appendix C.6 plots the τ-th percentile of pooled transfer errors as τ varies, demonstrating that forecast intervals constructed using other choices of τ (besides $\tau = 0.95$) would look similar to those shown in the main text. Online Appendix C.7 provides 71% forecast intervals for the ratio of the CPT error to the random forest error, and finds that the random forest error is sometimes much higher than the CPT error, but is rarely much worse. Online Appendix C.8 considers an alternative choice for the number of training domains, setting $n_T = 3$ instead of $n_T = 1$. While the results are similar, the contrast between the economic models and black boxes is not as large, suggesting that the relative performance of the black boxes improves given a larger number of training domains. Finally, Online Appendix C.3 provides forecast intervals when each of the 14 papers is treated as a different domain. Again we find that the black box methods transfer worse.

4.5 Do black boxes transfer poorly because they are too flexible?

One possible explanation of our empirical findings is that the black boxes are more flexible than the economic models, and thus learn idiosyncratic details that do not generalize across subject pools. For example, suppose Chinese subjects overvalue lotteries that contain the number ‘8’—the economic models we consider cannot learn this regularity, but the black box methods can. This would lead the black boxes to have better within-domain prediction for the Chinese data, but worse transfer performance if the regularity does not generalize across subject pools. While the high flexibility of black box algorithms is likely an important determinant of their transfer performance, there are at least two reasons this cannot be a complete explanation of our results.

First, the flexibility gap between the black boxes and economic models is not large: many conditional mean functions (for binary lotteries) can be well approximated by CPT for some choice of parameters values $\alpha, \beta, \delta, \gamma$ (Fudenberg et al., 2021).

Second, black boxes do not always transfer worse. One of the papers we use is a study consisting of samples of certainty equivalents from 30 countries (l’Haridon and Vieider, 2019). Crucially, of the 30 samples from this paper, 29 samples report certainty equivalents for the same 27 lotteries, and the remaining sample reports certainty equivalents for 23 of those lotteries. We repeat our analysis using these 30 samples as the domains, and find that the forecast intervals for transfer error are indistinguishable across the prediction methods (Panel (a) of Figure 3). There is some separation between the forecast intervals for the remaining two measures, but in both cases the CPT and random forest forecast intervals are more similar than in the original data.

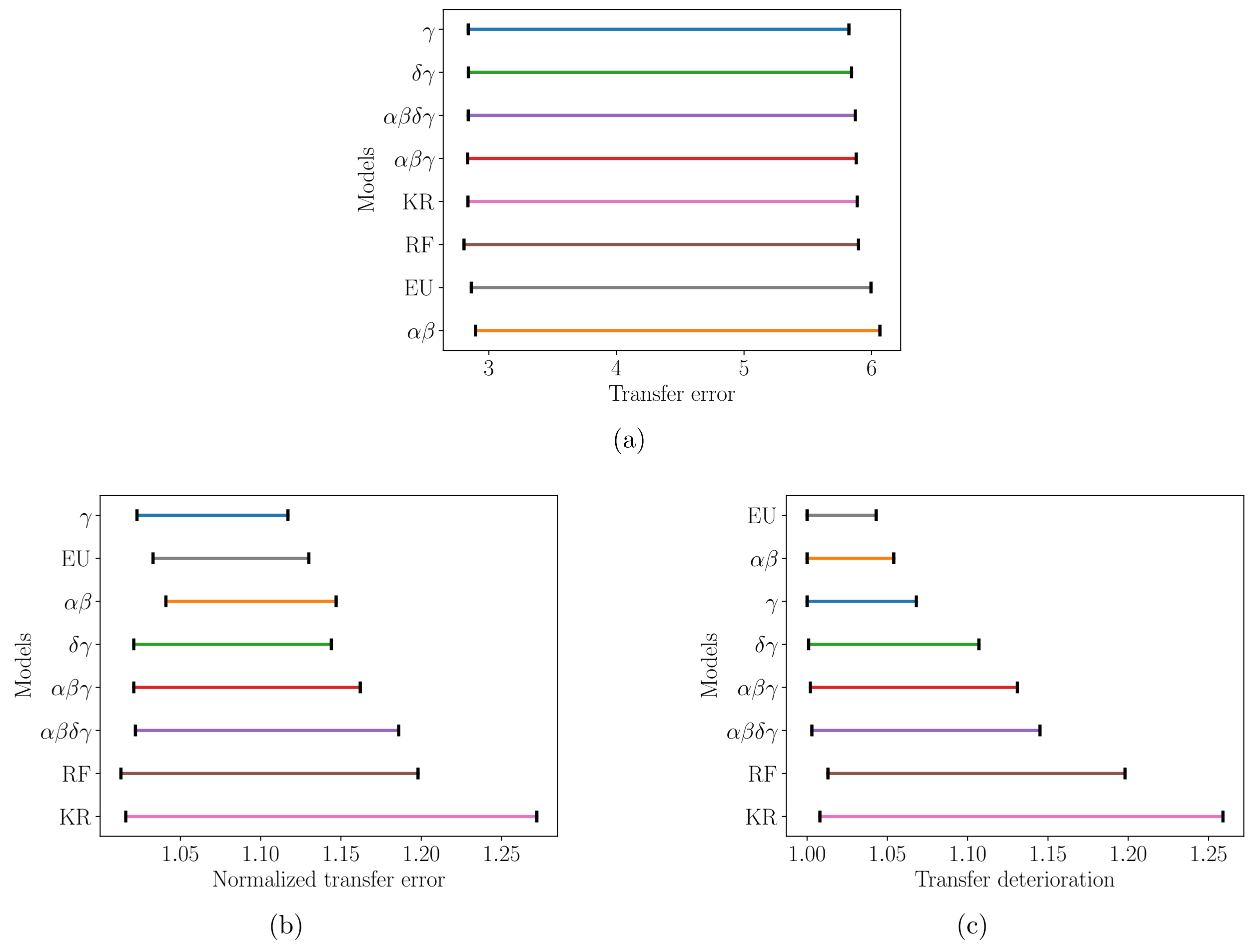


Figure 3: *71% forecast intervals for the 30 samples in l'Haridon and Vieider (2019).*

These observations tells us that poor transfer performance is not guaranteed for highly flexible prediction methods. Rather, there is something specific about black box algorithms that causes them to do poorly in certain kinds of transfer prediction tasks.

4.6 When do black boxes transfer poorly?

Our framework allows the distribution P governing the training sample and the distribution P' governing the test sample to differ from one another. At one extreme, P and P' may share a common marginal distribution on the feature space $\mathcal{X}$, but have very different conditional distributions $P_{Y|X}$ and $P'_{Y|X}$ (sometimes known as *model shift*). In our application, this would mean that the distribution over lotteries is the same, but the conditional distribution of reported certainty equivalents is different across domains. At another extreme, it may be that the conditional distributions $P_{Y|X}$ and $P'_{Y|X}$ are the same, but the marginal distributions over the feature space differ across domains, e.g., if different kinds of lotteries are used in different domains (sometimes known as *covariate shift*).

Our findings in Figure 3 suggest that black boxes do just as well as economic models at transfer prediction when the primary source of variation across distributions is a shift in the conditional distribution, rather than a shift in the marginal distribution over features. Intuitively, the observed training data necessarily involves only a small part of the feature space (e.g., a specific set of lotteries). The black box algorithms and economic models both search through a class of prediction rules to find the one that best fits the observed data. In our application, this means finding a prediction rule that does a good job of predicting certainty equivalents for the lotteries in the training data. If the conditional distributions $P_{Y|X}$ governing the test and training samples are different, the best predictions in the training sample will not be the best predictions in the test sample. But economic models and black box algorithms are disadvantaged for transfer prediction in the same way.

In contrast, when the set of lotteries varies across samples, then transfer prediction necessarily involves extrapolation. If the economic model has identified structure that is shared across settings, then fixing its parameters at values selected to perform best on the training data will improve predictions for the test lotteries. This need not be true for a black box algorithm that hasn't identified a global structure relating behavior across lotteries: Reproducing the average certainty equivalents for the training lotteries will not necessarily improve predictions on the test lotteries.

