THE TRANSFER PERFORMANCE OF ECONOMIC MODELS

ISAIAH ANDREWS[§], DREW FUDENBERG[‡], LIHUA LEI[#], ANNIE LIANG[†], AND CHAOFENG WU[*]

May 20, 2023

ABSTRACT. Economists often estimate models using data from a particular domain, e.g. estimating risk preferences in a particular subject pool or for a specific class of lotteries. Whether a model's predictions extrapolate well across domains 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 based on how well it performs 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, the black box algorithms we consider outperform standard economic models when estimated and tested on data from the same domain, but the economic models generalize across domains better than the black-box algorithms do.

[§]Department of Economics, Harvard.
[‡]Department of Economics, MIT.
[#]Stanford Graduate School of Business.
[†]Department of Economics, 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, we might estimate a pricing model on purchase data from one population of consumers and use it to predict demand in a new population with different demographics, or estimate a model of risk preferences on choices over insurance plans and use it to predict choice over state-contingent consumption bundles. This paper provides a tractable approach for evaluating cross-domain transfer performance, which (among other uses) can be applied to compare the generalizability of economic models and black box machine learning methods. Specifically, we derive finite-sample forecast intervals for a model's out-of-domain error, and then use them to evaluate the transferability of models and black box algorithms for predicting risk preferences over binary lotteries.[1] In this application, economic models transfer more reliably than two popular black box algorithms.

The question of how well a model trained on data from one domain will perform in a new domain dates back at least to Haavelmo (1944), and is a focus of work on external validity (Pearl and Bareinboim, 2011; Tipton and Olsen, 2018; Chassang and Kapon, 2022) and out-of-distribution prediction (Shen et al., 2021).[2] Understanding the transfer performance of various models is of increased relevance given the recent popularity of black-box machine learning methods within economics (Hofman et al., 2021). Black box methods (such as random forest algorithms) have been widely criticized for their lack of generalizability, and one reason some economists prefer structured economic models is the belief that such models are more likely to capture fundamental regularities that apply in a wide variety of domains (Coveney et al., 2016; Athey, 2017; Manski, 2021). On the other hand, economic models have also been criticized for failing to generalize. For example, although the theoretical literature on risk preferences includes some of the most widely used models in economics, it has been singled out as a setting where large datasets and flexible black box methods may

---

[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, but we could view these intervals as confidence intervals for these random targets.

[2] This literature considers other issues as well, such as whether parameter estimates or causal findings generalize to new domains.

help identify new regularities and more complete models (e.g., Plonsky et al. 2019, Ke et al. 2020, and Peterson et al. 2021).[3]

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 that yields easy-to-apply procedures which we then generalize.

We consider several measures for a model's transferability across domains. 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 then define a model's *normalized transfer error* to be 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 help select between models for making predictions in new domains; the third measure indicates the value of obtaining data from the target domain.

Section 3 shows how to construct forecast intervals with guaranteed coverage probability for these measures, using a meta-data set of samples from already observed domains. To construct a forecast interval for the estimated model's transfer error 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 yields an empirical distribution of transfer errors. We show that for every quantile τ, the interval bounded by the τ-th and $(1 - \tau)$-th quantiles of

[3]Peterson et al. (2021) writes: "We believe that use of large datasets coupled with machine-learning algorithms offers enormous potential for uncovering new cognitive and behavioral phenomena that would be difficult to identify without such tools." A second example is predicting play in games, where Hartford et al. (2016) says "[t]he recent success of deep learning has demonstrated that predictive accuracy can often be enhanced... by fitting highly flexible models that are capable of learning novel representations."

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. We next relax our iid sampling assumption, deriving a modified procedure for cases where the distributions in training domains are drawn iid from one distribution, while the distribution in the target domain is drawn from another. We implement this procedure and the other methods described in this paper in an `R` package (`transferUQ`), available on Github.[4]

Section 4 uses our 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 describes 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, and find 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. Transfer performance points to the latter: 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 Cumulative Prospect Theory normalized transfer errors (which constitute an 81% 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 worse within-domain for most of the subject pools, their predictions 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, so black box methods do not always transfer worse. Instead, black boxes seem to

---

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).

1.1. **Related Literature.** This paper is situated at the intersection of several literatures in economics, computer science, and statistics. These literatures consider several related but distinct tasks: synthesizing evidence across different domains, improving the quality of extrapolation from one domain to another, and quantifying the extent to which insights from one domain generalize to another. Our results are most closely related to this third strand.

The first objective, synthesizing results across different domains, is a particular focus of the literature on meta-analysis.[5] Our goal is instead to assess the cross-domain forecast accuracy of a model. These problems are related, and Meager (2019) and Meager (2022) in particular provide posterior predictive intervals for new domains in the context of her application. Unlike our approach, the predictive intervals reported in those papers rely on a parametric model for the distribution of effects across domains.

There is also a large literature that aims to extrapolate results from one domain to another. Within computer science, the literature on domain generalization (Blanchard et al. 2011 and Muandet et al. 2013) develops models that generalize well to new unseen domains (Zhou et al., 2021).[6] Similarly, several papers within economics (e.g., Hotz et al. 2005 and Dehejia et al. 2021) use knowledge about the distribution of covariates to extrapolate out-of-domain. In contrast, our focus is not on developing new models or algorithms with good out-of-domain guarantees, but rather on developing forecast intervals for the out-of-domain performance of models and algorithms that are used in practice.[7]

Finally, the literature on external validity studies the extent to which results obtained in one domain hold more generally. This paper does not focus on the generalizability of insights from randomized control trials (e.g. Deaton, 2010; Imbens, 2010) or laboratory experiments

---

[5]See Card and Krueger (1995), Benartzi et al. (2017), DellaVigna and Pope (2019), Hummel and Maedche (2019), Bandiera et al. (2021), Imai et al. (2020) and Vivalt (2020) among others.

[6]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 literature on *domain adaptation*, which aims to improve predictions when some data from the target domain is available – see Zhou et al. (2021).

[7]We focus on models that are estimated on a particular domain and ported to another without adjustment, but as discussed in Section P.1 our methods also apply to models that are re-estimated using some data from the target domain.

(e.g. Levitt and List, 2007; Al-Ubaydli and List, 2015), but instead on a model's generalizability across exchangeable domains.[8] Our use of exchangeability to construct bounds extends work on conformal inference (e.g. Vovk et al., 2005; Barber et al., 2021; Angelopoulos et al., 2022) by replacing the assumption of exchangeable observations with that of exchangeable domains.[9] Section 3.2 relaxes the exchangeability assumption; our results there connect to the literature on sensitivity analysis (e.g. Aronow and Lee, 2013; Andrews and Oster, 2019; Nie et al., 2021; Sahoo et al., 2022).

Finally, we join a small but growing body of work regarding the relative value of economic models and black box algorithms, and how the two approaches can be combined for better prediction and explanation of social science phenomena (Athey and Imbens, 2016; Fudenberg and Liang, 2019; Agrawal et al., 2020). Several recent papers compare the predictiveness of black box algorithms with that of more structured economic models.[10] While black box methods are often very effective given a large quantity of data from the domain of interest, our results suggest that they may be less effective for transferring across domains. Hofman et al. (2021) organizes recent work in this area and concludes that more work is needed on the question "how well does a predictive model fit to one data distribution generalize to another?" Our paper takes an important step in this direction.

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 $\mathcal{Y}$. A *prediction rule* is any mapping $\sigma : \mathcal{X} \to \mathcal{Y}$, and Σ denotes the set of all such mappings. Prediction rules are evaluated 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.

---

[8]Another set of papers study the generalizability of instrumental variables estimates (e.g. Angrist and Fernández-Val, 2013; Bertanha and Imbens, 2020) and causal effects (e.g. Pearl and Bareinboim, 2014; Park et al., 2023).

[9]This also differentiates our work from the out-of-distribution prediction literature in computer science (Shen et al., 2021), which bounds expected transfer error when the test and training distributions are close.

[10]See 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).

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 evaluated on a test sample S_{test} of observations drawn iid from the same distribution p. We generalize this framework by supposing that the analyst's choice of prediction rule is based on *meta-data*

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

consisting of samples across different domains,[11] and that the prediction rule is evaluated on a sample S_{n+1} from a new domain. Our main assumption (which we relax in Section 3.2) is that samples are generated iid from a meta-distribution $\mu \in \Delta(\mathcal{P} \times \mathbb{N})$ over distributions $\mathcal{P} \equiv \Delta(\mathcal{X} \times \mathcal{Y})$ and sample sizes $\mathbb{N}$. We define $S_d \sim_{iid} \mu$ to mean that each sample S_d is generated by first independently drawing a distribution and sample size $(P_d, m_d) \sim \mu$, and then independently drawing m_d observations $(x, y) \sim_{iid} P_d$.[12] This framework nests the usual one when μ assigns probability 1 to a single distribution in $p \in \mathcal{P}$ or when μ assigns probability 1 to $m = 1$, but we are primarily interested in settings where neither of these is the case.

Based on the metadata $\mathbf{M}$, the analyst chooses a prediction rule σ using a *decision rule*, which is a (potentially randomized) map ρ from the set of all finite metadata into the set of distributions on Σ. To develop forecast intervals for performance on the target sample S_{n+1}, it will be useful to select σ using strict subsets of the observed samples in the meta-data. For this purpose we let $\mathbf{M}_{\mathcal{T}} \equiv (S_t)_{t \in \mathcal{T}}$ denote the samples in metadata $\mathbf{M}$ that are indexed to a selection of training domains $\mathcal{T} \subseteq \{1, \ldots, n\}$.

We focus on two classes of decision rules ρ:

Estimation of an Economic Model/Empirical Risk Minimization. Let Σ^* be a set of prediction rules derived from an economic model (for example, Expected Utility or Cumulative Prospect Theory) and let ρ_{Σ^*} be a decision rule that maps training samples $\mathbf{M}_{\mathcal{T}}$ to the prediction rule in Σ^* that minimizes the average error on those samples (and breaks ties

---

[11]The domains are distinguished in $\mathbf{M}$. So, for example, the meta-data corresponding to observation of samples $S_1 = \{z_1\}$ and $S_2 = \{z_2, z_3\}$ is $\mathbf{M} = (S_1, S_2) = \{\{z_1\}, \{z_2, z_3\}\}$ rather than $\mathbf{M} = \{z_1, z_2, z_3\}$.

[12]Section 3.3 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.

according to a pre-specified rule):[13]

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

Training a Black Box Algorithm. 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.** Our main results consider a random index set of training samples $\mathbf{T}$, drawn from the uniform distribution over all subsets of $\{1, \ldots, n\}$ of a fixed (user-selected) size $r < n$. Let $\mathbf{M_T}$ denote this random selection of samples from $\mathbf{M}$. Below we propose several measures for transferability, which evaluate the performance of the prediction rule $\rho(\mathbf{M_T})$ on a new target sample $S_{n+1} \sim \mu$.

The first measure is 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*[14]

$$e(\rho(\mathbf{M_T}), S_{n+1}). \tag{2}$$

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 $\mathcal{R}$.[15]

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

$$\frac{e(\rho(\mathbf{M_T}), S_{n+1})}{\min_{\hat{\rho} \in \mathcal{R}} e(\hat{\rho}(S_{n+1}), S_{n+1})}. \tag{3}$$

---

[13]Under standard continuity and compactness assumptions the minimum is attained.

[14]Here and throughout the paper, when we write a transfer error for a random prediction rule, we mean the error with respect to a single draw of σ from the distribution.

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

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 $\mathcal{R}$. We expect the normalized transfer error to be bounded below by 1 for all decision rules used in practice.[16][17]

Our third and final measure evaluates 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*

$$\frac{e(\rho(\mathbf{M_T}), S_{n+1})}{e(\rho(S_{n+1}), S_{n+1})}. \tag{4}$$

This quantity tells us how many times larger a decision rule's transfer error is than its in-sample error $e(\rho(S_{n+1}), S_{n+1})$ in the test sample. We again expect this ratio to be bounded below by 1 for rules used in practice (see footnote 16); it is equal to 1 only if the transfer error of the model is the same as its in-sample error.[18] 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.

3. Theoretical Results

3.1. **IID Baseline.** Our goal is to develop forecast intervals for the measures defined in Section 2.2, i.e., interval-valued functions of the meta-data which cover the quantity with prescribed probability, regardless of the distribution μ that governs samples across domains. In many applications only a limited number of domains will be observed, so our focus is on finite-sample results.

In this section, we let $e_{\mathbf{T}, n+1}$ denote the transfer error of the prediction rule $\rho(\mathbf{M_T})$ on the target sample S_{n+1} as defined in (2). Our results and proof techniques hold identically for the other two measures, as well as the substantially broader class of random variables discussed in Section 3.3. We use the metadata $\mathbf{M} = (S_1, \ldots, S_n)$ to construct a forecast

---

[16] When $\mathcal{R}$ consists of empirical risk minimization rules (see Example 2.1), 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 $\mathcal{R}$. Normalized transfer error is not bounded below by 1 for all decision rules, for example random draws from Σ independent of the meta-data, but we do not expect such rules to be used in practice.

[17] Normalized transfer error compares an out-of-sample object in the numerator to an in-sample object in the denominator, which "stacks 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.

[18] 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.

interval for $e_{\mathbf{T},n+1}$ in the following way. For any choice of training samples $\mathcal{T} \subseteq \{1,\ldots,n\}$ and test sample $d \in \{1,\ldots,n\}\backslash\mathcal{T}$ from $\mathbf{M}$, let

$$e_{\mathcal{T},d}^{\mathbf{M}} = e(\rho(\mathbf{M}_{\mathcal{T}}), S_d)$$

be the (observed) transfer error from samples in $\mathcal{T}$ to sample d. Further define $\mathbb{T}_{s,t}$ to be the set of all vectors of length s that consist of distinct elements from $\{1,\ldots,t\}$. Then $\mathbb{T}_{r+1,n}$ is the set of all possible choices of r training samples and a single target sample from the metadata, and

$$\left\{ e_{\mathcal{T},d}^{\mathbf{M}} : (\mathcal{T},d) \in \mathbb{T}_{r+1,n} \right\}$$

is the pooled sample of transfer errors. Finally, define

$$F_{\mathbf{M}} = \frac{(n-r-1)!}{n!} \sum_{(\mathcal{T},d)\in\mathbb{T}_{r+1,n}} \delta_{e_{\mathcal{T},d}^{\mathbf{M}}} \tag{5}$$

to be the empirical distribution of transfer errors in this pooled sample (where δ denotes the Dirac measure).

Definition 5 (Upper and Lower Quantiles)**.** *For any distribution P let $\overline{Q}_\tau(P) = \inf\{b : P((-\infty,b]) \geq \tau\}$ and $\underline{Q}_\tau(P) = \sup\{b : P([b,\infty)) \geq 1-\tau\}$ denote the upper and lower τth quantiles, respectively.*

These quantiles coincide for continuously distributed variables.

Definition 6 (Quantiles of $F_{\mathbf{M}}$)**.** *For any $\tau \in (0,1)$, let $\bar{e}_\tau^{\mathbf{M}} \equiv \overline{Q}_\tau(F_{\mathbf{M}})$ and $\underline{e}_\tau^{\mathbf{M}} \equiv \underline{Q}_{1-\tau}(F_{\mathbf{M}})$ be the τth upper quantile and $(1-\tau)$th lower quantile of the empirical distribution of transfer errors in the pooled sample.*

These quantiles can be used to construct a valid forecast interval for the transfer error on the target sample:

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

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_\tau^{\mathbf{M}}\right) \geq \tau\left(\frac{n-r}{n+1}\right), \tag{6}$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in \left[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}\right]\right) \geq 2\tau\left(\frac{n-r}{n+1}\right) - 1.$$

9

Thus $\left(-\infty, \bar{e}_\tau^{\mathbf{M}}\right]$ is a level-$\left(\frac{\tau(n-r)}{n+1}\right)$ one-sided forecast interval for $e_{\mathbf{T},n+1}$, and $\left[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}\right]$ is a level-$\left(2\tau\left(\frac{n-r}{n+1}\right) - 1\right)$ forecast interval for $e_{\mathbf{T},n+1}$. The parameters r and τ are choice variables. 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 r determines how many samples in the meta-data are used for training versus testing. Larger choices of r mean that the model 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.

The next result shows the extent to which the guarantees in Proposition 1 are tight.

Claim 1. *Assume that* $\left(e_{\mathcal{T},d}^{\mathbf{M}} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n+1}\right)$ *almost surely has no ties. Then*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_\tau^{\mathbf{M}}\right) \leq \tau\left(\frac{n-r}{n+1}\right) + \frac{r+1}{n+1} + \frac{(n-r)!}{(n+1)!}.$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in \left[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}\right]\right) \leq 2\tau\left(\frac{n-r}{n+1}\right) - 1 + 2\left(\frac{r+1}{n+1} + \frac{(n-r)!}{(n+1)!}\right).$$

To gain intuition for the intervals in Proposition 1, fix a realization of the unordered set $\{S_1, \ldots, S_n, S_{n+1}\}$. Because by assumption all samples are exchangeable, the realization of $e_{\mathbf{T},n+1}$ (conditional on $\{S_d\}_{d=1}^{n+1}$) is a uniform draw from

$$\left\{e_{\mathcal{T},d}^{\mathbf{M}} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n+1}\right\}. \tag{7}$$

If we let e_τ^* denote the upper τ-th quantile of this empirical distribution, then by definition

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq e_\tau^* \mid \{S_d\}_{d=1}^{n+1}\right) \geq \tau. \tag{8}$$

In the case $r = 1$ where precisely one sample is used for training, the set of pooled errors (7) is the shaded cells in Figure 1 (either yellow or blue), and the inequality in (8) says that the probability that the value of a randomly drawn cell falls below the τth upper quantile of cells is at least τ.

The analyst does not observe the target sample S_{n+1}, and so does not know e_τ^*. As a surrogate, we use $\bar{e}_\tau^{\mathbf{M}}$, the τth upper quantile of the pooled sample of errors when transferring across samples in $\mathbf{M}$. In Figure 1, the probability that $e_{\mathbf{T},n+1} \leq \bar{e}_\tau^{\mathbf{M}}$ is the probability that the value of a randomly drawn shaded cell (yellow or blue) falls below the τth quantile of

FIGURE 1. *Matrix of transfer errors when training on one domain (row) and testing on another (column).*

the yellow cells. By a straightforward counting argument,

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_\tau^{\mathbf{M}} \mid \{S_i\}_{i=1}^{n+1}\right) \geq \tau \binom{n}{r+1} \bigg/ \binom{n+1}{r+1} = \tau \left(\frac{n-r}{n+1}\right).$$

Applying the law of iterated expectations (with respect to the sample $\{S_i\}_{i=1}^{n+1}$) yields the one-sided forecast interval in (6), and a similar argument yields the two-sided forecast interval.

3.2. **Relaxing the IID Assumption.** Our results so far assume that the distributions governing the different samples S_d are themselves independent and identically distributed. This assumption is not always appropriate. For example, if the samples in the metadata are from experiments run at different locations, the iid assumption fails if there is selection bias over where experiments are run.[19] We now relax this assumption to allow the distribution governing the training samples and the distribution governing the target sample to be drawn from different meta-distributions.

Specifically, suppose that the analyst's metadata consists of samples $S_1, \ldots, S_n \sim_{iid} \mu$ as in our main model, but S_{n+1} is independently drawn from some other density ν. Let

$$\omega(S) = \frac{\nu(S)}{\mu(S)}$$

denote their likelihood ratio. As before, $e_{\mathbf{T},n+1}$ is the transfer error when training on r samples drawn uniformly at random from $\{S_1, \ldots, S_n\}$, and testing on S_{n+1}.

---

[19]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.

We again construct a forecast interval for $e_{\mathbf{T},n+1}$ using the pooled sample of transfer errors across samples in the metadata, $\{e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T},d) \in \mathbb{T}_{r+1,n}\}$, giving different probabilities to each $e^{\mathbf{M}}_{T,d}$ instead of uniform weights. Under the iid assumption, each sample in the metadata is equally representative of the training and target distributions, while when we relax that assumption whether a sample S_d is more representative of the training or testing distribution depends on its relative likelihood under ν and μ.

A crucial quantity is the following:

Definition 7. *For every domain $d \in \{1, \ldots, n\}$, define*

$$W_d = \frac{(n-r-1)!}{(n-1)!} \frac{\omega(S_d)}{\sum_{j=1}^{n} \omega(S_j)}. \tag{9}$$

To interpret this quantity, consider an alternative data-generating process for the metadata (mimicking the larger environment) where for some permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, the samples $S_{\pi(1)}, \ldots, S_{\pi(n-1)} \sim_{iid} \mu$ while $S_{\pi(n)} \sim \nu$. Fix a realization of the metadata $(S_1, \ldots, S_n)$, and suppose the analyst does not observe the permutation π. Let Π denote the set of all permutations on $\{1, \ldots, n\}$, and for any vector of sample indices $(t_1, \ldots, t_r, d)$ let

$$\Pi_{(t_1, \ldots, t_r, d)} = \{\pi \in \Pi : (\pi(1), \ldots, \pi(r)) = (t_1, \ldots, t_r) \text{ and } \pi(n) = d\}$$

denote the permutations that specify $(t_1, \ldots, t_r)$ for training and d as the target. Then conditional on a realization of the metadata $(S_1, \ldots, S_n)$, the probability that $(S_{t_i})_{i=1}^{r}$ are the training samples and S_d is the test sample is[20]

$$\frac{\sum_{\pi \in \Pi_{(t_1, \ldots, t_r, d)}} \left(\nu(S_{\pi(n)}) \cdot \prod_{j=1}^{n-1} \mu(S_{\pi(j)}) \right)}{\sum_{\pi \in \Pi} \left(\nu(S_{\pi(n)}) \cdot \prod_{j=1}^{n-1} \mu(S_{\pi(j)}) \right)} = \frac{\sum_{\pi \in \Pi_{(t_1, \ldots, t_r, d)}} \omega(S_{\pi(n)})}{\sum_{\pi \in \Pi} \omega(S_{\pi(n)})}$$

$$= \frac{(n-r-1)! \cdot \omega(S_d)}{(n-1)! \cdot \sum_{j=1}^{n} \omega(S_j)} = W_d.$$

This quantity depends only on the identity of the target sample d, and not on the identity of the training samples $t_1, \ldots, t_r$. Finally, let

$$F^{\omega}_{\mathbf{M}} = \sum_{(\mathcal{T},d) \in \mathbb{T}_{r+1,n}} W_d \cdot \delta_{e^{\mathbf{M}}_{\mathcal{T},d}}$$

be the weighted empirical distribution of transfer errors, where each sample d is weighted according to W_d. When the two meta-distributions μ and ν are identical as in our main

---

[20]This is known as weighted exchangeability; see (Tibshirani et al., 2019).

model, then $W_d \equiv (n - r - 1)!/n!$ for every domain d, so the distribution $F_{\mathbf{M}}^{\omega}$ is simply $F_{\mathbf{M}}$ as defined in (5).

Definition 8 (Quantiles of $F_{\mathbf{M}}^{\omega}$). *For any likelihood ratio $\omega(\cdot)$ and quantile $\tau \in (0, 1)$, define $\bar{e}_{\tau}^{\mathbf{M},\omega} = \overline{Q}_{\tau}(F_{\mathbf{M}}^{\omega})$ and $\underline{e}_{\tau}^{\mathbf{M},\omega} = \underline{Q}_{1-\tau}(F_{\mathbf{M}}^{\omega})$ to be, respectively, the τth upper quantile and $(1 - \tau)$th lower quantile of the weighted distribution of transfer errors in the pooled sample.*

Theorem 1. *For any $\tau \in (0, 1)$,*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_{\tau}^{\mathbf{M},\omega}\right) \geq \tau \cdot \frac{n - r}{n} \mathbb{E}\left[\frac{\sum_{j=1}^{n} \omega(S_j)}{\sum_{j=1}^{n+1} \omega(S_j)}\right],$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in [\underline{e}_{\tau}^{\mathbf{M},\omega}, \bar{e}_{\tau}^{\mathbf{M},\omega}]\right) \geq 2\tau \cdot \frac{n - r}{n} \mathbb{E}\left[\frac{\sum_{j=1}^{n} \omega(S_j)}{\sum_{j=1}^{n+1} \omega(S_j)}\right] - 1.$$

This result strictly generalizes Proposition 1, since when $w(\cdot)$ is the identity then $\bar{e}_{\tau}^{\mathbf{M},\omega} = \bar{e}_{\tau}^{\mathbf{M}}$ and $\underline{e}_{\tau}^{\mathbf{M},\omega} = \underline{e}_{\tau}^{\mathbf{M}}$, and the bounds in this theorem reduce to those given in Proposition 1.

If the analyst does not know ω, but can uniformly upper and lower bound it across samples, the following result applies.

Definition 9 (Bounded Likelihood-Ratios). *For any $\Gamma \geq 1$, let $\mathcal{W}(\Gamma)$ be the class of density ratios that satisfy $\omega(S) \in [\Gamma^{-1}, \Gamma]$ for all samples S.*

Corollary 1. *Suppose $\omega \in \mathcal{W}(\Gamma)$. Then*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_{\tau}^{\mathbf{M},\omega}\right) \geq \tau \left(\frac{n - r}{n + \Gamma^2}\right),$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in [\underline{e}_{\tau}^{\mathbf{M},\omega}, \bar{e}_{\tau}^{\mathbf{M},\omega}]\right) \geq 2\tau \left(\frac{n - r}{n + \Gamma^2}\right) - 1.$$

When there is no natural bound for w, it can still be possible in some cases to compare the transferability of two methods. Fix any two decision rules ρ_1 and ρ_2 of interest, and let $\left[\underline{e}_{1,\tau}^{\mathbf{M},\omega}, \bar{e}_{1,\tau}^{\mathbf{M},\omega}\right]$ and $\left[\underline{e}_{2,\tau}^{\mathbf{M},\omega}, \bar{e}_{2,\tau}^{\mathbf{M},\omega}\right]$ denote their respective forecast intervals. For each rule ρ_i, let

$$\bar{e}_{i,\tau}^{\mathbf{M}}(\Gamma) = \sup_{\omega \in \mathcal{W}(\Gamma)} \bar{e}_{i,\tau}^{\mathbf{M},\omega} \tag{10}$$

be the worst-case upper bound across likelihood ratios in $\mathcal{W}(\Gamma)$. As shown in Appendix Q this quantity can be computed in $O(n^{r+1})$ time.

Definition 10 (Worst-Case Dominance). *Say that ρ_1 worst-case-upper-dominates ρ_2 at the τ-th quantile if*

$$\bar{e}_{1,\tau}^{\mathbf{M}}(\Gamma) \leq \bar{e}_{2,\tau}^{\mathbf{M}}(\Gamma) \quad \forall \Gamma \geq 1.$$

That is, ρ_1 worst-case-upper-dominates ρ_2 at the τ-th quantile if for every bound Γ, the worst-case upper bound for decision rule ρ_2 exceeds the worst-case for upper bound for decision rule ρ_1.

We can strengthen this comparison by requiring the upper bound of the forecast interval for ρ_1 to be smaller than the upper bound of the forecast interval for ρ_2 pointwise for each $\omega \in \mathcal{W}(\Gamma)$, rather than simply comparing worst-case upper bounds.

Definition 11 (Everywhere Dominance). *Say that ρ_1 everywhere-upper-dominates ρ_2 at the τ-th quantile if*

$$\bar{e}_{1,\tau}^{\mathbf{M},\omega} \leq \bar{e}_{2,\tau}^{\mathbf{M},\omega} \quad \forall \Gamma \geq 1, \forall \omega \in \mathcal{W}(\Gamma).$$

Many decision rules will not be comparable under either of these definitions, but both of these orders do have bite in our application. The even stronger requirement that $\bar{e}_{1,\tau}^{\mathbf{M},\omega} \leq \underline{e}_{2,\tau}^{\mathbf{M},\omega}$, i.e., that the upper bound of ρ_1's forecast interval is smaller than the lower bound of ρ_2's forecast interval, is likely too stringent to be useful in practice.[21]

3.3. **Discussion. What is a domain?** The specified domains determine the transfer question the analyst is interested in and the content of the iid sampling assumption. Suppose, for instance, that the meta-data consists of experimental results from multiple papers, where each paper reports results from experiments at multiple sites. If each site is treated as a separate domain, then iid sampling corresponds to drawing further sites, some from the papers already observed and others from as-yet-unobserved papers. Alternatively, if each paper is a separate domain, then iid sampling corresponds to drawing new papers, each with its own sites.

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

---

[21]This stronger order has bite only when the transfer error for ρ_1 across "the most dissimilar" training and testing domains is lower than the transfer error for ρ_2 for "the most similar" training and testing domains, which seems unlikely.

lead to different rankings. This is because normalized transfer error penalizes a model more for performing worse on samples where the best achievable error is low 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.[22] If we also had data on how well such extrapolations 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.

A complementary question is how to estimate a "best" prediction rule for transfer. One might use cross-validation across domains to tune parameters, build a Bayesian hierarchical model that explicitly models a distribution over domains, or choose parameters with worst-case guarantees (as in the literature on distributionally robust optimization, e.g. Rahimian and Mehrotra 2019). The development of estimation procedures that lead to better transfer performance is interesting, but it is distinct from providing forecast intervals for transfer performance. In our application we examine the transfer performance of estimation procedures that have previously been used in the literature. However, each of the approaches mentioned above can be formalized as a decision rule (see Appendix P.2 for details), so our results imply forecast intervals for their transfer performance.

Non-Prediction Targets. While we have framed our discussion in terms of prediction, we can apply our results to many other problems by re-defining $e_{\mathcal{T},d}^{\mathbf{M}}$. For instance, if we were interested in estimating some parameter across domains, we could take $e_{\mathcal{T},d}^{\mathbf{M}}$ to measure the difference between the estimates based on $\mathbf{M}_{\mathcal{T}}$ and those based on S_d. In this case, Proposition 1 provides a predictive interval for the estimate in a new domain. Alternatively,

---

[22]For instance, 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.

we could consider approaches designed to extrapolate across domains, as in Hotz et al. (2005)'s extrapolation of the efficacy of job training programs to different locations using differences in the distribution of covariates. Here we could define $e^{\mathbf{M}}_{\mathcal{T},d}$ as the difference between the effect predicted based on the training locations $\mathcal{T}$ and the covariate distribution at location d and the effect estimated at location d. In this case, Proposition 1 provides a forecast interval for the cross-domain extrapolation error. More generally, Proposition 1 provides a forecast interval for any function of the data in the training and target domains; See Online Appendix P.1.