For a simple, stylized, example of this contrast, consider three domains with degenerate distributions over observations. In domain 1, the distribution is degenerate at the lottery $(\overline{z}, \underline{z}, p) = (10, 0, 1/2)$ and certainty equivalent $y = 3$. In domain 2, the distribution is degenerate at the lottery $(\overline{z}, \underline{z}, p) = (10, 0, 1/2)$ and certainty equivalent $y = 4$. In domain

3, the distribution is degenerate at a new lottery $(\overline{z}, \underline{z}, p) = (20, 10, 1/10)$ and certainty equivalent $y = 11$. Suppose EU and a decision tree are both trained on a sample from domain 1. The CRRA parameter $\eta \approx 0.64$ perfectly fits the observation $(10, 0, 1/2; 3)$, as does the trivial decision tree that predicts $y = 3$ for all lotteries. The estimated EU model and decision tree are equivalent for predicting observations in domain 2: both predict $y = 3$ and achieve a mean-squared error of 1. But their errors are very different on domain 3: the EU prediction for the new lottery is approximately 10.8 with a mean-squared error of approximately 0.05, while the decision tree's prediction is 3 with a mean-squared error of 64.

4.7 Predicting the relative transfer performance of black boxes and economic models

The preceding sections suggest that the relative transfer performance of black boxes and economic models is determined primarily by shifts in which lotteries are sampled, rather than shifts in behavior conditional on those lotteries. To further test this conjecture, we examine how well we can predict the ratio of the random forest transfer error to the CPT transfer error, $e(\rho^{RF}(\mathbf{M}_T), S_d)/e(\rho^{CPT}(\mathbf{M}_T), S_d)$, using only properties of the training lotteries in $\mathbf{M}_T$ and the test lotteries in S_d.

For each sample $S = \{(\overline{z}_i, \underline{z}_i, p_i; y_i)\}_{i=1}^m$, we consider the following features:

- the mean, standard deviation, max, and min value of $\overline{z}$ among the lotteries in S
- the mean, standard deviation, max, and min value of $\underline{z}$ among the lotteries in S
- the mean, standard deviation, max, and min value of p among the lotteries in S
- the mean, standard deviation, max, and min value of $1 - p$ among the lotteries in S
- the mean, standard deviation, max, and min of $p\overline{z} + (1 - p)\underline{z}$ among the lotteries in S
- the size of S
- an indicator variable for whether $\underline{z}, \overline{z} \geq 0$ for all lotteries in S

We consider three possible feature sets: (a) *Training Only,* which includes all features derived from the training sample S_T; (b) *Test Only,* which includes all features derived from the test sample S_d, (c) *Both,* which includes all features derived from the training sample S_T

and the test sample S_d. We consider two different prediction methods: OLS and a random forest algorithm. Table 2 reports tenfold cross-validated errors for each of these feature sets and prediction methods. As a benchmark, we also consider the best possible constant prediction.

	Train Only	Test Only	Both
Constant	2.57	2.57	2.57
OLS	1.00	2.61	0.94
RF	0.98	2.52	0.76

Table 2: *Cross-Validated MSE*

We find that there is substantial predictive power in features describing the training and test lotteries. The best constant prediction is 1.71 and achieves a mean-squared error of 2.57. Using features of the training set alone, it is possible to more than halve the error of the constant model. Using features of both the training and test sets, the random forest algorithm reduces error to 30% of the constant model. Crucially, the random forest algorithm is permitted to learn nonlinear combinations of the input features, and thus discover relationships between the training and test lotteries that are relevant to the relative performance of the black box and the economic model.

The random forest algorithm is too opaque to deliver insight into how these better predictions are achieved, but we can obtain some understanding by examining the best 1-split decision tree shown in Figure 4 below.

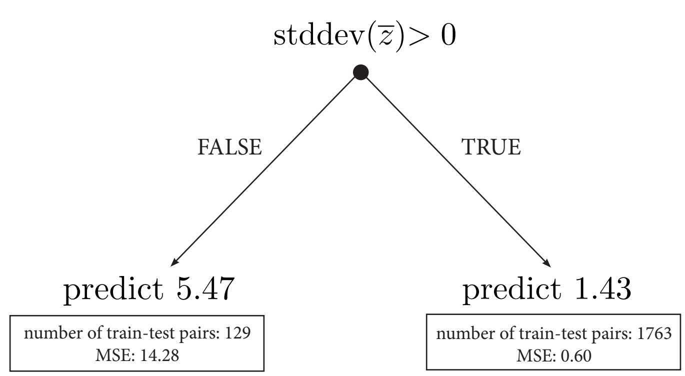


Figure 4: *Best 1-split decision tree based on training and test features.*

This decision tree achieves a cross-validated MSE of 1.75, reducing the error of the constant model by 32%. It partitions the set of (train,test) domain pairs into two groups depending on whether the standard deviation of $\overline{z}$ (the larger prize) in the training set of lotteries exceeds zero. There are three domains in which the prizes $(\overline{z}, \underline{z})$ are held constant across all training lotteries (although the probabilities vary). In the 129 transfer prediction tasks where one of these three domains is used for training, the decision tree predicts the ratio of the random forest error to the CPT error to be 5.47. For all other transfer prediction tasks, the decision tree predicts 1.43.

This finding reinforces our previous intuition that the relative performance of the black boxes and economic models is driven in part by whether the training sample covers the relevant part of the feature space. If the training observations concentrate on an unrepresentative part of the feature space (such as all lotteries that share a common pair of prize outcomes), then the black boxes transfer much more poorly than economic models.

These results also clarify the contrast between transfer performance and classical out-of-sample performance. In out-of-sample testing, the marginal distribution on $\mathcal{X}$ is the same for the training and test samples, so the set of training lotteries is likely to be representative of the set of test lotteries as long as the training sample is sufficiently large. When test and training samples are governed by distributions with different marginals on $\mathcal{X}$, the set of training lotteries can be unrepresentative of the set of test lotteries regardless of the number of training observations. Training on observations pooled across many domains alleviates the potential unrepresentativeness of the training data, but the number of domains needed will depend on properties of the distribution. For example, a difficult environment for the black box algorithms may be one in which each domain puts weight on exactly one lottery which is itself sampled iid,[26] while an easier environment may be one in which the marginal distribution is degenerate on the same lottery in all domains. There is no analog in out-of-sample testing for the role played by variation in the marginal distribution on $\mathcal{X}$ across domains. Moving beyond our specific application, we expect this variation to be an important determinant of the relative transfer performance of black box algorithms and economic models in general.

[26] In this edge case, the number of domains needed for black boxes to achieve good transfer performance is likely comparable to the number of observations needed for good out-of-sample performance, which can be quite large.

5 Extensions and further results

Our main results focus on forecasting realized transfer errors, which is useful when we want to know the range of plausible errors in transferring a given model a new domain. In many other contexts, however, inference is focused on population quantities, such as quantiles and means, rather than realized quantities such as transfer errors. This section develops methods that for inference on two quantities related to the population of domains. Section 5.1 provides confidence intervals for quantiles of the transfer error distribution. Section 5.2 provides a one-sided confidence interval for the expected transfer error.

5.1 Confidence intervals for quantiles of transfer error

Our focus so far has been on forecast intervals, i.e., intervals that cover the realized transfer error $e(\rho(\mathbf{M}_T), S_t)$ with a given probability when both the training data $\mathbf{M}_T$ and the target data S_t are drawn from μ. Although forecast intervals are useful for assessing the plausible range of transfer errors for a given decision rule, they have some potentially undesirable properties. First, as discussed above, the distributions of $e_{T,n+1}$ for different decision rules (and hence possibly their forecast intervals) will often overlap even if one method has smaller transfer errors in the sense of first order stochastic dominance. Further, since forecast intervals must account for the variation in $e(\rho(\mathbf{M}_T), S_t)$ from the random selection of training and target samples, the width of these intervals does not go to zero even if we have data from many samples.

This section provides a complementary approach that lets us construct traditional confidence intervals for quantiles of the transfer error distribution. Since these quantiles can be perfectly recovered given data from an infinite number of domains, we expect the lengths of these intervals to vanish as the number of observed domains grows large. We compute confidence intervals for quantiles of the distribution of CPT transfer errors and random forest transfer errors, and compare these confidence intervals across the two decision rules.

Formally, for μ^{n_T+1} the product measure corresponding to $n_T + 1$ iid draws from μ, let

$$q_\beta \equiv Q_{\beta,(\mathbf{M}_T,S_t)\sim\mu^{n_T+1}}(e_{T,t})$$

denote the β-th quantile of the transfer error when both the training and target data are drawn from μ. For ease of exposition we assume that $e_{T,t}$ is continuously distributed when $(\mathbf{M}_T, S_t) \sim \mu^{n_T+1}$, so the upper and lower quantiles of $e_{T,t}$ coincide. As with our forecast intervals above, our results in this section apply both to the raw transfer error (corresponding to $e_{T,t} = e(\rho(\mathbf{M}_T), S_t))$ and to transformations of the transfer error, e.g. the transfer deterioration (corresponding to $e_{T,t} = \frac{e(\rho(\mathbf{M}_T),S_t)}{e(\rho(S_t),S_t)}$).

To construct confidence intervals for q_β, define $J = \lfloor \frac{n}{n_T+1} \rfloor$ and let $\mathcal{T} = \{(T_j, t_j)\}_{j=1}^J$ collect disjoint subsets of $[n]$, with $T_j \cap t_j = \emptyset$ and $\{T_j \cup t_j\} \cap \{T_k \cup t_k\} = \emptyset$ for all $j \neq k$. Let $c_{\mathcal{T}}(q) = \sum_{j=1}^J 1\{e_{T_j,t_j} \leq q\}$ be the count of (T, t) pairs where the transfer error from T to t is less than q. If $\mathbf{M} \sim \mu^n$ then $c_{\mathcal{T}}(q_\beta)$ follows a $Bin(J, \beta)$ distribution, that is, a binomial distribution with J draws and success probability β. Let $F_{Bin}(\cdot; J, \beta)$ denote the $Bin(J, \beta)$ distribution function, and let

$$F_{Bin}^-(x; J, \beta) = \mathbb{P}_{X \sim Bin(J,\beta)}\left(X < x\right) = 1 - F_{Bin}(J - x; J, 1 - \beta)$$

denote the analog of the distribution function defined using a strict, rather than weak, inequality. Finally, let $\varphi(n_T, [n])$ denote the uniform distribution over collections $\mathcal{T}$ satisfying the conditions stated above. The following result provides one- and two-sided confidence intervals for the β-th quantile, where the choice of τ controls the confidence level.[27]

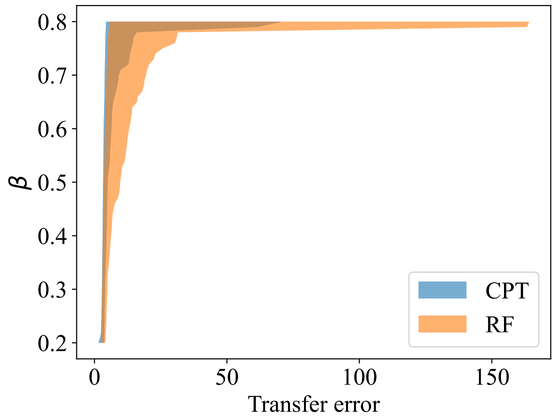
Proposition 2. *For $\overline{q}_{\beta,\tau} = \sup\left\{q : \mathbb{E}_{\mathcal{T} \sim \varphi(n_T,[n])}[F^-(c_{\mathcal{T}}(q); J, \beta)] < \tau\right\}$ the largest value of q such that the average of $F^-(c_{\mathcal{T}}(q); J, \beta)$ across $\mathcal{T} \sim \varphi(n_T, [n])$ is strictly below τ,*

$$\mathbb{P}\left(q_\beta > \overline{q}_{\beta,\tau}\right) \leq 2(1 - \tau).$$

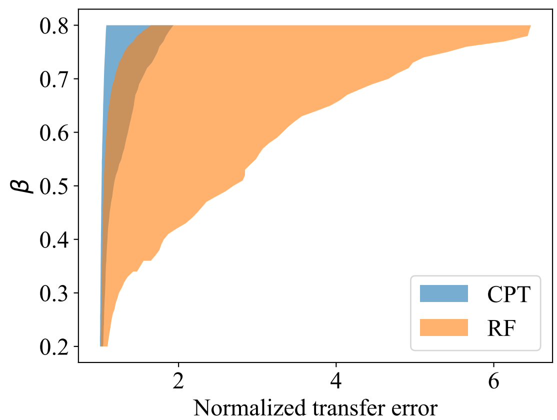
Further, for $\overline{q}^*_{\beta,\tau}$ the analog of $\overline{q}_{\beta,\tau}$ computed using $-e_{T,t}$, and $\underline{q}_{\beta,\tau} = -\overline{q}^*_{1-\beta,\tau}$,

$$\mathbb{P}\left(q_\beta \notin [\underline{q}_{\beta,\tau}, \overline{q}_{\beta,\tau}]\right) \leq 4(1 - \tau).$$

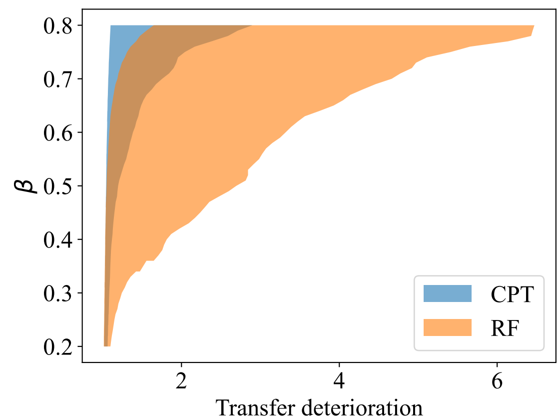
[27]For instance, to obtain a two-sided 90% confidence interval for the median transfer error, choose $\beta = 0.5$ and $\tau = 0.975$, so $4(1 - \tau) = 0.9$.



(a) *Confidence intervals for transfer error*



(b) *Confidence intervals for normalized transfer error*



(c) *Confidence intervals for transfer deterioration*

Figure 5: *80%-confidence intervals for the β-quantile of (a) transfer error, (b) normalized transfer error, and (c) transfer deterioration.*

It is not obvious how to obtain a closed-form expression for quantities like $\overline{q}_{\beta,\tau}$. However, as the next result shows, we can approximate these confidence sets by simulation, e.g. by repeatedly randomly pairing domains and averaging to approximate $\mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[F^{-}(c_{\mathcal{T}}(q); J, \beta)]$.

Corollary 1. *For $A \in \mathbb{N}$, let $(\mathcal{T}_1, ..., \mathcal{T}_A)$ be A independent draws from $\varphi(n_T, [n])$. Define $\overline{q}^B_{\beta,\tau} = \sup\left\{q : \frac{1}{A}\sum_{a\in A}[F^{-}(c_{\mathcal{T}_a}(q); J, \beta)] < \tau\right\}$ as the analog of $\overline{q}_{\beta,\tau}$ which replaces $\mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[\cdot]$ by a sample average over $\mathcal{T}_a$, and define $\underline{q}^A_{\beta,\tau}$ similarly. Then*

$$\mathbb{P}\left(q_\beta > \overline{q}^A_{\beta,\tau}\right) \leq 2(1-\tau).$$

$$\mathbb{P}\left(q_\beta \notin [\underline{q}^A_{\beta,\tau}, \overline{q}^A_{\beta,\tau}]\right) \leq 4(1-\tau).$$

The function $\frac{1}{A}\sum_{a\in A}[F^{-}(c_{\mathcal{T}_a}(q);J,\beta)]$ is weakly increasing and left-continuous in q, so calculating $\overline{q}^{B}_{\beta,\tau}$ reduces to finding the root of a monotonic function, and is tractable in applications. Figure 5 uses Corollary 1 to compute a two-sided 80% confidence interval for the transfer performance of CPT and random forest in Section 4's application. Although the confidence intervals overlap at every quantile level—reflecting the limited number of domains in our data—we again find that the lower and upper bounds of the random forest confidence intervals are higher at every quantile level.

5.2 Bounds on expected transfer error

Section 5.1 provides confidence intervals for quantiles of the transfer error distribution. This section provides complementary results for bounding the expected transfer error. Consider the special case of our framework where all observations (X, Y) are governed by the single distribution P. The well-known *Rademacher complexity* is a measure for the size or expressiveness of a set of prediction rules Σ^*. Formally, for any sample S, the Rademacher complexity of Σ^* with respect to S is

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

where $w \sim \{\pm 1\}^m$ means that w is drawn uniformly at random from the set of length-m vectors with entries equal to ± 1. We adapt this idea for our multiple-domain setting as follows.

Definition 5. *The* transfer Rademacher complexity *of* Σ^* *given meta-data* $\mathbf{M} = (S_1, \ldots, S_n)$ *is*

$$\mathcal{R}(\Sigma^*, \mathbf{M}) = \frac{1}{n}\mathbb{E}_{w\sim\{\pm 1\}^n}\left[\sup_{\sigma\in\Sigma^*}\sum_{i=1}^{n} w_i \cdot \left(\frac{1}{|S_i|}\sum_{(x_j,y_j)\in S_i}\ell(\sigma(x_j), y_j)\right)\right].$$

Transfer Rademacher complexity can be used to upper bound the expected transfer error of a prediction rule, where the more expressive the set of prediction rules Σ^* is (in the sense of higher transfer Rademacher complexity), the greater the possible difference between the error of $\sigma^*_{\mathbf{M}}$ on the meta-data $\mathbf{M}$ and its error on a sample from a new domain.