Number of Domains Versus Observations. 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 $e_{\mathbf{T},n+1}$: In the limit of infinitely many observations per domain, the error $e_{\mathbf{T},n+1}$ 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 $e_{\mathbf{T},n+1}$ 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 $e_{\mathbf{T},n+1}$, but instead allows this distribution to be estimated more precisely.

Counterfactual Predictions. One important way that economic models are sometimes used is to form predictions for outcomes under policy changes that have yet to be implemented. For instance, McFadden (1974) predicted the demand impacts of the then-new BART rapid transit system in the San Franciso Bay Area, and Pathak and Shi (2013) predicted demand for schools under changes to the Boston school choice system. If we view the pre-intervention period as the "training" data, and the post-intervention period as the "test" data, the two seem unlikely to satisfy our exchangeability assumption. However, our results do extend to such settings when there are iid *pairs* of train-test domains, where the training domain represents the pre-intervention data used to form a counterfactual prediction, and the test domain represents the post-intervention data used to evaluate the prediction.

4. Application

To illustrate our methods, we evaluate the transferability of predictions of certainty equivalents for binary lotteries, where the domains correspond to different subject pools. Section

4.1 describes our metadata, and Section 4.2 describes the decision rules we consider. Section 4.3 conducts "within-domain" out-of-sample tests, where the training and test data are drawn from the same domain. Section 4.4 constructs forecast intervals for transfer error, normalized transfer error, and transfer deterioration.

4.1. **Data.** Our metadata consists of 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 R.2, we repeat our analysis with each paper treated as a separate domain, and show that the results are qualitatively similar.[23] Our samples range in size from 72 observations to 8906 observations, with an average of 2752.7 observations per sample.[24] We convert all prizes to dollars using purchasing power parity exchange rates (from OECD 2022) in the year of the paper's publication

Within each sample, observations take the form $(z_1, z_2, p; y)$, where z_1 and z_2 denote the possible prizes of the lottery (and we adopt the convention that $|z_1| > |z_2|$), p is the probability of z_1, 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.[25] Since different subjects report different certainty equivalents for the same lottery, the best achievable error 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 parameterized by $\eta \geq 0$. For $\eta \neq 1$, define

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

---

[23]In both cases, the domains sometime combines observations from different experimental treatments. In Etchart-Vincent and l'Haridon (2011), we pool reported certainty equivalents across three payment conditions: real losses, hypothetical-losses, and losses-from-initial-endowment.

[24]Online Appendix R.1 describes our data sources in more detail.

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

17

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

$$\sigma_\eta(z_1, z_2, p) = v_\eta^{-1}\big(p \cdot v_\eta(z_1) + (1 - p) \cdot v_\eta(z_2)\big).$$

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 $\mathbf{M}_\mathcal{T}$ 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).

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 (z_1, z_2, p) is assigned a utility

$$w(p)v(z_1) + (1 - w(p))v(z_2)$$

where

$$v(z) = \begin{cases} z^\alpha & \text{if } z \geq 0 \\ -(-z)^\beta & \text{if } z < 0 \end{cases} \tag{11}$$

is a value function for money, and

$$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1 - p)^\gamma} \tag{12}$$

is a probability weighting function.

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

$$\sigma_{(\alpha, \beta, \gamma, \delta)}(z_1, z_2, p) = v^{-1}\big(w(p)v(z_1) + (1 - w(p))v(z_2)\big).$$

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 $\mathbf{M}_\mathcal{T}$ 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 (11);

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.[26] 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 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.[27] 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 radial basis function 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 $\vec{w} = (\mathbb{K} + \lambda I_N)^{-1}\vec{y}$, where $\mathbb{K}$ is the $N \times N$ matrix whose (i,j)-th entry is $\kappa(x_i, x_j)$, I_N is the $N \times N$ identity matrix, and $\vec{y} = (y_1, \ldots, y_N)'$ is the vector of observed outcomes in the training data. The estimated prediction rule is $\sigma(x) = \sum_{i=1}^N w_i \kappa(x, x_i)$.

There are at least two approaches for cross-validating hyper-parameters such as the size of the trees in the random forest algorithm. First, in settings with multiple training domains one can cross-validate across training domains. This procedure is not relevant to our analysis in Section 4.4, which considers transfer from a single training domain to a single test domain, but we use it in Appendix R.7 when we consider multiple training domains. Second, one can cross-validate across observations within the training domains. Since we are interested in cross-domain performance, rather than within-domain performance, it is not clear that this will improve performance, and indeed we find that choosing the hyper-parameters via within-domain cross-validation leads to worse transfer performance than using default values. Thus

---

[26]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.

[27]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 split is chosen to greedily minimize mean squared error.

in our main analysis with a single training domain, we set all hyper-parameters to default values.[28]

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.[29] 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 2 reports the CDF of tenfold cross-validated errors for the random forest, kernel regression, EU, and CPT.[30]
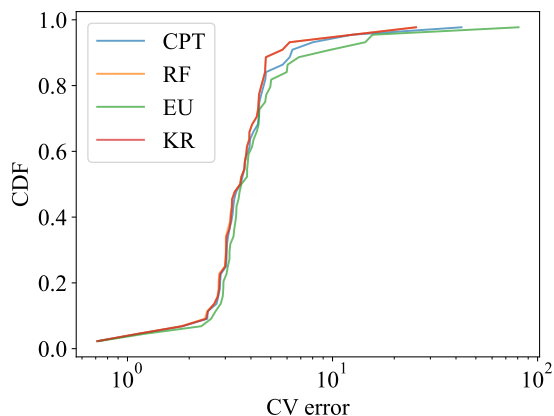

FIGURE 2. *CDF of Cross-Validated Errors. (The CDFs for kernel regression and random forest overlap.)*

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.[31]

---

[28]Specifically, we set $\lambda = 1$ and $\gamma = 1/(\#\text{covariates}) = 1/3$ in the kernel regression algorithm. See Pedregosa et al. (2011) and Chapter 14 of Murphy (2012) for further reference. For the random forest model, we set the maximum depth to none, which means that nodes are expanded until all leaves are pure or until all leaves contain 1 samples.

[29]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.

[30]Online Appendix R.3 shows that the CDFs for in-sample errors are likewise very close.

[31]The numbers in Table 1 are very similar if we normalized by the kernel regression error instead.

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 relative to random forest.*

These results suggest that the different prediction methods we consider are comparable for within-domain prediction, with the black boxes performing slightly better. But the 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 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 the 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 $r = 1$ of these to use as the training domain. This choice of r corresponds to the question, "If I draw one domain at random, and then try to generalize to another domain, how well do I do?"

Figure 3 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 $\tau = 0.95$, 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.[32] Applying Proposition 1, these are 81% 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 R.5, including a 91% forecast interval.

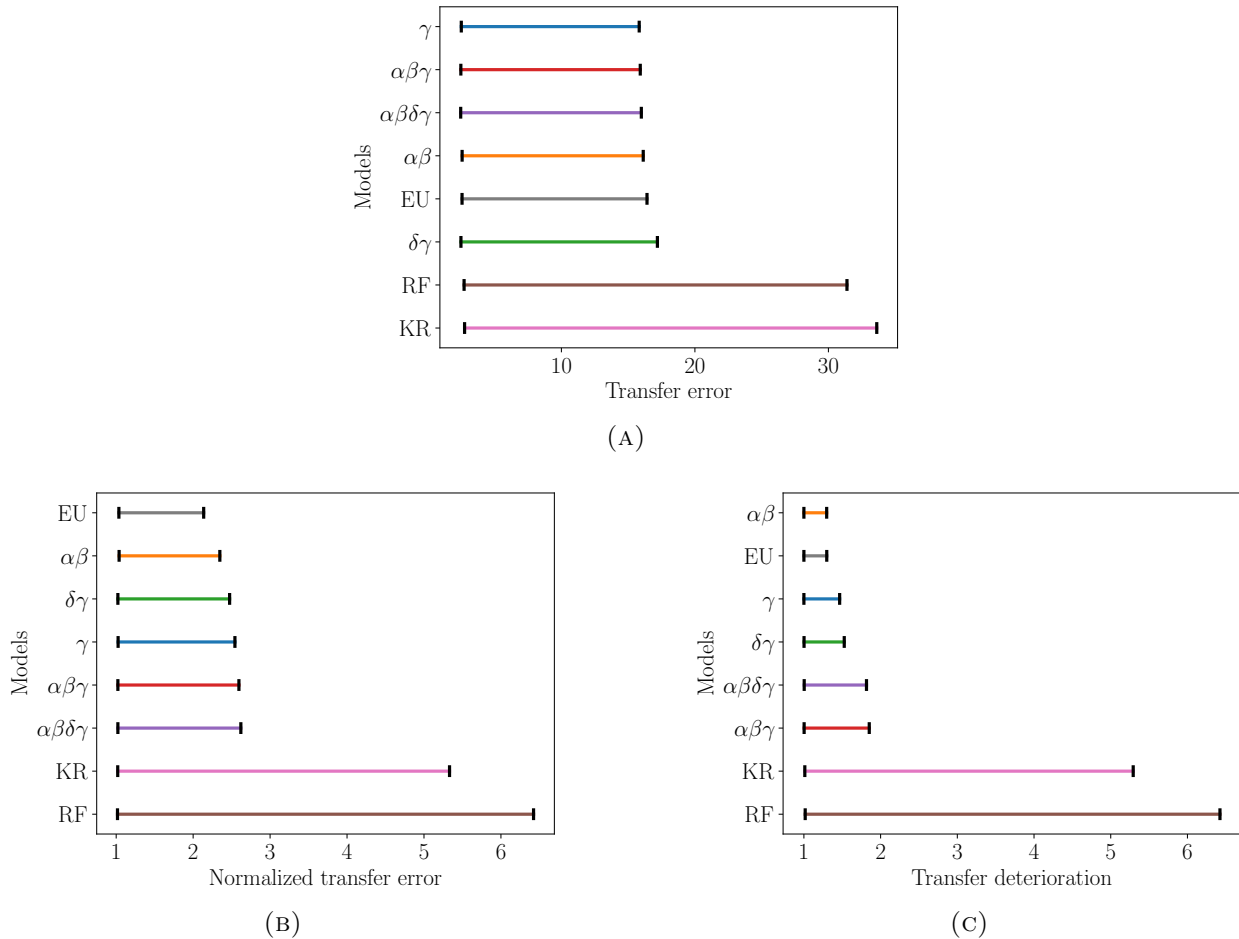[32]See Table 5 in Appendix R.4 for the exact numbers.

FIGURE 3. *81% (n=44, τ = 0.95) 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.*

Our main takeaway from Figure 3 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.

Panel (a) of Figure 3 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 have very different performances in the test domain, and the prediction rules selected by the economic models generalize substantially better.

Panel (c) of Figure 3 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 $CPT(\alpha, \beta)$, which has lower transfer deterioration than the single-parameter EU model and also $CPT(\gamma)$).

All of 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 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. Section 5 provides confidence intervals for different population quantities, including quantiles of the transfer error distribution and the expected transfer error, whose width we do expect to vanish as the number of domains grow large. There, we find similar conclusions with regards to the relative performance of the black box algorithms and economic models.

The appendix provides several robustness checks and complementary analyses. Online Appendix R.5 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 R.6 provides 81% 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 lower. Online Appendix R.7 considers an alternative choice for the number of training domains, setting $r = 3$ instead of $r = 1$. 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. Online Appendix R.2 provides forecast intervals when each of the 14 papers is treated as a different domain; once again the black box methods transfer worse than the economic models do.

We next use our theoretical results from Section 3.2 to study the consequences of relaxing the iid assumption in our comparison of $CPT(\alpha, \beta, \delta, \gamma)$ and RF. Since the main differences

observed above concerned the upper bounds of our forecast intervals, we limit attention to $\tau \geq 0.5$, and compare the methods in terms of worst-case and everywhere upper-dominance with respect to all three measures of the transfer performance. These results are summarized in Table 2.

Type	transfer error	normalized transfer error	transfer deterioration
Worst-case dominance	$\tau \geq 0.5$	$\tau \geq 0.5$	$\tau \geq 0.5$
Everywhere dominance	$\tau \geq 0.954$	$\tau \geq 0.866$	$\tau \geq 0.647$

TABLE 2. *Comparison between CPT and RF in terms of worst-case and everywhere upper-dominance. Each cell gives the range of τ at which CPT dominates RF.*

Table 2 shows that CPT worst-case-upper-dominates RF at all quantiles $\tau \geq 0.5$ and for all three measures of transfer performance. Hence, our finding that the upper tail of transfer errors is larger for RF than for CPT is fully robust to relaxing the assumption that the training and test domains are drawn from the same distribution, provided that for a given degree of relaxation we are comfortable comparing the upper bound for one method to the upper bound for the other. In Appendix R.8, we provide a more detailed view of worst-case-upper-dominance by plotting $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ as functions of τ and Γ, respectively.

We can also consider the more demanding everywhere-upper-dominance criterion, which asks what happens if we relax our iid sampling assumption in a way which is as favorable to RF (and as unfavorable to CPT) as possible. We find a substantial degree of robustness even under this highly demanding criterion: CPT everywhere-upper-dominates RF in transfer error for all $\tau \geq 0.954$, everywhere-dominates in normalized transfer error for $\tau \geq 0.866$, and everywhere dominates in transfer deterioration for $\tau \geq 0.647$.

4.5. **Do black boxes transfer poorly because they are too flexible?** One tempting explanation of our empirical findings is that because the black boxes are more flexible than the economic models, they can learn idiosyncratic details that do not generalize across subject pools, such as that some subject pools tend to value lotteries depending on specific digits they contain.[33] This would lead the black boxes to have better within-domain prediction for those subject pools, but worse transfer performance if the regularity does not generalize across subject pools.

---

[33]For example, Fortin et al. (2014) find that in neighborhoods with a higher than average percentage of Chinese residents, homes with address numbers ending in "4" are sold at a 2.2% discount and those ending in "8" are sold at a 2.5% premium.

While the 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 based on 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 4). 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.

These observations show that flexible prediction methods do not always transfer poorly, although they do perform poorly in certain kinds of transfer prediction tasks. The next section explores one potential explanation.

4.6. **Two kinds of transfer problems.** Our framework allows the distribution P governing the training sample and the distribution P' governing the test sample to differ. 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 $P'_{Y|X}$ (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, the conditional distributions $P_{Y|X}$ and $P'_{Y|X}$ might be the same, but the marginal distributions over the feature space could differ across domains, e.g., if different kinds of lotteries are used in different domains (known as *covariate shift*).