Proposition 3. *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 $\sigma \in \Sigma^*$,*

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

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

We omit the proof since this result follows directly from a standard result for the one-domain setting, see for example Theorem 26.5 in Shalev-Shwartz and Ben-David (2019). For our application in Section 4, the upper bound on the loss, c, is so large that the bound in Proposition 3 is uninformative, but the bound may be useful in other applications.

6 Conclusion

Our measures of transfer error quantify how well a model's performance on one domain extrapolates to other settings. We applied these measures to show that the predictions of expected utility theory and cumulative prospect theory outperform those of black box models on out-of-domain tests, even though the black boxes generally have lower out-of-sample prediction errors within a given domain. The relatively worse transfer performance of the black boxes seems to be because the black box algorithms have not identified structure that is commonly shared across domains, and thus cannot effectively extrapolate behavior from one set of features to another. Our finding that the economic models transfer better provides support for the intuition that economic models recover regularities that are general across a variety of domains.

A Other decision rules

We describe here additional examples of decision rules.

Example 8 (Domain Cross-Validation). Let R^* be a family of decision rules, for example random forest algorithms with decision trees restricted to different depths. Let

$$\rho^*_{\mathbf{M}_T} \in \operatorname*{argmin}_{\widetilde{\rho}\in R^*} \sum_{t\in T} e(\widetilde{\rho}(\mathbf{M}_{T\setminus\{t\}}), S_t) \quad \forall \mathbf{M}_T \in \mathcal{M}, |T| \geq 2$$

and set $\rho(\mathbf{M}_T) = \rho^*(\mathbf{M}_T)$. That is, we pick the decision rule which, among the set of candidates R^*, attains the best average leave-one-domain-out error across our training domains, and then apply that rule to the full training data.

Example 9 (Distributionally Robust Optimization (DRO)). DRO chooses model parameters to minimize worst-case error with respect to a neighborhood of distributions around the training data. First specify a distance function $D(P, P')$ for the distance between any two probability measures $P, P' \in \Delta(\mathcal{X} \times \mathcal{Y})$ and choose a neighborhood size ε. For any meta-data realization $\mathbf{M}_T$, let $P_{\mathbf{M}_T}$ denote the empirical distribution of the pooled set of observations across the samples in the meta-data. As in Example 1, consider a set of prediction rules Σ^*. We define ρ_{Σ^*} to satisfy

$$\rho_{\Sigma^*}(\mathbf{M}_T) \in \operatorname*{argmin}_{\sigma\in\Sigma^*} \left(\max_{P\in B_\varepsilon\left(P_{\mathbf{M}_T}\right)} \mathbb{E}_P[\ell(\sigma(x), y)] \right).$$

That is, the chosen prediction rule is the one that minimizes the worst-case error on distributions in the neighborhood of the observed data.

Example 10 (Hierarchical Bayesian Model). A Hierarchical Bayesian model posits a parametric form for the data generating process in each domain d, with parameters θ_d that can vary across domains. For instance, we might model the outcome y as normally distributed conditional on the features x, with conditional variance ς_d and conditional mean $\sigma_{\theta_d}(x)$, and specify a family of priors $\pi(\theta_d, \varsigma_d; \gamma)$ for the domain-level parameters (θ_d, ς_d), as well as a prior $\pi(\gamma)$ on the hyperparameter γ which is common across domains, and indexes the distribution of (θ_d, ς_d). In this case the log-likelihood for the training data in domain d, fixing ς_d and varying θ_d, is proportional to squared error loss. The hierarchical Bayes decision rule

$\rho^{HB}(\mathbf{M}_T)$ for squared error loss then computes the posterior on $(\gamma, \{\theta_d : d \in T\})$ based on the training data $\mathbf{M}_T$ and sets the prediction rule based on features x in domain d' equal to the posterior mean of $\sigma_{\theta_{d'}}(x)$.

B Proofs

B.1 Proof of Proposition 1

To save on notation let $\phi \equiv \phi(n_T, [n])$ be the distribution that draws n_T elements from the set $[n]$. The set of training domains is distributed $T \sim \phi$. For a given realization of T, let $X_T(a) = \sum_{d \in [n] \setminus T} 1\{e_{T,d} < a\}$ count the number of test samples where the transfer error from T to these domains is strictly below a. The assumption that $(\mathcal{P}_d, m_d)$ is drawn iid from μ in each domain implies that if we evaluate $X_T(\cdot)$ at the transfer error of the target domain, $e_{T,n+1}$, the distribution of $X_T(e_{T,n+1})$ (for T fixed) is dominated by a discrete uniform distribution over $\{0, ..., n - n_T\}$ in the sense of first-order stochastic dominance. Hence, for $X_T^*(a) = \frac{1}{n - n_T + 1} X_T(a)$, the distribution of $X_T^*(e_{T,n+1})$ is dominated by a $U[0,1]$ 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 $X_T^*(e_{T,n+1}) \leq U_T$.[28] For $T \sim \phi$, the distribution of U_T is a mean-preserving spread of the distribution of $\mathbb{E}_{T \sim \phi}[U_T]$.

We now use Lemma 1 from Meng (1994), which applies to any random variable W with mean $1/2$ that is second-order stochastically dominated by a $U[0,1]$-random variable. Setting $W = 1 - \mathbb{E}_{T \sim \phi}[U_T]$ and $\alpha = 1 - \psi$ in the lemma's conclusion shows that

$$\mathbb{P}\left(\mathbb{E}_{T \sim \phi}[U_T] \geq \psi\right) \leq 2(1 - \psi).$$[29]

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

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

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

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

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

below ψ,

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

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

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

Hence,

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

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.

B.2 Proof of Proposition 2

Since $F^-_{Bin}(x; J, \beta)$ is weakly increasing and left-continuous in x, it follows that $F^-_{Bin}(x; J, \beta) < \tau$ if and only if $x \leq x_\tau$, where

$$x_\tau \equiv \sup\{x : F^-_{Bin}(x; J, \beta) < \tau\}.$$

Hence,

$$\mathbb{P}_{X\sim Bin(J,\beta)}\left(F^-_{Bin}(X; J, \beta) < \tau\right) = F_{Bin}(x_\tau; J, \beta) \geq \tau.$$

Hence, the distribution of $F^-(c_{\mathcal{T}}(q_\beta); J, \beta)$ is dominated by a $U[0,1]$ in the sense of first-order stochastic dominance, and for each $\mathcal{T}$ we can construct a $U[0,1]$ random variable $U_{\mathcal{T}}$ such that $F^-(c_{\mathcal{T}}(q_\beta); J, \beta) \leq U_{\mathcal{T}}$ almost surely.

As in the proof of Proposition 1, Lemma 1 in Meng (1994) implies that

$$\mathbb{P}\left(\mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[U_{\mathcal{T}}] \geq \tau\right) \leq 2(1-\tau),$$

so since $\mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[F^-(c_\mathcal{T}(q_\beta);J,\beta)] \le \mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[U_\mathcal{T}]$ almost surely,

$$\mathbb{P}\left(\mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[F^-(c_\mathcal{T}(q_\beta);J,\beta)] \ge \tau\right) \le 2(1-\tau).$$

Thus, if we collect the set of values q where $\mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[F^-(c_\mathcal{T}(q);J,\beta)]$ falls below τ,

$$CS = \{q : \mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[F^-(c_\mathcal{T}(q);J,\beta)] < \tau\},$$

we have that

$$\mathbb{P}\left(q_\beta \in CS\right) \ge 1 - 2(1-\tau) = 2\tau - 1.$$

Note further, that since $F^-(\cdot;J,\beta)$ and $c_\mathcal{T}$ are left-continuous and weakly increasing, and the support of $\mathcal{T}$ is finite, $\mathbb{E}_{\mathcal{T}\sim\varphi(n_T,[n])}[F^-(c_\mathcal{T}(q);J,\beta)]$ is also left-continuous and weakly increasing. Hence, we can write $CS = (-\infty, \bar{q}_\beta]$ for $\bar{q}_\beta$ as defined in the statement of the proposition.

B.3 Proof of Corollary 1

For each $a \in [A]$, $U_{\mathcal{T}_a}$ is a mean-preserving spread of $\frac{1}{A}\sum_{a'} U_{\mathcal{T}_{a'}}$. Hence, the same argument as in the proof of Proposition 2 establishes that

$$\mathbb{P}\left(\frac{1}{A}\sum_a [F^-(c_{\mathcal{T}_a}(q_\beta);J,\beta)] \ge \tau\right) \le 2(1-\tau),$$

from which the result follows, again by the same argument as in Proposition 2.

References

Abdellaoui, M., P. Klibanoff, and L. Placido (2015): "Experiments on compound risk in relation to simple risk and to ambiguity," *Management Science*, 61, 1306–1322.

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

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

Benartzi, S., J. Beshears, K. L. Milkman, C. R. Sunstein, R. H. Thaler,

M. Shankar, W. Tucker-Ray, W. J. Congdon, and S. Galing (2017): "Should Governments Invest More in Nudging?" *Psychological Science*, 28, 1041–1055, pMID: 28581899.

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

Blanchard, G., G. Lee, and C. Scott (2011): "Generalizing from Several Related Classification Tasks to a New Unlabeled Sample," in *Advances in Neural Information Processing Systems*, ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Curran Associates, Inc., vol. 24.

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

Bruhin, A., H. Fehr-Duda, and T. Epper (2010): "Risk and rationality: Uncovering heterogeneity in probability distortion," *Econometrica*, 78, 1375–1412.

Camerer, C. F., G. Nave, and A. Smith (2019): "Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning," *Management Science*, 65, 1867–1890.

Dean, M. and P. Ortoleva (2019): "The empirical relationship between nonstandard economic behaviors," *Proceedings of the National Academy of Sciences*, 116, 16262–16267.

DellaVigna, S. and D. Pope (2019): "Stability of Experimental Results: Forecasts and Evidence," .

Etchart-Vincent, N. and O. 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, 61–83.

Fan, Y., D. V. Budescu, and E. Diecidue (2019): "Decisions with compound lotteries." *Decision*, 6, 109.

Fehr-Duda, H., A. Bruhin, T. Epper, and R. Schubert (2010): "Rationality on the rise: Why relative risk aversion increases with stake size," *Journal of Risk and Uncertainty*, 40, 147–180.

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

Fudenberg, D., W. Gao, and A. Liang (2021): "How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories," Working Paper.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2022): "Measuring the Completeness of Economic Models," Forthcoming in the Journal of Political Economy.

Fudenberg, D. and A. Liang (2019): "Predicting and Understanding Initial Play," *American Economic Review*, 109, 4112–4141.

Goldstein, W. M. and H. J. Einhorn (1987): "Expression theory and the preference reversal phenomena," *Psychological review*, 94, 236–254.

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

Hummel, D. and A. 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, 47–58.

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

Ke, S., C. Zhao, Z. Wang, and S.-L. Hsieh (2020): "Behavioral Neural Networks," Working Paper.

Külpmann, P. and C. Kuzmics (2022): "Comparing Theories of One-Shot Play Out of Treatment," .

Lattimore, P. K., J. R. Baker, and A. D. Witte (1992): "The influence of probability on risky choice: A parametric examination," *Journal of Economic Behavior & Organization*, 17, 315–436.

Lefebvre, M., F. M. Vieider, and M. C. Villeval (2010): "Incentive effects on risk attitude in small probability prospects," *Economics Letters*, 109, 115–120.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman (2018): "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094–1111.

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

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

Meng, X.-L. (1994): "Posterior Predictive p-Values," *Annals of Statistics*, 22, 1142–1160.

Muandet, K., D. Balduzzi, and B. Schölkopf (2013): "Domain Generalization via Invariant Feature Representation," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, JMLR.org, ICML'13, I–10–I–18.

Murad, Z., M. Sefton, and C. Starmer (2016): "How do risk attitudes affect measured confidence?" *Journal of Risk and Uncertainty*, 52, 21–46.

Murphy, K. (2012): *Machine Learning: a Probabilistic Perspective*, MIT Press.

Noti, G., E. Levi, Y. Kolumbus, and A. Daniely (2016): "Behavior-Based Machine-Learning: A Hybrid Approach for Predicting Human Decision Making," *CoRR*, abs/1611.10228.

OECD (2022): "Purchasing Power Parities," .

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011): "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

Plonsky, O., R. Apel, E. Ert, M. Tennenholtz, D. Bourgin, J. Peterson, D. Reichman, T. Griffiths, S. Russell, E. Carter, J. Cavanagh, and I. Erev (2019): "Predicting human decisions with behavioral theories and machine learning," *CoRR*, abs/1904.06866.

Plonsky, O., I. Erev, T. Hazan, and M. Tennenholtz (2017): "Psychological forest: Predicting human behavior," *AAAI Conference on Artificial Intelligence*, 31, 656–662.

Rahimian, H. and S. Mehrotra (2019): "Distributionally Robust Optimization: A Review," Working Paper.

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

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

Sutter, M., M. G. Kocher, D. Glätzle-Rützler, and S. T. Trautmann (2013): "Impatience and uncertainty: Experimental decisions predict adolescents' field behavior," *American Economic Review*, 103, 510–31.

Vovk, V. and R. Wang (2020): "Combining p-values via Averaging," *Biometrika*, 107, 791–808.

Zhou, K., Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy (2021): "Domain Generalization: A Survey," Working Paper.

Online appendix to the paper

The Transfer Performance of Economic Models

Isaiah Andrews Drew Fudenberg Annie Liang Chaofeng Wu

July 28, 2022

This online appendix provides additional empirical results to complement those reported in Section 4 of the main text.

C Supplementary material for Section 4

C.1 Description of data

We briefly describe the individual samples in our meta-data. There are 44 domains in total.

Table 3

Source of Data	# Obs	# Subj	# Lottery	Country	Gains Only
Abdellaoui et al. (2015)	801	89	3	France	Y
Fan et al. (2019)	4750	125	19	US	Y
Bouchouicha and Vieider (2017)	3162	94	66	UK	N
Sutter et al. (2013)	661	661	4	Austria	Y
Etchart-Vincent and l'Haridon (2011)	3036	46	20	France	N
Fehr-Duda et al. (2010)	8560	153	56	China	N
Lefebvre et al. (2010)	72	72	2	France	Y
Halevy (2007)	366	122	2	Canada	Y
Anderhub et al. (2001)	183	61	1	Israel	Y
Murad et al. (2016)	2131	86	25	UK	Y
Dean and Ortoleva (2019)	1032	179	3	US	Y
Bernheim and Sprenger (2020)	1071	153	7	US	Y
Bruhin et al. (2010)	8906	179	50	Switzerland	N
Bruhin et al. (2010)	4669	118	40	Switzerland	N
l'Haridon and Vieider (2019)	1708	61	27	Australia	N
l'Haridon and Vieider (2019)	2548	95	27	Belgium	N
l'Haridon and Vieider (2019)	2350	84	27	Brazil	N

l'Haridon and Vieider (2019)	2240	80	27	Cambodia	N
l'Haridon and Vieider (2019)	2687	96	27	Chile	N
l'Haridon and Vieider (2019)	5711	204	27	China	N
l'Haridon and Vieider (2019)	3072	128	23	Colombia	N
l'Haridon and Vieider (2019)	2968	106	27	Costa Rica	N
l'Haridon and Vieider (2019)	2770	99	27	Czech Republic	N
l'Haridon and Vieider (2019)	3906	140	27	Ethiopia	N
l'Haridon and Vieider (2019)	2604	93	27	France	N
l'Haridon and Vieider (2019)	3639	130	27	Germany	N
l'Haridon and Vieider (2019)	2352	84	27	Guatemala	N
l'Haridon and Vieider (2019)	2492	89	27	India	N
l'Haridon and Vieider (2019)	2352	84	27	Japan	N
l'Haridon and Vieider (2019)	2716	97	27	Kyrgyzstan	N
l'Haridon and Vieider (2019)	1791	64	27	Malaysia	N
l'Haridon and Vieider (2019)	3360	120	27	Nicaragua	N
l'Haridon and Vieider (2019)	5638	202	27	Nigeria	N
l'Haridon and Vieider (2019)	2660	95	27	Peru	N
l'Haridon and Vieider (2019)	2491	89	27	Poland	N
l'Haridon and Vieider (2019)	1959	70	27	Russia	N
l'Haridon and Vieider (2019)	1819	65	27	Saudi Arabia	N
l'Haridon and Vieider (2019)	1988	71	27	South Africa	N
l'Haridon and Vieider (2019)	2240	80	27	Spain	N
l'Haridon and Vieider (2019)	2212	79	27	Thailand	N
l'Haridon and Vieider (2019)	2070	74	27	Tunisia	N
l'Haridon and Vieider (2019)	2240	80	27	UK	N
l'Haridon and Vieider (2019)	2701	97	27	US	N
l'Haridon and Vieider (2019)	2436	87	27	Vietnam	N

C.2 Random assignment of observations to domains

We now consider 44 new domains that are generated by pooling all observations in the meta-data and randomly reassigning them to 44 domains, where the number of observations per domain is fixed to be the same as in our original samples. As discussed in Section 2.3, these domains are governed by the same distribution, so the transfer prediction task is not essentially different from classical out-of-sample prediction. We verify below that in this case, transfer performance closely resembles performance for out-of-sample (but within-domain) prediction.