Our findings in Figure 4 suggest that black boxes do 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,
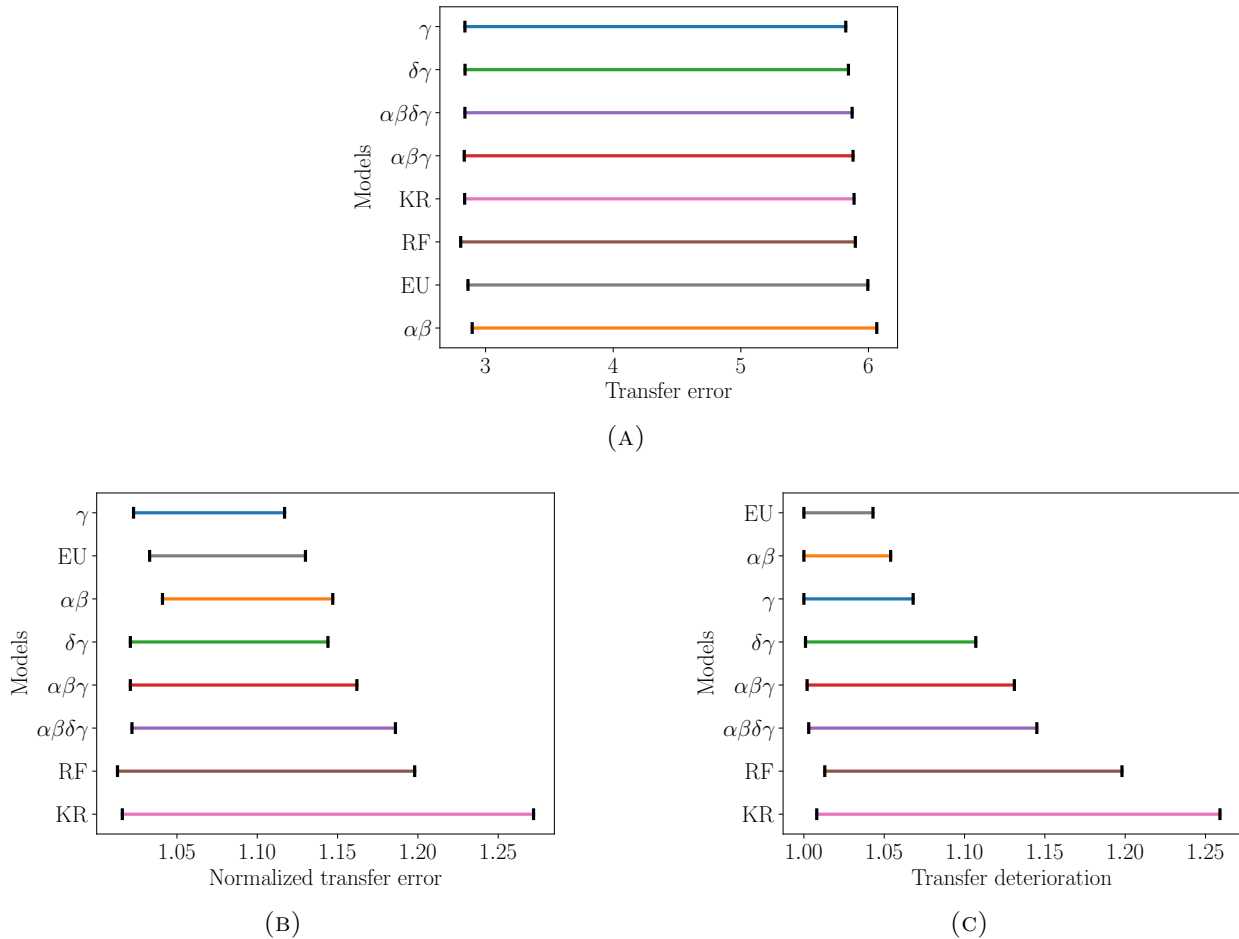
(A)

(B)

(C)

Figure 4. 78% (n=30, τ=0.95) forecast intervals using samples in l'Haridon and Vieider (2019).

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 an algorithm that hasn't identified a structure that relates behavior across 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

$(z_1, z_2, p) = (10, 0, 1/2)$ and certainty equivalent $y = 3$. In domain 2, the distribution is degenerate at the lottery $(z_1, z_2, p) = (10, 0, 1/2)$ and certainty equivalent $y = 4$. In domain 3, the distribution is degenerate at a new lottery $(z_1, z_2, p) = (20, 10, 1/10)$ 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, $e(\rho^{RF}(\mathbf{M}_\mathcal{T}), S_d)/e(\rho^{CPT}(\mathbf{M}_\mathcal{T}), S_d)$, given information about the training and test lotteries but not about the distribution of certainty equivalents in either sample. If the relative performance of these methods depended importantly on behavioral shifts in the two domains—i.e., a change in the distribution of certainty equivalents for the same lotteries— then we would expect prediction of the relative performance based on lottery information alone to be poor. We find instead that lottery information has substantial predictive power for this ratio.

For each sample $S = \{(z_{1,i}, z_{2,i}, p_i; y_i)\}_{i=1}^m$, we consider the following features:

- the mean, standard deviation, max, and min value of z_1 among the lotteries in S
- the mean, standard deviation, max, and min value of z_2 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 $pz_1 + (1 - p)z_2$ among the lotteries in S
- the size of S
- an indicator variable for whether $z_1, z_2 \geq 0$ for all lotteries in S

We consider three possible feature sets: (a) *Training Only,* which includes all features derived from the training sample $\mathbf{M}_\mathcal{T}$; (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 $\mathbf{M}_\mathcal{T}$ and the test sample S_d. We evaluate two prediction methods: OLS and a random forest algorithm. Table 3 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

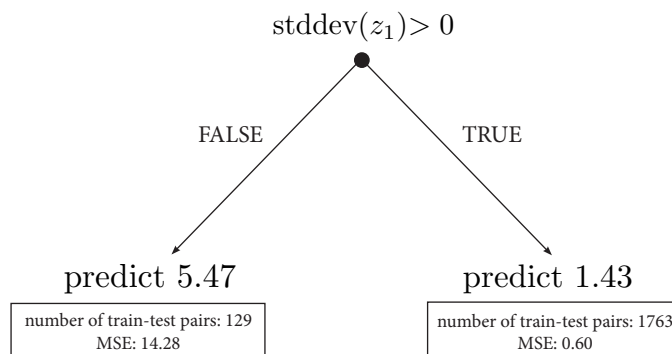TABLE 3. *Cross-Validated MSE*

FIGURE 5. *Best 1-split decision tree based on training and test features.*

The best constant prediction achieves a mean-squared error of 2.57, which can be more than halved using features of the training set alone. 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 it achieves these better predictions, but we can obtain some understanding by examining the best 1-split

decision tree, shown in Figure 5 below. 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 z_1 (the larger prize) in the training set of lotteries exceeds zero. There are three domains in which the prizes (z_1, z_2) 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 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. When 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.

Our results also clarify a 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: An environment where each domain puts weight on exactly one lottery that is itself sampled iid may be difficult for black-box algorithms,[34] while an environment where the marginal distribution is degenerate on the same lottery in all domains may be easier. 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.

---

[34]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 to a new domain. We now complement those results with procedures for inference focused on population quantities: Section 5.2 provides confidence intervals for quantiles of the transfer error distribution, and Section 5.3 provides a confidence interval for the expected transfer error. Since these quantities 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, unlike the forecast intervals from Section 3.

5.1. Preliminary Lemma.

We start by establishing a bound that will be useful in the subsequent construction of confidence intervals. Let

$$U = \frac{(n-k)!}{n!} \sum_{(i_1, \dots, i_k) \in \mathbb{T}_{r+1,n}} \phi(Z_{i_1}, \dots, Z_{i_k})$$

be an arbitrary U-statistic of degree k with a bounded (and potentially asymmetric) kernel ϕ that takes values in $[0, 1]$.

Definition 12. *For every $n, k \in \mathbb{Z}_+$ and $x, y \in \mathbb{R}$, define*

$$B_{n,k}(x; y) \equiv \min \left\{ b_{n,k}^1(x; y), b_{n,k}^2(x; y), b_{n,k}^3(x; y) \right\}$$

where

$$b_{n,k}^1(x; y) \equiv \exp \left\{ -\lceil n/k \rceil \left(x \wedge y \log \left(\frac{x \wedge y}{y} \right) + (1 - x \wedge y) \log \left(\frac{1 - x \wedge y}{1 - y} \right) \right) \right\}$$

$$b_{n,k}^2(x; y) \equiv e \cdot \mathbb{P} \left(\text{Binom}(\lceil n/k \rceil; y) \leq \lceil \lceil n/k \rceil \cdot x \rceil \right)$$

$$b_{n,k}^3(x; y) \equiv \min_{\lambda > 0} \frac{n\lambda}{k} \left(x - \frac{\lambda}{\lambda + kG(\lambda)} y \right)$$

Lemma 1. *If $\phi(Z_1, \dots, Z_k) \in [0, 1]$ almost surely, then $P(U \leq x) \leq B_{n,k}(x; \mathbb{E}(U))$ for every $x \in [0, 1]$.*

5.2. Quantiles of transfer error.

Let F denote the CDF of $e_{\mathbf{T}, n+1}$, which we assume is continuous. This section builds a confidence interval for the β-th quantile of F, denoted q_β.

For arbitrary $q \in \mathbb{R}$ and realized metadata $\mathbf{M} = \{S_1, \dots, S_n\}$, define

$$\varphi(q, \mathbf{M}) = \frac{(n-r-1)!}{n!} \sum_{(d_1, \dots, d_{r+1}) \in \mathbb{T}_{r+1,n}} \mathbb{I}(e_{(d_1, \dots, d_r), d_{r+1}} \leq q)$$

where $\mathbb{I}(\cdot)$ is the indicator function, recalling that $e_{(d_1, \dots, d_r), d_{r+1}}$ denotes the observed transfer error from samples $(S_{d_1}, \dots, S_{d_r})$ to sample $S_{d_{r+1}}$. This is the fraction of observed transfer errors in the metadata (from r training samples to one test sample) that are less than q. Then $U_\beta \equiv \varphi(q_\beta, \mathbf{M})$ is a U-statistic where by definition, $\mathbb{E}[U_\beta] = \beta$. Lemma 1 then implies

$$\mathbb{P}(U_\beta \leq x) \leq B_{n,r+1}(x, \beta) \qquad \mathbb{P}(U_\beta \geq x) = \mathbb{P}(1 - U_\beta \leq 1 - x) \leq B_{n,r+1}(1-x, 1-\beta). \quad (13)$$

Definition 13. *For any quantile $\beta \in (0,1)$ and confidence level $1 - \alpha \in (0,1)$, let $\hat{u}_\beta^+(\alpha) = \inf\{u : B_{n,r+1}(u; \beta) \geq \alpha\}$ and $\hat{u}_\beta^-(\alpha) = \sup\{u : B_{n,r+1}(1-u; 1-\beta) \geq \alpha\}$. Further define $\hat{q}_\beta^L(\alpha) \equiv \min\{q : \varphi(q, \mathbf{M}) \geq \hat{u}_\beta^+(\alpha)\}$ and $\hat{q}_\beta^U(\alpha) \equiv \max\{q : \varphi(q, \mathbf{M}) \leq \hat{u}_r^-(\alpha)\}.$*

Since $B_{n,r+1}(u; \cdot)$ is right-continuous in u, it follows from (13) that $\mathbb{P}(U_\beta < \hat{u}_\beta^+(\alpha)) \leq \alpha$ and $\mathbb{P}(U_\beta > \hat{u}_\beta^-(\alpha)) \leq \alpha$. Since $\varphi(q, \mathbf{M})$ is monotonically increasing in q, the event $\{U_\beta < \hat{u}_\beta^+(\alpha)\}$ is equivalent to $\{q_\beta < \hat{q}_\beta^L(\alpha)\}$, while $\{U_\beta > \hat{u}_\beta^-(\alpha)\}$ is equivalent to $\{q_\beta > \hat{q}_\beta^U(\alpha)\}$. Thus we obtain:

Proposition 2. *For any quantile $\beta \in (0,1)$ and confidence level $1 - \alpha \in (0,1)$,*

$$\mathbb{P}(q_\beta \leq \hat{q}_\beta^U(\alpha)) \geq 1 - \alpha$$

and

$$\mathbb{P}\left(q_\beta \in \left[\hat{q}_\beta^L(\alpha/2), \hat{q}_\beta^U(\alpha/2)\right]\right) \geq 1 - \alpha.$$

Figure 6 applies Proposition 2 to construct two-sided 81% confidence interval for the median transfer error, median normalized transfer error, and median transfer deterioration. As in Figure 3, these confidence intervals are substantially wider for the black box algorithms, and have higher upper bounds.

5.3. **Expected transfer error.** This section constructs confidence intervals for the expected transfer error, $\mu \equiv \mathbb{E}(e_{\mathbf{T}, n+1})$, under the assumption that transfer errors are uniformly bounded (in which case it is without loss to set $e_{\mathbf{T}, n+1} \in [0, 1]$). Define the U-statistic
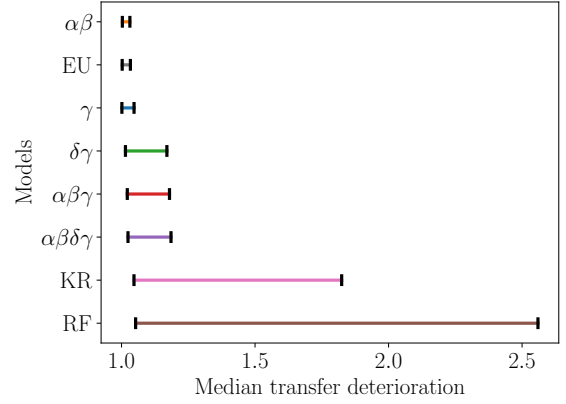
$$U = \frac{(n-r-1)!}{n!} \sum_{(d_1, \dots, d_{r+1}) \in \mathbb{T}_{r+1,n}} e_{(d_1, \dots, d_r), d_{r+1}}.$$

31
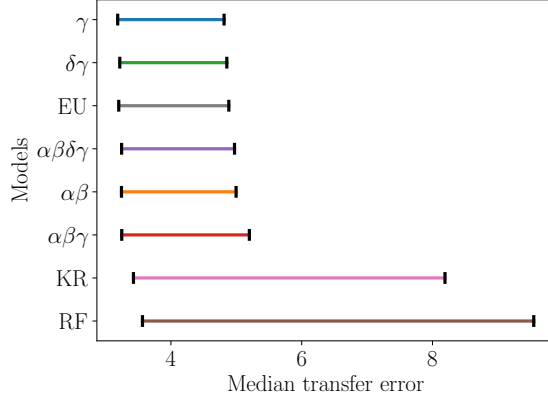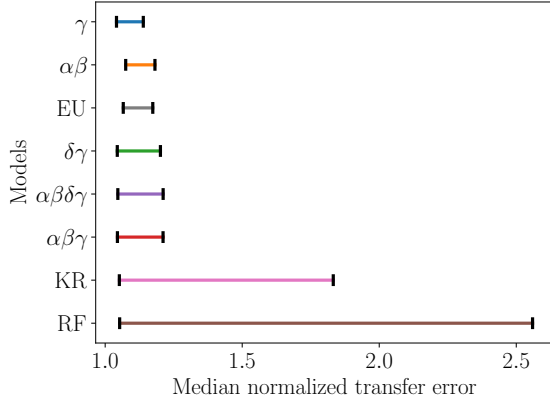
FIGURE 6. *81% confidence intervals for the median of (a) transfer error, (b) normalized transfer error, and (c) transfer deterioration.*

Because $\mathbb{E}[U] = \mu$, Lemma 1 implies that $\mathbb{P}(U \leq x) \leq B_{n,r+1}(x,\mu)$ and $\mathbb{P}(U \geq x) \leq B_{n,r+1}(1-x, 1-\mu)$ for all $x \in \mathbb{R}$.

Definition 14. *For any confidence guarantee* $1-\alpha \in (0,1)$*, let* $\hat{\mu}^+(\alpha) = \sup\{\mu : B_{n,r+1}(U;\mu) \geq \alpha\}$ *and* $\hat{\mu}^-(\alpha) = \inf\{\mu : B_{n,r+1}(1-U;1-\mu) \geq \alpha\}$.

It follows from (13) that $\mathbb{P}(U < \hat{u}^+(\alpha)) \leq \alpha$ and $\mathbb{P}(U > \hat{u}^-(\alpha)) \leq \alpha$, which implies:

Proposition 3. *If* $e_{T,d} \in [0,1]$ *almost surely, then*

$$\mathbb{P}\left(\mu \leq \hat{\mu}^+(\alpha)\right) \geq 1 - \alpha$$

and

$$\mathbb{P}\left(\mu \in [\hat{\mu}^-(\alpha/2), \hat{\mu}^+(\alpha/2)]\right) \geq 1 - \alpha.$$

Figure 7 applies this result to construct a two-sided 81% confidence intervals for our transferability measures. Since our main measures are not bounded, we report instead confidence intervals for the expectation of the inverse of normalized transfer error, i.e.,

$$\frac{\min_{\hat{\rho}\in\mathcal{R}} e(\hat{\rho}(S_{n+1}), S_{n+1})}{e(\rho(\mathbf{M_T}), S_{n+1})} \in [0, 1]$$

and the inverse of transfer deterioration, i.e.,

$$\frac{e(\rho(S_{n+1}), S_{n+1})}{e(\rho(\mathbf{M_T}), S_{n+1})} \in [0, 1].$$

Lower values for these measures correspond to worse transfer performance. We find again that the confidence intervals for the black box algorithms are qualitatively worse than those for the economic models.
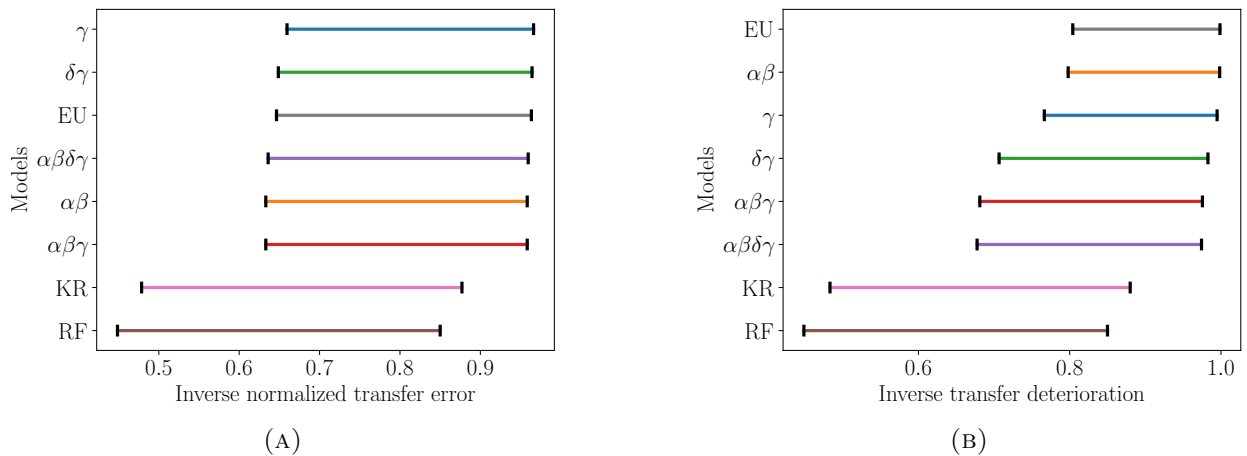

FIGURE 7. *81% forecast intervals for (a) expected inverse normalized transfer error, (b) expected inverse transfer deterioration.*

6. CONCLUSION

Our measures of transfer error quantify how well a model's performance on one domain extrapolates to other domains. 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 supports the intuition that economic models can recover regularities that are general across a variety of domains.

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.

AGRAWAL, M., J. C. PETERSON, AND T. L. GRIFFITHS (2020): "Scaling up psychology via Scientific Regret Minimization," *Proceedings of the National Academy of Sciences*, 117, 8825–8835.

AL-UBAYDLI, O. AND J. A. LIST (2015): "420On the Generalizability of Experimental Results in Economics," in *Handbook of Experimental Economic Methodology*, Oxford University Press.

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.

ANDREWS, I. AND E. OSTER (2019): "A simple approximation for evaluating external validity bias," *Economics Letters*, 178, 58–62.

ANGELOPOULOS, A. N., S. BATES, A. FISCH, L. LEI, AND T. SCHUSTER (2022): "Conformal risk control," *arXiv preprint arXiv:2208.02814*.

ANGRIST, J. D. AND I. FERNÁNDEZ-VAL (2013): *ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework*, Cambridge University Press, vol. 3 of *Econometric Society Monographs*, 401–434.

ARONOW, P. M. AND D. K. LEE (2013): "Interval estimation of population means under unknown but bounded probabilities of sample selection," *Biometrika*, 100, 235–240.

ATHEY, S. (2017): "Beyond prediction: Using big data for policy problems," *Science*, 355, 483–485.

ATHEY, S. AND G. IMBENS (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

BANDIERA, O., G. FISCHER, A. PRAT, AND E. YTSMA (2021): "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments," *American Economic Review: Insights*, 3, 435–54.

BARBER, R. F., E. J. CANDES, A. RAMDAS, AND R. J. TIBSHIRANI (2021): "Predictive inference with the jackknife+," *The Annals of Statistics*, 49, 486–507.

BATES, S., A. ANGELOPOULOS, L. LEI, J. MALIK, AND M. JORDAN (2021): "Distribution-free, risk-controlling prediction sets," *Journal of the ACM (JACM)*, 68, 1–34.

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.

BENTKUS, V. (2004): "On Hoeffding's inequalities," *The Annals of Probability*, 32, 1650–1673.

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.

BERTANHA, M. AND G. W. IMBENS (2020): "External Validity in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 38, 593–612.

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.

CARD, D. AND A. B. KRUEGER (1995): "Time-Series Minimum-Wage Studies: A Meta-analysis," *The American Economic Review*, 85, 238–243.

CHARNES, A. AND W. W. COOPER (1962): "Programming with linear fractional functionals," *Naval Research logistics quarterly*, 9, 181–186.

CHASSANG, S. AND S. KAPON (2022): "Designing Randomized Controlled Trials with External Validity in Mind," .

COVENEY, P. V., E. R. DOUGHERTY, AND R. R. HIGHFIELD (2016): "Big data need big theory too," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374, 20160153.

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

DEATON, A. (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48, 424–55.

DEHEJIA, R., C. POP-ELECHES, AND C. SAMII (2021): "From Local to Global: External Validity in a Fertility Natural Experiment," *Journal of Business & Economic Statistics*, 39, 217–243.

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.

FORTIN, N. M., A. HILL, AND J. HUANG (2014): "SUPERSTITION IN THE HOUSING MARKET," *Economic Inquiry*, 52, 974–993.

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.

HAAVELMO, T. (1944): "The Probability Approach in Econometrics," *Econometrica*, 12, iii–115.

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

HARTFORD, J. S., J. R. WRIGHT, AND K. LEYTON-BROWN (2016): "Deep Learning for Predicting Human Strategic Behavior," in *Advances in Neural Information Processing Systems*, ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Curran Associates, Inc., vol. 29.

HOEFFDING, W. (1963): "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30.

HOFMAN, J. M., D. J. WATTS, S. ATHEY, F. GARIP, T. L. GRIFFITHS, J. KLEINBERG, H. MARGETTS, S. MULLAINATHAN, M. J. SALGANIK, S. VAZIRE, A. VESPIGNANI, AND T. YARKONI (2021): "Integrating explanation and prediction in computational social science," *Nature*, 595, 181–188.

HOTZ, V. J., G. W. IMBENS, AND J. H. MORTIMER (2005): "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics*, 125, 241–270.

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.

IMAI, T., T. A. RUTTER, AND C. F. CAMERER (2020): "Meta-Analysis of Present-Bias Estimation using Convex Time Budgets," *The Economic Journal*, 131, 1788–1814.

IMBENS, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423.

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.

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.

LEVITT, S. D. AND J. A. LIST (2007): "Viewpoint: On the Generalizability of Lab Behaviour to the Field (A propos de lapossibilité de généraliser les comportements de laboratoire à ce qui se passe sur le terrain)," *The Canadian Journal of Economics / Revue canadienne d'Economique*, 40, 347–370.

L'HARIDON, O. AND F. M. VIEIDER (2019): "All over the map: A worldwide comparison of risk preferences," *Quantitative Economics*, 10, 185–215.

MANSKI, C. F. (2021): "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89, 2827–2853.

MAURER, A. (2006): "Concentration inequalities for functions of independent variables," *Random Structures & Algorithms*, 29, 121–138.

MCFADDEN, D. (1974): "The measurement of urban travel demand," *Journal of Public Economics*, 3, 303–328.

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.

——— (2022): "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature," *American Economic Review*, 112, 1818–47.

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.

NIE, X., G. IMBENS, AND S. WAGER (2021): "Covariate Balancing Sensitivity Analysis for Extrapolating Randomized Trials across Locations," .

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," .

PARK, S., A. TAFTI, AND G. SHMUELI (2023): "Transporting Causal Effects Across Populations Using Structural Causal Modeling: The Example of Work-From-Home Productivity," *Information Systems Research*.

PATHAK, P. AND P. SHI (2013): "Simulating Alternative School Choice Options in Boston - Main Report," Working Paper.

PEARL, J. AND E. BAREINBOIM (2011): "Transportability across studies: A formal approach," Tech. rep., CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.

——— (2014): "External Validity: From Do-Calculus to Transportability Across Populations," *Statistical Science*, 29, 579 – 595.

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.

PETERSON, J. C., D. D. BOURGIN, M. AGRAWAL, D. REICHMAN, AND T. L. GRIFFITHS (2021): "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, 372, 1209–1214.

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.

SAHOO, R., L. LEI, AND S. WAGER (2022): "Learning from a biased sample," *arXiv preprint arXiv:2209.01754*.

SHEN, Z., J. LIU, Y. HE, X. ZHANG, R. XU, H. YU, AND P. CUI (2021): "Towards Out-Of-Distribution Generalization: A Survey," *CoRR*, abs/2108.13624.

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.

TIBSHIRANI, R. J., R. FOYGEL BARBER, E. CANDES, AND A. RAMDAS (2019): "Conformal prediction under covariate shift," *Advances in neural information processing systems*, 32.

TIPTON, E. AND R. B. OLSEN (2018): "A review of statistical methods for generalizing from evaluations of educational interventions," *Educational Researcher*, 47, 516–524.

VIVALT, E. (2020): "How Much Can We Generalize From Impact Evaluations?" *Journal of the European Economic Association*, 18, 3045–3089.

VOVK, V., A. GAMMERMAN, AND G. SHAFER (2005): *Algorithmic learning in a random world*, vol. 29, Springer.

ZHOU, K., Z. LIU, Y. QIAO, T. XIANG, AND C. C. LOY (2021): "Domain Generalization: A Survey," Working Paper.

APPENDIX A. PROOFS

A.1. **Notation.** Throughout let $\mathcal{N} \equiv \{1, \ldots, n\}$. The set $\mathbb{T}_{r,n}$ consists of all vectors of length r with distinct values in $\mathcal{N}$. For any $(d_1, \ldots, d_{r+1}) \in \mathbb{T}_{r+1,n+1}$, let $f(d_1, \ldots, d_{r+1}) = e_{(d_1,\ldots,d_r),d_{r+1}}$ denote the transfer error from training samples $S_{d_1}, \ldots, S_{d_r}$ to test sample $S_{d_{r+1}}$.

A.2. **Proofs of Proposition 1 and Claim 1.** Since $\mathbf{T}$ is a random subset of $\{1, \ldots, n\}$ that is independent of $\{S_1, \ldots, S_{n+1}\}$,

$$\mathbb{P}(e_{\mathbf{T},n+1} \in A) = \mathbb{P}(e_{(1,\ldots,r),n+1} \in A)$$

for any event A that is independent of $\mathbf{T}$. Thus, it suffices to prove Proposition 1 by replacing $\mathbf{T}$ with $(1, \ldots, r)$. We start by proving the one-sided guarantee:

$$\mathbb{P}\left(e_{(1,\ldots,r),n+1} \leq \bar{e}_\tau^{\mathbf{M}}\right) \geq \tau \left(1 - \frac{r+1}{n+1}\right).$$

Throughout the proof we condition on the unordered samples $\{S_1, \ldots, S_{n+1}\}$ and denote by $\{S_{(1)}, \ldots, S_{(n+1)}\}$ any typical realization. Ranging over all possible choices of r (ordered) training environments and a single test environment from $\{1, \ldots, n+1\}$ yields the multiset

of transfer errors[35]

$$\mathcal{C} = \left\{ f(d_1, \ldots, d_{r+1}) : (d_1, \ldots, d_{r+1}) \in \mathbb{T}_{r+1,n+1} \right\}, \tag{A.1}$$

which has size $(n+1)!/(n-r)!$. The subset of these transfer errors that don't use sample k for either training or testing is

$$\mathcal{C}_{-k} = \left(f(d_1, \ldots, d_{r+1}) : (d_1, \ldots, d_{r+1}) \in \mathbb{T}_{r+1,n+1}, d_j \neq k, j = 1, \ldots, r+1 \right),$$

which has size $n!/(n-r-1)!$. For any $\tau \in [0,1]$, let $\bar{E}^*_\tau$ denote the $\lceil \tau n!/(n-r-1)! \rceil$-th smallest element in $\mathcal{C}$, and let $\bar{E}_{k,\tau}$ denote the $\lceil \tau n!/(n-r-1)! \rceil$-th smallest element in $\mathcal{C}_{-k}$. Since $\mathcal{C}_{-k} \subseteq \mathcal{C}$,

$$\bar{E}_{k,\tau} \geq \bar{E}^*_\tau \text{ for all } k \text{ and } \tau. \tag{A.2}$$

Let $\mathcal{F}$ denote the sigma-field generated by the unordered set $\{S_{(1)}, \ldots, S_{(n+1)}\}$. Exchangeability implies that

$$\left(e_{(1,\ldots,r),n+1}, \bar{e}_\tau \right) \mid \mathcal{F} \stackrel{d}{=} \left(f(\boldsymbol{\pi}^*(1), \ldots, \boldsymbol{\pi}^*(r), \boldsymbol{\pi}^*(n+1)), \bar{E}_{\boldsymbol{\pi}^*(n+1),\tau} \right)$$

where $\boldsymbol{\pi}^*$ denotes a random permutation drawn from the uniform distribution over all permutations on $\{1, \ldots, n+1\}$. Thus

$$\mathbb{P}\left(e_{(1,\ldots,r),n+1} \leq \bar{e}_\tau \mid \mathcal{F} \right) = \mathbb{P}_{\boldsymbol{\pi}^*}\left(f(\boldsymbol{\pi}^*(1), \ldots, \boldsymbol{\pi}^*(r), \boldsymbol{\pi}^*(n+1)) \leq \bar{E}_{\boldsymbol{\pi}^*(n+1),\tau} \right). \tag{A.3}$$

It follows that

$$\mathbb{P}\left(e_{(1,\ldots,r),n+1} \leq \bar{e}_\tau \mid \mathcal{F} \right) \geq \mathbb{P}_{\boldsymbol{\pi}^*}\left(f(\boldsymbol{\pi}^*(1), \ldots, \boldsymbol{\pi}^*(r), \boldsymbol{\pi}^*(n+1)) \leq \bar{E}^*_\tau \right) \qquad \text{by (A.2) and (A.3)}$$

$$= \frac{(n-r)!}{(n+1)!} \sum_{(d_1,\ldots,d_{r+1}) \in \mathbb{T}_{r+1,n+1}} \mathbb{I}(f(d_1,\ldots,d_{r+1}) \leq \bar{E}^*_\tau)$$

$$= \frac{(n-r)!}{(n+1)!} \sum_{f \in \mathcal{C}} \mathbb{I}(f \leq \bar{E}^*_\tau)$$

$$\geq \frac{(n-r)!}{(n+1)!} \left\lceil \tau \frac{n!}{(n-r-1)!} \right\rceil \geq \tau \left(1 - \frac{r+1}{n+1} \right).$$

The one-sided guarantee then follows from the law of iterated expectation.