Figure 6 is analogous to Figure 1 from the main text, and reports the CDFs of out-of-sample errors across the 44 domains. Figure 7 is analogous to Figure 2, and reports forecast intervals for our three measures of transferability. We find that in both cases, the black boxes slightly underperform the economic models. These results contrast with our main findings (with subject pool domains), where black boxes slightly outperform the economic models in within-domain tests (Figure 1) but perform more poorly in transfer prediction (Figure 2).

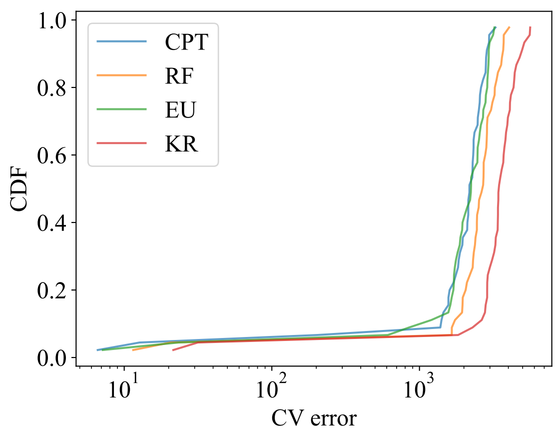


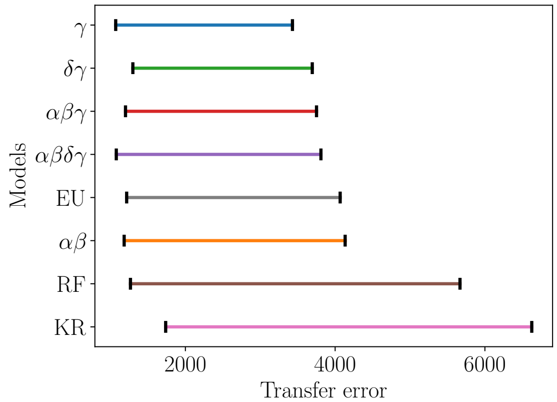
Figure 6: *CDFs of out-of-sample errors.*

C.3 Papers as domains

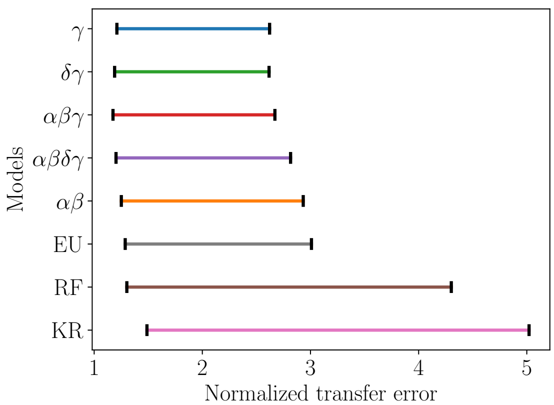
We now consider an alternative definition of domains, with each of the 14 papers representing a different domain. This changes the content of the iid assumption imposed in Section 3, where we now assume that samples are iid across papers, but may be dependent across subject pools within the same paper. We repeat our main analysis and report forecast intervals below, which are qualitatively similar to those reported in Figure 2.

C.4 In-sample errors

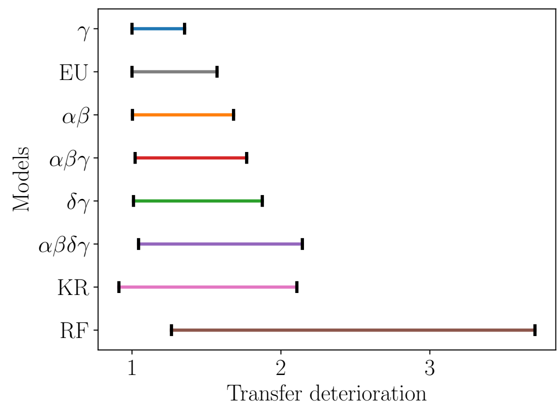
Figure 9 displays the CDFs of the in-sample errors of EU, CPT, the random forest algorithm, and kernel regression. As with Figure 1 these curves are nearly indistinguishable.



(a) *Transfer error*



(b) *Normalized transfer error*



(c) *Transfer deterioration*

Figure 7: *71% forecast intervals for (a) transfer error, (b) normalized transfer error, and (c) transfer deterioration, when observations are randomly assigned to domains.*

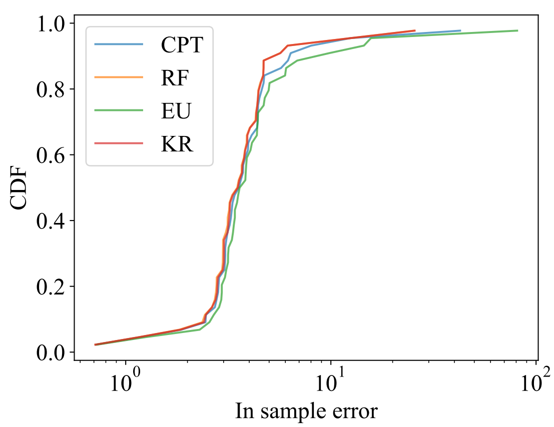
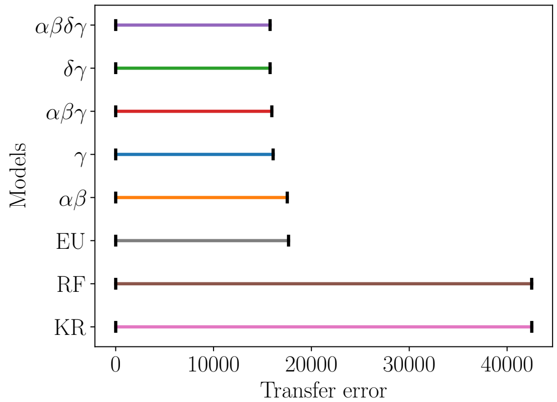
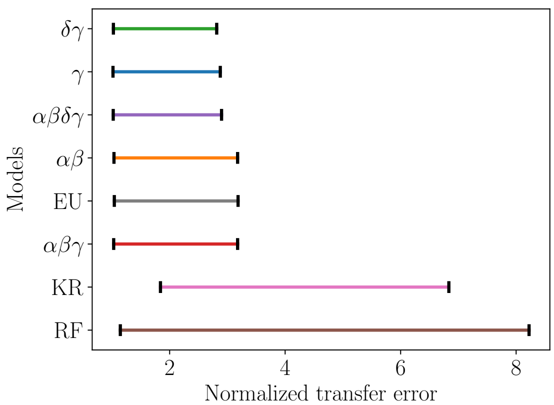


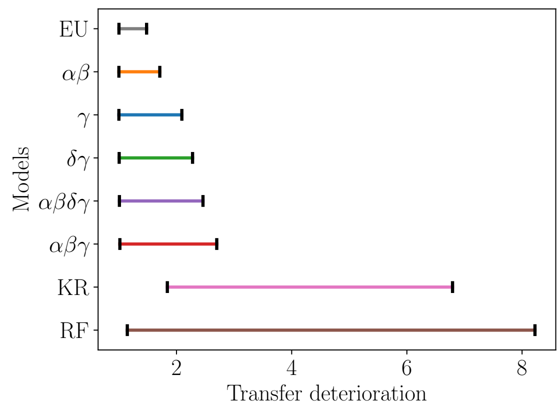
Figure 9: *CDF of in-sample errors*



(a) *Transfer error*



(b) *Normalized transfer error*



(c) *Transfer deterioration*

Figure 8: *53% forecast intervals for each of the three measures, treating each paper as a separate domain.*

C.5 Supplementary tables and figures for main analysis

Table 4 reports the forecast intervals that are depicted in Figure 2.

Model	Transfer Error	Normalized Error	Deterioration
CPT variants			
γ	[2.50,15.83]	[1.03,2.54]	[1.00,1.47]
α, β	[2.56,16.13]	[1.04,2.35]	[1.00,1.30]
δ, γ	[2.48,17.19]	[1.02,2.47]	[1.00,1.53]
α, β, γ	[2.47,15.91]	[1.02,2.60]	[1.00,1.85]
$\alpha, \beta, \delta, \gamma$	[2.46,15.99]	[1.02,2.62]	[1.00,1.82]
EU models			
EU	[2.56,16.41]	[1.04,2.14]	[1.00,1.30]
ML algorithms			
Random Forest	[2.71,31.39]	[1.02,6.42]	[1.02,6.42]
Kernel Regression	[2.75,33.62]	[1.02,5.33]	[1.01,5.29]

Table 4: *71% forecast intervals*

C.6 Alternative forecast intervals

In this section, we report alternative forecast intervals for our three measures. Table 5 constructs two-sided forecast 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% forecast intervals. Table 6 constructs one-sided forecast intervals whose upper bounds are the 95% transfer error; applying Proposition 1, these are 86% forecast intervals. All of the forecast intervals are qualitatively similar to the 71% two-sided forecast intervals reported in the main text.

Model	Transfer Error	Normalized Error	Deterioration
CPT main variants			
γ	[0.81,23104.96]	[1.01,7.31]	[1.00,7.22]
α, β	[0.71,19999.41]	[1.00,5.28]	[1.00,5.27]
δ, γ	[0.71,23052.76]	[1.00,7.25]	[1.00,7.18]
α, β, γ	[0.71,28122.26]	[1.00,5.65]	[1.00,5.60]
$\alpha, \beta, \delta, \gamma$	[0.71,27959.10]	[1.00,6.01]	[1.00,5.95]
EU models			
EU	[0.72,22787.99]	[1.00,4.44]	[1.00,1.75]
ML algorithms			
Random Forest	[0.96,42520.49]	[1.01,33.17]	[1.01,33.17]
Kernel Regression	[1.01,42519.23]	[1.01,6.835]	[1.00,6.79]

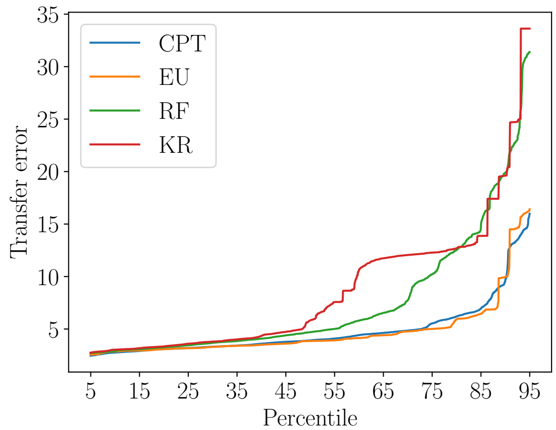
Table 5: *90% two-sided forecast intervals*

Model	Transfer Error	Normalized Error	Deterioration
CPT main variants			
γ	[0,15.83]	[1,2.54]	[1,1.47]
α, β	[0,16.13]	[1,2.35]	[1,1.30]
δ, γ	[0,17.19]	[1,2.47]	[1,1.53]
α, β, γ	[0,15.91]	[1,2.60]	[1,1.85]
$\alpha, \beta, \delta, \gamma$	[0,15.99]	[1,2.62]	[1,1.82]
EU models			
EU	[0,16.41]	[1,2.14]	[1,1.30]
ML algorithms			
Random Forest	[0,31.39]	[1,6.42]	[1,6.42]
Kernel Regression	[0,33.62]	[1,5.33]	[1,5.29]

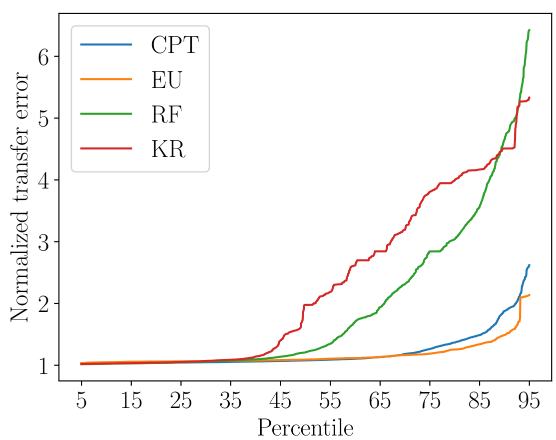
Table 6: *86% one-sided forecast intervals*

Finally, Figure 10 plots the τ-th percentile of the pooled transfer errors as τ varies. It is clear that the qualitative conclusions we have drawn about the relative performance of black boxes and economic models is not specific to any choice of τ.[30] In fact, in Panels (a) and (c), the black box curves lie everywhere above the CPT and EU curves, so both the lower and upper bounds of the black boxes' forecast intervals must be higher than those of the economic models, for any choice of τ.

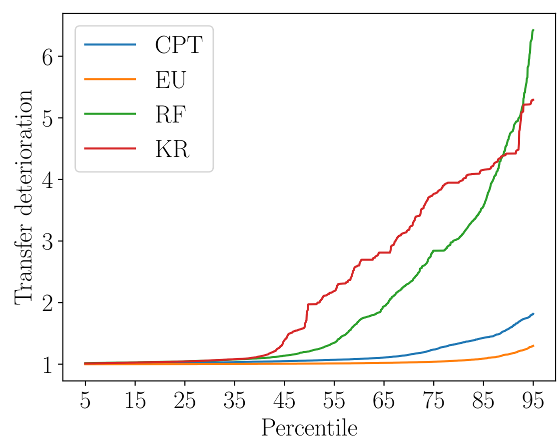
[30]To improve readability, we remove extreme numbers by truncating $\tau \in [5, 95]$, and show results only for the $\alpha\beta\gamma\delta$ specification of the CPT model.



(a) *Transfer error*



(b) *Normalized transfer error*



(c) *Transfer deterioration*

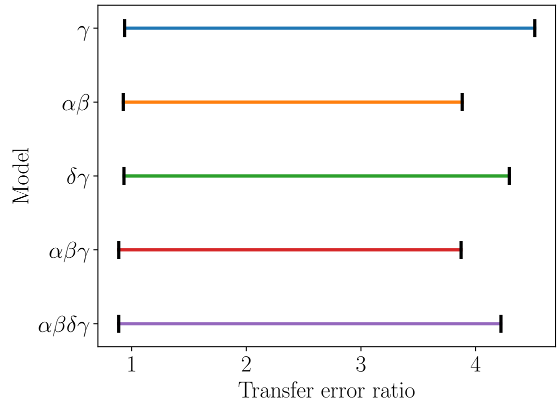
Figure 10: *Error percentiles from 5 to 95 (truncated for readability).*

C.7 Forecast intervals for the ratio of CPT and RF errors

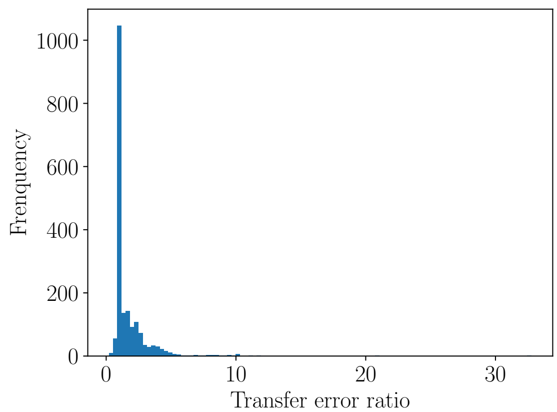
Let ρ^{CPT} denote the decision rule corresponding to CPT, and ρ^{RF} denote the decision rule corresponding to the random forest algorithm. Define

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

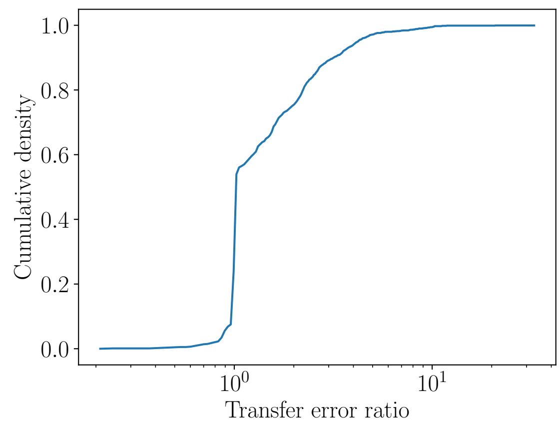
to be the ratio of the random forest transfer error to the CPT transfer error, henceforth the *transfer error ratio*.



(a) *71% Forecast intervals*



(b) *Density*



(c) *CDF*

Figure 11: *Forecast intervals, density, and cdf for the ratio of the random forest transfer error to the CPT transfer error.*

Panel (a) of Figure 11 reports 71% two-sided forecast intervals for the transfer error ratio for each CPT specification. The lower bound for each CPT model is approximately 0.9, while the upper bound is as large as 4.5. Panel (b) of the figure is a histogram of transfer error ratios for the 4-parameter CPT model when the training domains T and the target domains d are drawn uniformly at random from the set of domains in the meta-data. This distribution has a large cluster of ratios around 1 (i.e., CPT transfer errors are similar to the random forest errors) and a long right tail of ratios achieving a max value of 32.8 (i.e., the random forest error can be up to 32 times as large as the CPT error). The cumulative distribution function of $e_{T,d}$, reported in Panel (c) of Figure 11, shows that the random forest algorithm outperforms CPT in approximately 35% of (T, d) pairs, although CPT rarely has a much worse transfer error than the random forest and is sometimes much better.