---

[35]A multiset is a set-like, unordered collection where repeated values are multiply counted.

Turning to the two-sided guarantee, we first notice that the one-sided guarantee implies

$$\mathbb{P}\left(e_{(1,\ldots,r),n+1} > \bar{e}_\tau^{\mathbf{M}}\right) \leq 1 - \tau + \tau \frac{r+1}{n+1},$$

and, by symmetry,

$$\mathbb{P}\left(e_{(1,\ldots,r),n+1} < \underline{e}_\tau^{\mathbf{M}}\right) \leq 1 - \tau + \tau \frac{r+1}{n+1}.$$

Thus,

$$\mathbb{P}\left(e_{(1,\ldots,r),n+1} \in \left[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}\right]\right) \geq 2\tau - 1 - 2\tau \frac{r+1}{n+1}.$$

This completes the proof of Proposition 1.

To prove Claim 1, define $\bar{E}_\tau'$ to be the $\lfloor (1-\tau)n!/(n-r-1)! + 1 \rfloor$-th largest element in $\mathcal{C}$ as defined in (A.1). Since $|\mathcal{C}_{-k}| = n!/(n-r-1)!$, the quantity $\bar{E}_{k,\tau}$ is also the $\lfloor (1-\tau)n!/(n-r-1)! + 1 \rfloor$-th largest element in $\mathcal{C}_{-k}$. So $\bar{E}_{k,\tau} \leq \bar{E}_\tau'$ for any k and τ.

By (A.3), $\mathbb{P}\left(e_{(1,\ldots,r),n+1} \leq \bar{e}_\tau \mid \mathcal{F}\right) \leq \mathbb{P}_{\boldsymbol{\pi}^*}\left(f(\boldsymbol{\pi}^*(1),\ldots,\boldsymbol{\pi}^*(r), \boldsymbol{\pi}^*(n+1)) \leq \bar{E}_\tau'\right)$

$$= \frac{(n-r)!}{(n+1)!} \sum_{(d_1,\ldots,d_{r+1}) \in \mathbb{T}_{r+1,n+1}} \mathbb{I}(f(d_1,\ldots,d_{r+1}) \leq \bar{E}_\tau')$$

$$= \frac{(n-r)!}{(n+1)!} \sum_{f \in \mathcal{C}} \mathbb{I}(f \leq \bar{E}_\tau').$$

Claim 1 assumes that $\mathcal{C}$ has no ties almost surely, so $\sum_{f \in \mathcal{C}} \mathbb{I}(f > \bar{E}_\tau') = \left\lfloor (1-\tau)\frac{n!}{(n-r-1)!} \right\rfloor$. Thus

$$\mathbb{P}\left(e_{(1,\ldots,r),n+1} \leq \bar{e}_\tau \mid \mathcal{F}\right) \leq \frac{(n-r)!}{(n+1)!} \left(|\mathcal{C}| - \left\lfloor (1-\tau)\frac{n!}{(n-r-1)!} \right\rfloor\right)$$

$$\leq \frac{(n-r)!}{(n+1)!} \left(\frac{(n+1)!}{(n-r)!} - (1-\tau)\frac{n!}{(n-r-1)!} + 1\right)$$

$$= \tau + (1-\tau)\frac{r+1}{n+1} + \frac{(n-r)!}{(n+1)!}.$$

The two-sided guarantee follows by symmetry.

A.3. **Proof of Theorem 1.** Again let $\mathcal{F}$ denote the sigma-field generated by the unordered set $\{S_1, \ldots, S_{(n+1)}\}$. Under the assumed data-generating process,

$$e_{(d_1,\ldots,d_r),n+1} \mid \mathcal{F} \overset{d}{=} f(\boldsymbol{\pi}^w(d_1), \ldots, \boldsymbol{\pi}^w(d_r), \boldsymbol{\pi}^w(n+1)), \quad \forall (d_1, \ldots, d_r) \in \mathbb{T}_{r,n}.$$

where $\boldsymbol{\pi}^w$ is a random permutation on $\{1, \ldots, n+1\}$ distributed according to

$$\mathbb{P}(\boldsymbol{\pi}^w = \pi) = \frac{1}{n!} \frac{w(S_{\pi(n+1)})}{\sum_{j=1}^{n+1} \omega(S_j)}. \tag{A.4}$$

On the other hand, $\mathbf{T} \overset{d}{=} (\boldsymbol{\pi}^n(1), \ldots, \boldsymbol{\pi}^n(r))$ where $\boldsymbol{\pi}^n$ denotes a uniform random permutation on $\{1, \ldots, n\}$, so $e_{\mathbf{T},n+1} \mid \mathcal{F} \overset{d}{=} f(\boldsymbol{\pi}^w \circ \boldsymbol{\pi}^n(1), \ldots, \boldsymbol{\pi}^w \circ \boldsymbol{\pi}^n(r), \boldsymbol{\pi}^w(n+1))$. By (A.4), we have

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \overset{d}{=} f(\boldsymbol{\pi}^w(1), \ldots, \boldsymbol{\pi}^w(r), \boldsymbol{\pi}^w(n+1)). \tag{A.5}$$

For any $(d_1, \ldots, d_r, k) \in \mathbb{T}_{r+1,n+1}$,

$$\mathbb{P}((\boldsymbol{\pi}^w(1), \ldots, \boldsymbol{\pi}^w(r), \boldsymbol{\pi}^w(n+1)) = (d_1, \ldots, d_r, k)) = \frac{(n-r)!}{n!} \frac{w(S_k)}{\sum_{j=1}^{n+1} \omega(S_j)} \triangleq W_k'.$$

Thus,

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \sim \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} W_k' \cdot \delta_{f(d_1,\ldots,d_r,k)}.$$

Note that for any random variable $Z \sim F$, $\mathbb{P}(Z \leq \bar{Q}_\tau(F)) \geq \tau$. Then

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} W_k' \cdot \delta_{f(d_1,\ldots,d_r,k)}\right) \mid \mathcal{F}\right) \geq \tau. \tag{A.6}$$

The value $f(d_1, \ldots, d_r, k)$ can be observed iff $d_1, \ldots, d_r, k$ are all not equal to $n+1$. The above unobserved upper confidence bound can be replaced by setting all unobserved f-values to ∞, i.e.,

$$\bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} W_k' \cdot \delta_{f(d_1,\ldots,d_r,k)} + \Omega_{n+1} \cdot \delta_\infty\right).$$

where

$$\Omega_{n+1} = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}\backslash\mathbb{T}_{r+1,n}} W_k'.$$

Via a simple counting argument, we obtain that

$$\begin{aligned}
\Omega_{n+1} &= \frac{n!}{(n-r)!} W_{n+1}' + \frac{r(n-1)!}{(n-r)!} \sum_{k=1}^{n} W_k' \\
&= \frac{r}{n} + \frac{(n-1)!}{(n-r-1)!} W_{n+1}' \\
&= \frac{r}{n} + \frac{n-r}{n} \frac{w(S_{n+1})}{\sum_{j=1}^{n+1} \omega(S_j)}. \tag{A.7}
\end{aligned}$$

42

where the second to last equality uses $\sum_{k=1}^{n+1} W_k' = \frac{(n-r)!}{n!}$. By (A.7) and (9), we have $W_k = \frac{W_k'}{1-\Omega_{n+1}}$, so

$$\bar{Q}_\tau \left(\sum_{(d_1,\dots,d_r,k)\in\mathbb{T}_{r+1,n}} W_k' \cdot \delta_{f(d_1,\dots,d_r,k)} + \Omega_{n+1} \cdot \delta_\infty \right)$$
$$= \bar{Q}_{\frac{\tau}{1-\Omega_{n+1}}} \left(\sum_{(d_1,\dots,d_r,k)\in\mathbb{T}_{r+1,n}} W_k \cdot \delta_{f(d_1,\dots,d_r,k)} \right),$$

where the RHS is ∞ if $\tau \le \Omega_{n+1}$.

Replacing τ by $\tau(1-\Omega_{n+1})$, (A.6) implies that

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \le \bar{Q}_\tau \left(\sum_{(d_1,\dots,d_r,k)\in\mathbb{T}_{r+1,n}} W_k \cdot \delta_{f(d_1,\dots,d_r,k)} \right) \mid \mathcal{F} \right) \ge \tau(1-\Omega_{n+1}).$$

The theorem now follows from the law of iterated expectation.

A.4. **Proof of Lemma 1.** Hoeffding (1963) shows that $P(U \le x) \le b_{n,k}^1(x, \mathbb{E}(U))$, and Bates et al. (2021) shows that $P(U \le x) \le b_{n,k}^2(x, \mathbb{E}(U))$. To show that if $x \in [0,1]$ then $P(U \le x) \le b_{n,k}^3(x, \mathbb{E}(U))$ we use a series of intermediate result to extend a result of Bates et al. (2021) on U-statistics of degree 2 with bounded kernels to U-statistics with bounded kernels for any order $k \ge 2$.

Let $Z_1, \dots, Z_n$ be iid random variables and $\phi : \mathbb{R}^k \to [0,1]$ be a bounded function. Then a U-statistic of degree k is defined as

$$U = \frac{(n-k)!}{n!} \sum_{i_1,\dots,i_k} \phi(Z_{i_1}, \dots, Z_{i_k}), \tag{A.8}$$

where $\sum_{i_1,\dots,i_k}$ denotes the sum over all k-tuples in $\mathcal{N}$ with mutually distinct elements. The average of Z_i is a special case of (A.8) with $k = 1$ and $\phi(z) = z$.

Let $m = \lfloor n/k \rfloor$ and $\boldsymbol{\pi}^n : \mathcal{N} \mapsto \mathcal{N}$ be a uniformly random permutation. For each permutation π, define

$$W_\pi = \frac{1}{m} \sum_{j=1}^m \phi\left(Z_{\pi((j-1)k+1)}, \dots, Z_{\pi(jk)} \right).$$

Note that the summands in W_π are independent given π. Then $U = \mathbb{E}_{\boldsymbol{\pi}^n}[W_{\boldsymbol{\pi}^n}]$, where $\mathbb{E}_{\boldsymbol{\pi}^n}$ denotes the expectation with respect to $\boldsymbol{\pi}^n$ when conditioning on $Z_1, \dots, Z_n$. By Jensen's inequality, for any convex function ψ, $\mathbb{E}[\psi(U)] = \mathbb{E}[\psi(\mathbb{E}_{\boldsymbol{\pi}^n}[W_{\boldsymbol{\pi}^n}])] \le \mathbb{E}[\mathbb{E}_{\boldsymbol{\pi}^n}\psi(W_{\boldsymbol{\pi}^n})] =$

43

$\mathbb{E}_{\boldsymbol{\pi}^n}[\mathbb{E}\psi(W_{\boldsymbol{\pi}^n})]$. Since W_π has identical distributions for all π,

$$\mathbb{E}[\psi(U)] \leq \mathbb{E}[\psi(W_{\mathbf{id}})] \tag{A.9}$$

where **id** is the permutation that maps each element to itself.

Recalling that Hoeffding's inequality is derived from the moment-generating function $\psi(z) = e^{\lambda z}$ (Hoeffding, 1963), and the Bentkus inequality is derived from the piecewise linear function $\psi(z) = (z - t)_+$ (Bentkus, 2004), the following tail inequalities for U-statistics are a direct consequence of (A.9).

Proposition A.1. *Let U be a U-statistic of order k with a bounded kernel $\phi \in [0, 1]$ in the form of* (A.8) *and $m = \lfloor n/k \rfloor$. Then*

(1) (Hoeffding inequality for U-statistics, Section 5 of Hoeffding 1963)

$$\mathbb{P}(U \leq x) \leq \exp\left\{-mh_1\left(x \wedge \mathbb{E}[U]; \mathbb{E}[U]\right)\right\},$$

where

$$h_1(y; \mu) = y \log\left(\frac{y}{\mu}\right) + (1 - y) \log\left(\frac{1 - y}{1 - \mu}\right).$$

(2) (Bentkus inequality for U-statistics, modified from Bentkus 2004)

$$\mathbb{P}\left(U \leq x\right) \leq e\mathbb{P}\left(\mathrm{Bin}\left(m; \mathbb{E}[U]\right) \leq \lceil mx \rceil\right).$$

Other concentration inequalities can be derived from the leave-one-out property. Write $U(Z_1, \ldots, Z_n)$ for U and let $U_i = \inf_{z_i} U(Z_1, \ldots, Z_{i-1}, z_i, Z_{i+1}, \ldots, Z_n)$. Note that U_i is independent of Z_i. Since $\phi(\cdot) \geq 0$, we have $0 \leq U - U_i \leq \frac{(n-k)!}{n!} \sum_{j=1}^k \sum_{i_1, \ldots, i_k, i_j = i} \phi(Z_{i_1}, \ldots, Z_{i_k})$ so $\frac{n}{k}(U - U_i) \leq 1$ and

$$\sum_{i=1}^n (U - U_i)^2 \leq \frac{((n-k)!)^2}{(n!)^2} \sum_{i=1}^n \left(\sum_{j=1}^k \sum_{i_1, \ldots, i_k, i_j = i} \phi(Z_{i_1}, \ldots, Z_{i_k})\right)^2$$

$$\overset{(i)}{\leq} \frac{k(n-k)!}{n \cdot n!} \sum_{j=1}^k \sum_{i=1}^n \sum_{i_1, \ldots, i_k, i_j = i} \phi(Z_{i_1}, \ldots, Z_{i_k})^2$$

$$\overset{(ii)}{\leq} \frac{k(n-k)!}{n \cdot n!} \sum_{j=1}^k \sum_{i=1}^n \sum_{i_1, \ldots, i_k, i_j = i} \phi(Z_{i_1}, \ldots, Z_{i_k})$$

$$= \frac{k^2}{n} U,$$

where (i) applies the Cauchy-Schwarz inequality and (ii) uses the fact that $\phi(\cdot) \leq 1$. If we let $W = (n/k)U$ and $W_i = (n/k)U_i$, then $W - W_i \leq 1$, $\quad \sum_{i=1}^{n}(W - W_i)^2 \leq kW$. This implies that W as a function of $Z_1, \ldots, Z_n$ satisfies the assumptions for the claim (34) in Theorem 13 of Maurer (2006) with constant $a = k$.[36]

Proposition A.2 (Theorem 13, Maurer 2006). *Let* $G(\lambda) = (e^\lambda - \lambda - 1)/\lambda$. *Then for any* $\lambda > 0$,

$$\log \mathbb{E}[e^{\lambda(\mathbb{E}[W]-W)}] \leq \frac{k\lambda G(\lambda)}{\lambda + kG(\lambda)}\mathbb{E}[W].$$

This further implies that for any $x \in (0, \mathbb{E}[U])$,

$$\mathbb{P}(U \leq x) \leq \exp\left\{\min_{\lambda>0} \frac{n\lambda}{k}\left(x - \frac{\lambda}{\lambda + kG(\lambda)}\mathbb{E}[U]\right)\right\}.$$

Putting Proposition A.1 and Proposition A.2 together yields Lemma 1.