C.8 Alternative Choice of n_T

Here we consider an alternative choice for the number of training domains, setting $n_T = 3$ instead of $n_T = 1$.[31] This corresponds to randomly choosing 3 of the 44 domains to be the training domains, finding the best prediction rule for this pooled data, and using the estimated prediction rule to predict the remaining 41 samples.

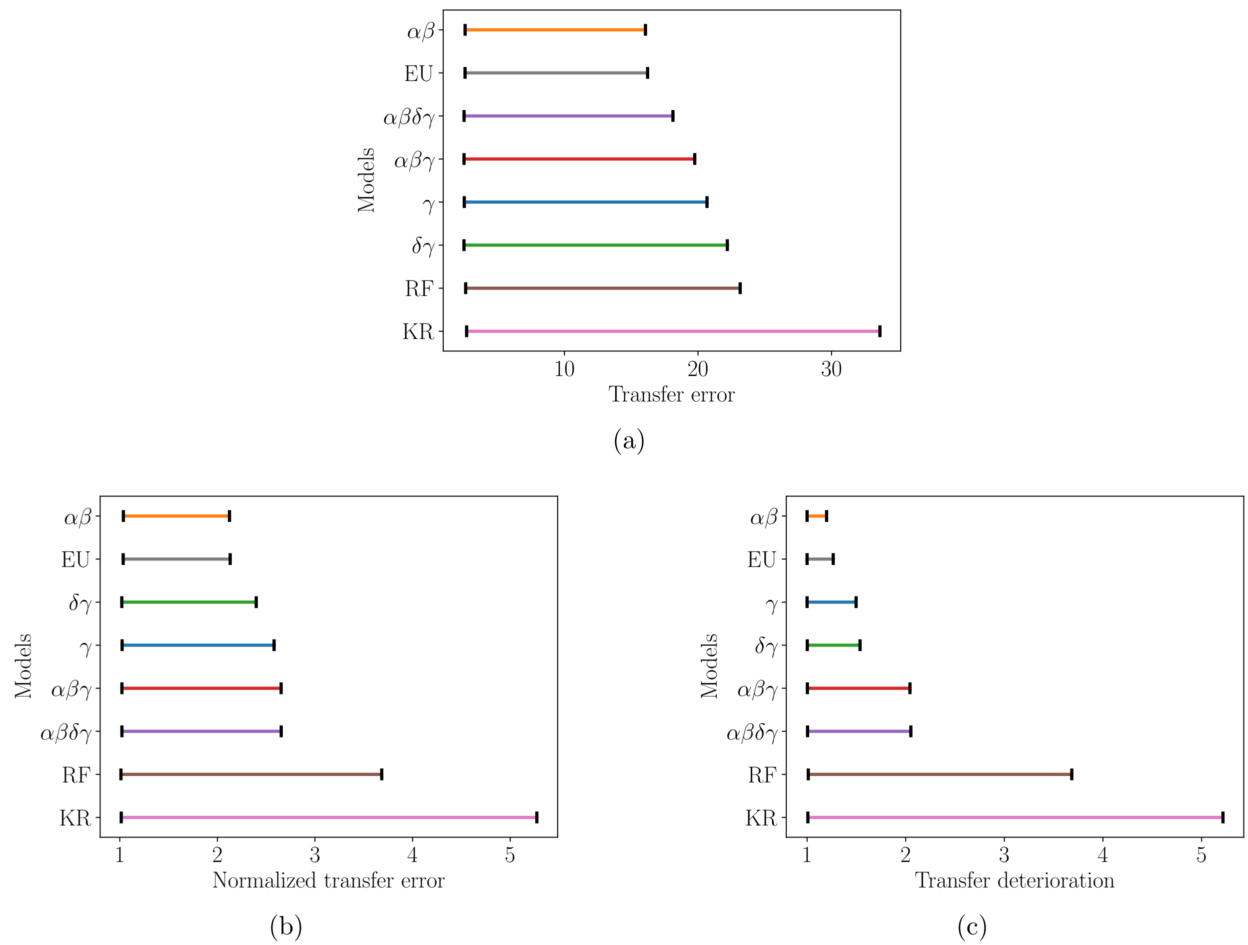
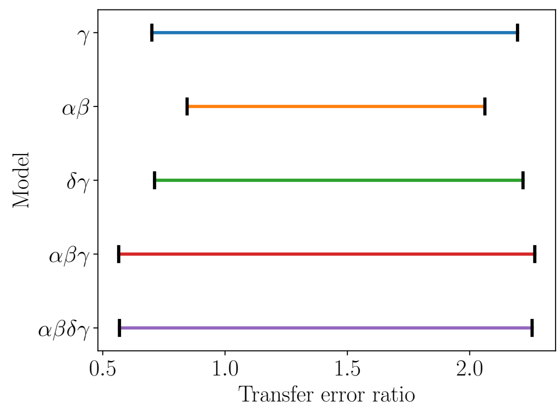


Figure 12: *71% forecast intervals for (a) transfer error, (b) normalized transfer error, and (c) transfer deterioration, with the choice of* $n_T = 3$.

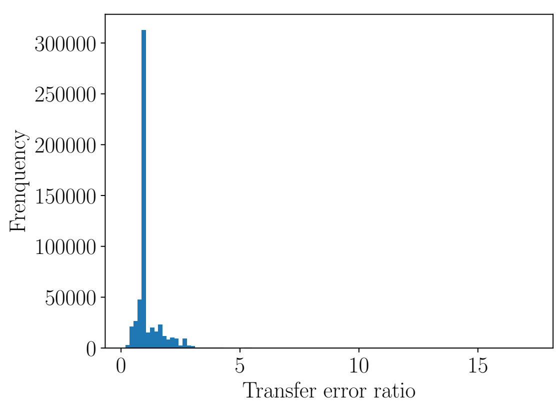
Figure 12 is the analog of Figure 2. Again we choose $\tau = 0.95$, thus constructing forecast intervals whose lower bounds are the 5% percentile of pooled transfer errors, and whose upper bounds are the 95% percentile of pooled transfer errors. Applying Proposition 1, these are 71% forecast intervals. The most notable change is that the random forest forecast interval

[31] We chose $n_T = 3$ to preserve a large number of test domains, and to keep the number of tests manageable. (The choice of $n_T = 3$ already requires computing $\binom{44}{3} \cdot 41 = 543{,}004$ transfer errors.)

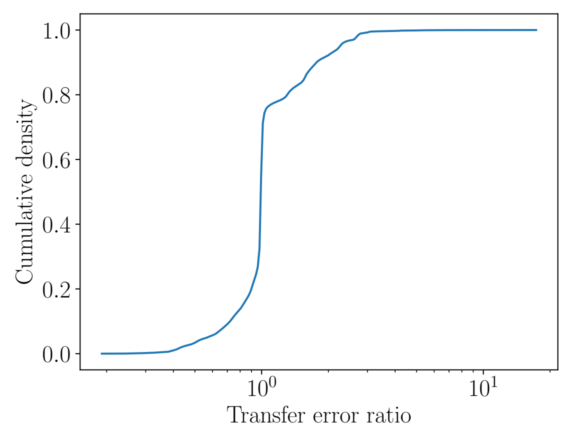
shrinks considerably, which suggests that the transfer error of the random forest algorithm becomes less variable when it is trained on more domains. Otherwise, all of the qualitative statements in the main text for $n_T = 1$ continue to hold. In particular, as with $n_T = 1$, we find that the forecast intervals for all three of our measures have higher lower and upper bounds for the black box algorithms than for the CPT specifications.



(a) *54% Forecast intervals*



(b) *Density*



(c) *CDF*

Figure 13: *Forecast intervals, density, and cdf of the ratio of the random forest transfer error to the CPT transfer error for $n_T = 3$.*

Figure 13 is the analog of Figure 11. The findings are again similar to the main text, but we see further evidence that the random forest algorithm improves when trained on samples from more domains. For example, the forecast intervals for the transfer error ratio (Panel (a)) now range from approximately 0.6 to approximately 2.2, instead of 0.9 to 4.5. Comparing Panel (b) of Figures 11 and 13, 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 findings described in Zhou et al. (2021).