---

[36]Theorem 13 of Maurer (2006) states a weaker result that $\log \mathbb{E}[e^{\lambda(\mathbb{E}[W]-W)}] \leq \frac{k\mathbb{E}[W]}{2}\lambda^2$. The stronger version stated here can be found in the second last display in the proof of Theorem 13 of Maurer (2006).

Online appendix to the paper

The Transfer Performance of Economic Models

Isaiah Andrews   Drew Fudenberg   Lihua Lei   Annie Liang   Chaofeng Wu

May 20, 2023

Appendix P. Supplementary Material to Section 2

P.1. Other choices for $e_{\mathbf{T},d}$.

Example 1 (Ratios of Model Errors). For any two decision rules ρ and ρ', set

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

to be the ratio of the transfer errors of the two decision rules.

Example 2 (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* $\mathbf{M}_{\mathcal{T}}$ *to* S is $e(\theta_{\mathbf{M}_{\mathcal{T}}}, S)$, where

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

is a best-fitting value of θ on the training data $\mathbf{M}_{\mathcal{T}}$.[37] Set $e_{T,d} \equiv e(\theta_{\mathbf{M}_{\mathcal{T}}}, S_d)$ to be the partial transfer error on sample S_d.

P.2. Other decision rules.

Example 3 (Domain Cross-Validation). Let $\mathcal{R}^*$ be a family of decision rules, for example random forest algorithms with decision trees restricted to different depths. Let

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

---

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

and set $\rho(\mathbf{M}_\mathcal{T}) = \rho^*(\mathbf{M}_\mathcal{T})$. 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 4 (Distributionally Robust Optimization (DRO)). DRO chooses model parameters to minimize worst-case error with respect to a neighborhood of distributions around the training data. 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 $\mathbf{M}_\mathcal{T}$, let $P_{\mathbf{M}_\mathcal{T}}$ denote the empirical distribution of the pooled set of observations across the samples in the meta-data. As in Example 2.1, consider a set of prediction rules Σ^*. We define ρ_{Σ^*} to satisfy

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

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

Appendix Q. Supplementary Material to Section 3.2

Q.1. **Algorithm for evaluating worst-case-upper-dominance.** We provide an algorithm that computes $\bar{e}_\tau(\Gamma)$ with a single τ in $O(rn^{r+1} \log n)$ time and computes $\bar{e}_\tau(\Gamma)$ for all $\tau \in (0, 1)$ in $O(rn^{r+1} \log n + n^{r+2})$ time. Recall the definition of $\mathcal{C}$ in (A.1). First, sort the

elements in $\mathcal{C}$ as

$$f_{(1)} \leq f_{(2)} \leq \ldots \leq f_{(|\mathbb{T}_{r+1,n}|)},$$

where

$$f_{(j)} = f(d^{(j)}), \quad d^{(j)} = (d_1^{(j)}, \ldots, d_{r+1}^{(j)}) \in \mathbb{T}_{r+1,n}.$$

Let $\psi^{(j)} \in \{0,1\}^n$ with

$$\psi_i^{(j)} = I\left(d_{r+1}^{(j)} = i\right).$$

Further define the cumulative sum of $\psi^{(j)}$ as

$$\Psi^{(j)} = \sum_{\ell=1}^{j} \psi^{(\ell)}.$$

Let $w = (\omega(S_1), \ldots, \omega(S_n))^T$ and $\mathbf{1}_n = (1, 1, \ldots, 1)^T$. By (9), for each j,

$$f_{(j)} \geq \bar{e}_\tau^{\mathbf{M},\omega} \iff \frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} \geq \tau.$$

Therefore,

$$\bar{e}_\tau^{\mathbf{M},\omega} = f_{J^\omega}, \quad \text{where } J_\tau^\omega = \min\left\{j : \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} \geq \tau \frac{(n-1)!}{(n-r-1)!}\right\}.$$

By definition, the set of w generated by all $\omega \in \mathcal{W}(\Gamma)$ is $[\Gamma^{-1}, \Gamma]^n$. Thus,

$$\bar{e}_\tau^{\mathbf{M}}(\Gamma) = f_{J_\tau(\Gamma)}, \quad \text{where } J_\tau(\Gamma) = \min\left\{j : \min_{w \in [\Gamma^{-1}, \Gamma]^n} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} \geq \tau \frac{(n-1)!}{(n-r-1)!}\right\}. \quad \text{(Q.1)}$$

Via some algebra, we can further simplify the expression of $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$.

Theorem Q.1. *Let $\bar{\Psi}_k^{(j)}$ be the average of the k-smallest coordinates of $\Psi^{(j)}$ and*

$$Q_j(\Gamma) = \frac{j}{n} + \min_{k \in \mathcal{N}} \frac{\bar{\Psi}_k^{(j)} - \frac{j}{n}}{1 + \frac{n}{k(\Gamma^2 - 1)}}.$$

Then $Q_j(\Gamma)$ is strictly increasing in both j and Γ. Moreover, $\bar{e}_\tau^{\mathbf{M}}(\Gamma) = f_{(J_\tau(\Gamma))}$, where

$$J_\tau(\Gamma) = \min\left\{j \geq \tau \frac{n!}{(n-r-1)!} : Q_j(\Gamma) \geq \tau \frac{(n-1)!}{(n-r-1)!}\right\}.$$

Proof. First, we prove that

$$\min_{w \in [\Gamma^{-1}, \Gamma]^n} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} = \min_{w \in \{\Gamma^{-1}, \Gamma\}^n} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n}. \quad \text{(Q.2)}$$

3

Let $g_j(w) = w^T \Psi^{(j)}/w^T \mathbf{1}_n$. Then g_j is continuous and bounded on the closed set $[\Gamma^{-1}, \Gamma]^n$ and thus the minimum can be achieved. Let

$$w^{(j)}(\Gamma) = \underset{w:g_j(w)=\min_{w\in[\Gamma^{-1},\Gamma]^n} g_j(w)}{\operatorname{argmin}} \sum_{i=1}^n \min\left\{|w_i - \Gamma|, |w_i - \Gamma^{-1}|\right\}.$$

Suppose there exists $i \in \mathcal{N}$ such that $w_i^{(j)}(\Gamma) \in (\Gamma^{-1}, \Gamma)$. Then

$$g_j(w_i, w_{-i}) = \frac{\Psi_i^{(j)} w_i + \Psi_{-i}^{(j)T} w_{-i}}{w_i + \mathbf{1}_{n-1}^T w_{-i}} = \Psi_i^{(j)} + \frac{\Psi_{-i}^{(j)T} w_{-i} - \Psi_i^{(j)} \cdot \mathbf{1}_{n-1}^T w_{-i}}{w_i + \mathbf{1}_{n-1}^T w_{-i}},$$

where $\Psi_{-i}^{(j)}$ and w_{-i} are the leave-i-th-entry subvectors of $\Psi^{(j)}$ and w. Clearly, g_j is a monotone function of w_i for any given w_{-i}. Since $w^{(j)}(\Gamma)$ is a minimizer and $w_i^{(j)}(\Gamma) \in (\Gamma^{-1}, \Gamma)$, we must have $\Psi_{-i}^{(j)T} w_{-i} - \Psi_i^{(j)} \cdot \mathbf{1}_{n-1}^T w_{-i} = 0$. Define $\widetilde{w}^{(j)}(\Gamma)$ with

$$\widetilde{w}_i^{(j)}(\Gamma) = \Gamma, \quad \widetilde{w}_{-i}^{(j)}(\Gamma) = w_{-i}^{(j)}(\Gamma).$$

Then

$$g_j(\widetilde{w}^{(j)}(\Gamma)) = g_j(w^{(j)}(\Gamma)) = \min_{w\in[\Gamma^{-1},\Gamma]^n} g_j(w),$$

while

$$\sum_{i=1}^n \min\left\{|\widetilde{w}_i^{(j)}(\Gamma) - \Gamma|, |\widetilde{w}_i^{(j)}(\Gamma) - \Gamma^{-1}|\right\} < \sum_{i=1}^n \min\left\{|w_i^{(j)}(\Gamma) - \Gamma|, |w_i^{(j)}(\Gamma) - \Gamma^{-1}|\right\}.$$

This contradicts the definition of $w^{(j)}(\Gamma)$, so $w^{(j)}(\Gamma) \in \{\Gamma^{-1}, \Gamma\}^n$, which completes the proof of (Q.2).

For any $w \in \{\Gamma^{-1}, \Gamma\}^n$ with $|\{i : w_i = \Gamma\}| = k$, the Fréchet-Hoeffding inequality implies that Γ's are allocated to the k smallest entries of $\Psi^{(j)}$. Thus,

$$\min_{w\in\{\Gamma^{-1},\Gamma\}^n} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} = \min_{k\in\mathcal{N}\cup\{0\}} \frac{\Gamma k \bar{\Psi}_k^{(j)} + \Gamma^{-1}(\mathbf{1}_n^T \Psi_i^{(j)} - k \bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n - k)}.$$

By definition, $\mathbf{1}_n^T \Psi_i^{(j)} = j$. Then for each k, the above expression can be simplified as

$$\frac{\Gamma k \bar{\Psi}_k^{(j)} + \Gamma^{-1}(\mathbf{1}_n^T \Psi_i^{(j)} - k \bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n - k)} = \frac{\Gamma k \bar{\Psi}_k^{(j)} + \Gamma^{-1}(j - k \bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n - k)}$$

$$= \frac{(\Gamma - \Gamma^{-1})k \bar{\Psi}_k^{(j)} + \Gamma^{-1} j}{(\Gamma - \Gamma^{-1})k + \Gamma^{-1} n} = \frac{j}{n} + \frac{(\Gamma - \Gamma^{-1})k \left(\bar{\Psi}_k^{(j)} - \frac{j}{n}\right)}{(\Gamma - \Gamma^{-1})k + \Gamma^{-1} n}$$

$$= \frac{j}{n} + \frac{\bar{\Psi}_k^{(j)} - \frac{j}{n}}{1 + \frac{n}{k(\Gamma^2 - 1)}}.$$

The above expression is j/n for both $k = n$ and $k = 0$, so we can remove 0 from the minimum, and thus

$$\min_{w \in [\Gamma^{-1}, \Gamma]^n} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} = Q_j(\Gamma).$$

By (Q.1),

$$\bar{e}_\tau^{\mathbf{M}}(\Gamma) = \min \left\{ j : Q_j(\Gamma) \geq \tau \frac{(n-1)!}{(n-r-1)!} \right\}.$$

Finally, we can restrict to $j \geq \tau n!/(n-r-1)!$ because $Q_j(\Gamma) \leq \frac{j}{n}$ by taking $k = n$. $\quad\square$

Since $Q_j(\Gamma)$ is increasing in j, $J_\tau(\Gamma)$ can be found via binary search with iteration complexity $O(\log n^{r+1}) = O(r \log n)$. Each iteration costs at most $O(n)$ operations to sort the entries of $\Psi^{(j)}$ based on the ordered version of $\Psi^{(j-1)}$, since there is only entry updated, and $O(n)$ additional operations to compute $Q_j(\Gamma)$. Thus, the overall computational overhead after obtaining $(f_{(1)}, \ldots, f_{(|\mathbb{T}_{r+1,n}|)})$ is just $O(rn \log n)$, which is much smaller than the cost of sorting f-values $O(n^{r+1} \log n^{r+1}) = O(rn^{r+1} \log n)$.

In some cases, we want to compute $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ for all $\tau \in [0,1]$ at once. The following result links $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ to an induced distribution on the f's.

Corollary Q.1. *For any $\Gamma \geq 1$, let μ_Γ be a weighted measure with*

$$\mu_\Gamma = \sum_{j=1}^{|\mathbb{T}_{r+1,n}|} \frac{(n-r-1)!}{(n-1)!} (Q_j(\Gamma) - Q_{j-1}(\Gamma)) \cdot \delta_{f_{(j)}},$$

where $Q_0(\Gamma) = 0$. Then $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ is the τ-th quantile of μ_Γ.

Since the ordering takes $O(rn^{r+1} \log n)$ time and computing each $Q_j(\Gamma)$ takes $O(n)$ time, the total computational cost to compute $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ for all $\tau \in [0,1]$ is $O(rn^{r+1} \log n + n^{r+2})$.

Q.2. **Algorithm for evaluating everywhere dominance.** Let $f_{(j),1}$ and $f_{(j),2}$ be the j-th largest transfer errors for method 1 and 2, respectively. Similarly, the count vectors for two methods are denoted by $\Psi^{(j),1}$ and $\Psi^{(j),2}$. Then method 1 does NOT everywhere-upper-dominate method 2 at the τ-th quantile if and only if there exists $j_1, j_2 \in \{1, \ldots, |\mathbb{T}_{r+1,n}|\}$ and $W \in [0,\infty)^n$ such that

$$f_{(j_1),1} > f_{(j_2),2}, \quad \frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(j_1-1),1}}{w^T \mathbf{1}_n} < \tau \leq \frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(j_2),2}}{w^T \mathbf{1}_n}. \tag{Q.3}$$

Above $\Psi^{(0),1} = (0, 0, \ldots, 0)^T$.

To avoid pairwise comparisons, which incur $O(n^{2(r+1)})$ computation, we can check (Q.3) by only focusing on $j_1 = m(j), j_2 = j$ where

$$m(j) = \min\{j' : f_{(j'),1} > f_{(j),2}\}.$$

It is easy to see that (Q.3) holds for some pair $(j_1, j_2) \in \{1, \ldots, |\mathbb{T}_{r+1,n}|\}^2$ if and only if it holds for $(m(j), j)$ for some $j \in \{1, \ldots, |\mathbb{T}_{r+1,n}|\}$. For any given j, (Q.3) reduces to

$$\frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(m(j)-1),1}}{w^T \mathbf{1}_n} < \tau \leq \frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(j),2}}{w^T \mathbf{1}_n}, \quad w \in [0, \infty)^n.$$

This is equivalent to solving the following linear fractional programming problem and then checking if the objective is below τ:

$$\min \frac{w^T a^{(j)}}{w^T \mathbf{1}_n}, \quad \text{s.t.,} \quad \frac{w^T b^{(j)}}{w^T \mathbf{1}_n} \geq \tau, \quad w \in [0, \infty)^n,$$

where

$$a^{(j)} = \Psi^{(m(j)),1} \cdot \frac{(n-r-1)!}{n-1)!}, \quad b^{(j)} = \Psi^{(j),2} \cdot \frac{(n-r-1)!}{(n-1)!}.$$

We can apply the Charnes-Cooper transformation (Charnes and Cooper, 1962) by introducing $v = w/w^T \mathbf{1}_n$ to transform it into a linear programming problem:

$$\min v^T a^{(j)}, \quad \text{s.t.,} \quad v^T b^{(j)} \geq \tau, v^T \mathbf{1}_n = 1, v \in [0, \infty)^n. \tag{Q.4}$$

Solving these $O(n^{r+1})$ LP problems can be accelerated by the following two observations:

(1) Using the same argument as in the last step of the proof of Theorem Q.1, we can restrict

$$j \geq \tau \frac{n!}{(n-r-1)!}.$$

(2) When $a_i^{(j)} \geq b_i^{(j)}$ for every $i \in \mathcal{N}$, then the objective of (Q.4) can never be below τ.

APPENDIX R. SUPPLEMENTARY MATERIAL FOR SECTION 4

R.1. **Description of data.** We briefly describe the individual samples in our meta-data. There are 44 domains in total.

TABLE 4

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

R.2. **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 65% two-sided forecast intervals in Figure 8. These intervals are qualitatively similar to those reported in Figure 3.
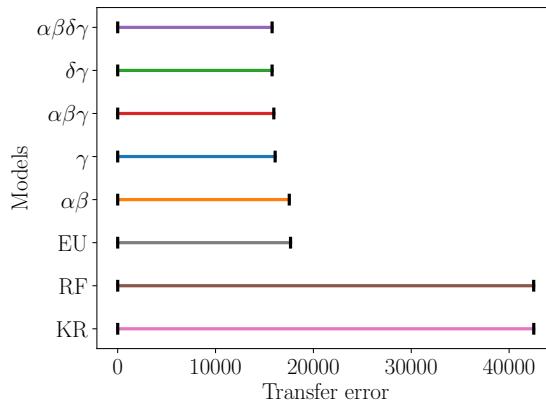
(A) *Transfer error*
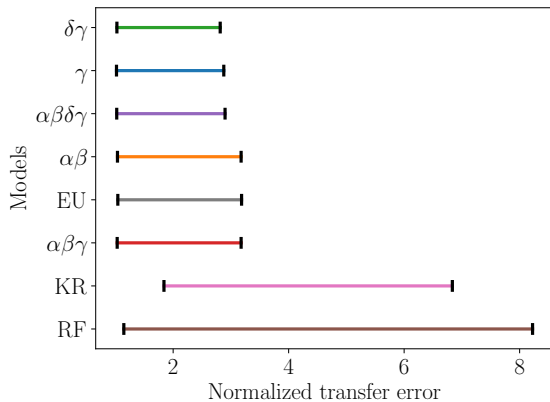
(B) *Normalized transfer error*
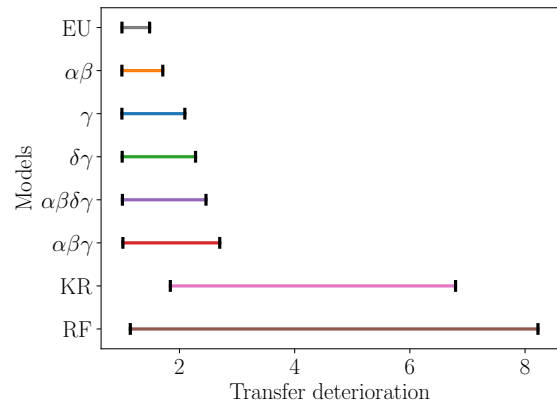
(C) *Transfer deterioration*

FIGURE 8. *65% (n=14, $\tau = 0.95$) forecast intervals for each of the three measures, treating each paper as a separate domain.*

R.3. **In-sample errors.** Figure 9 displays the CDFs of the in-sample errors of EU, CPT, the random forest algorithm, and kernel regression. As in Figure 2 (which reported out-of-sample errors), these curves are nearly indistinguishable.
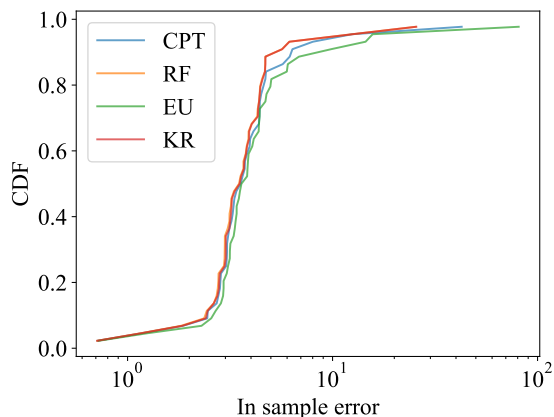

FIGURE 9. *CDF of in-sample errors*

R.4. **Supplementary tables and figures for main analysis.** Table 5 reports the forecast intervals that are depicted in Figure 3.

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 5. *81% (n=44, $\tau = 0.95$) forecast intervals*

R.5. **Alternative forecast intervals.** In this section, we report alternative forecast intervals for our three measures. Table 6 constructs 91% two-sided forecast intervals (setting $\tau = 1$),[38] and Table 7 reports 91% one-sided forecast intervals (setting $\tau = 0.95$). All of the

[38]The lower bounds of these intervals are the minimum transfer error (among the pooled transfer errors) and the upper bounds are the maximum transfer error.

forecast intervals are qualitatively similar to the 81% 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 6. *91% (n=44, $\tau = 1$) two-sided forecast intervals*

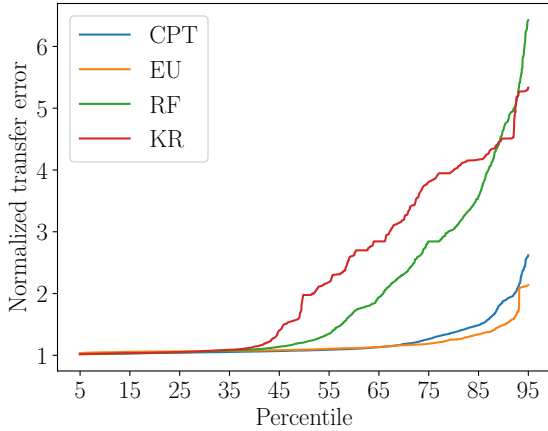| Model	Transfer Error	Normalized Error	Deterioration
CPT main variants			
γ	[0,15.83]	[1,2.54]	[1,1.47]
α,β	[0,16.13]	[1,2.35]	[1,1.30]
δ,γ	[0,17.19]	[1,2.47]	[1,1.53]
α,β,γ	[0,15.91]	[1,2.60]	[1,1.85]
$\alpha,\beta,\delta,\gamma$	[0,15.99]	[1,2.62]	[1,1.82]
EU models			
EU	[0,16.41]	[1,2.14]	[1,1.30]
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 7. *91% (n=44, $\tau = 0.95$) one-sided forecast intervals*

Finally, Figure 10 plots the τ-th percentile of the pooled transfer errors as τ varies. The figure shows that the qualitative conclusions we have drawn about the relative performance of black boxes and economic models are not specific to any choice of τ.[39] 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 are higher than those of the economic models for every choice of τ.
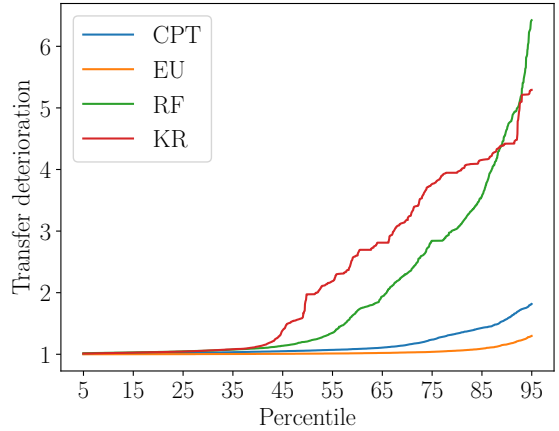
---

[39]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.
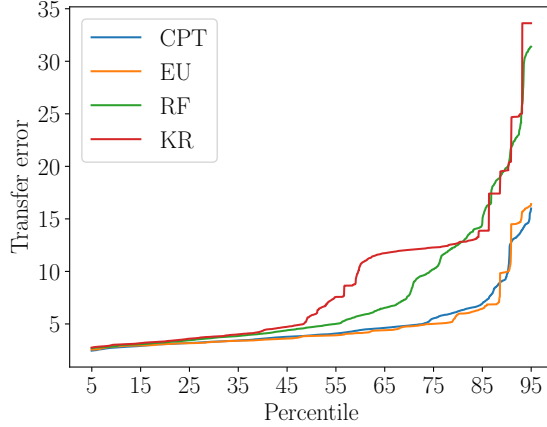
(A) *Transfer error*

(B) *Normalized transfer error*

(C) *Transfer deterioration*

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

R.6. **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
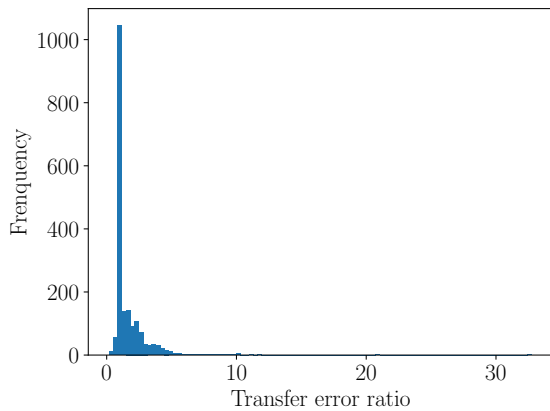
$$e_{\mathcal{T},d} = \frac{e(\rho^{RF}(\mathbf{M}_{\mathcal{T}}), S_d)}{e(\rho^{CPT}(\mathbf{M}_{\mathcal{T}}), S_d)}$$

to be the ratio of the random forest transfer error to the CPT transfer error, henceforth the *transfer error ratio.*
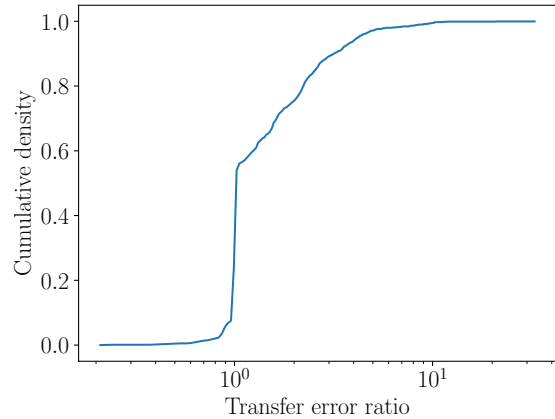
Panel (a) of Figure 11 reports 81% two-sided forecast intervals for the transfer error ratio for each CPT specification. The lower bound for each CPT model is approximately 0.9, while the upper bound is as large as 4.5. Panel (b) of the figure is a histogram of transfer

11
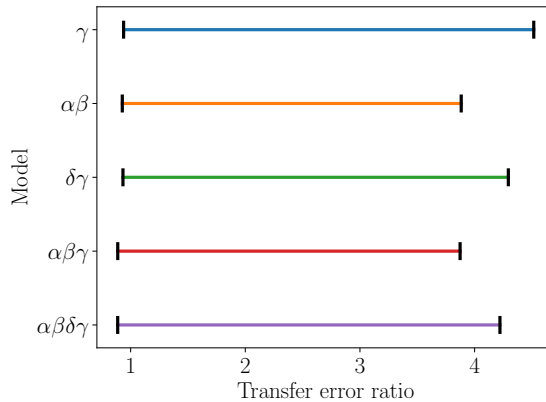
(A) *81% (n=44, τ = 0.95) Fore-cast intervals*

(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.*

error ratios for the 4-parameter CPT model when the training domains $\mathcal{T}$ 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.

R.7. **Alternative Choice of r.** Here we consider an alternative choice for the number of training domains, setting $r = 3$ instead of $r = 1$. 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. For this analysis we use domain cross-validation to select tuning parameters for the black box algorithms, as described in Example 3.
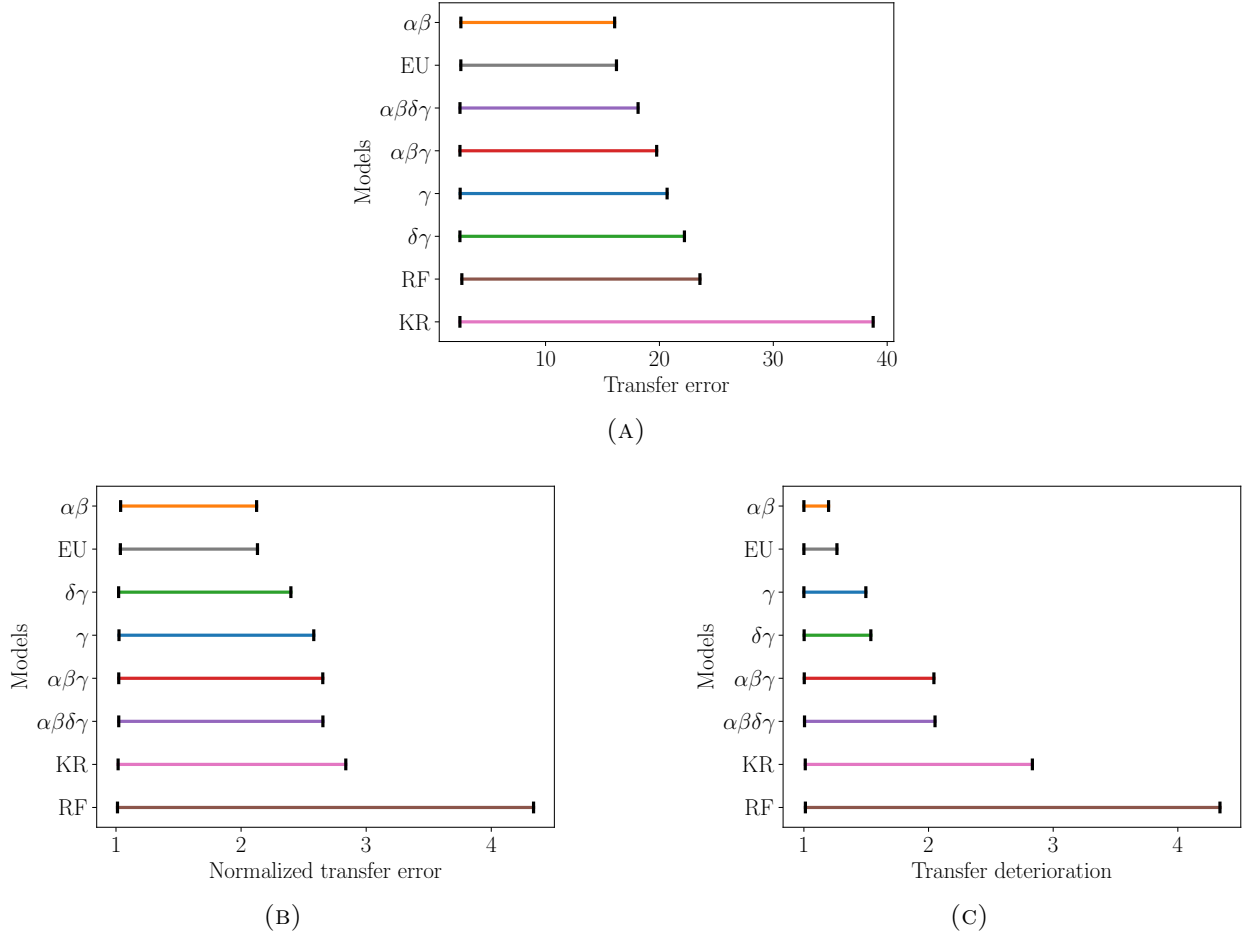

FIGURE 12. *73% (n=44, $\tau = 0.95$) forecast intervals for (a) transfer error, (b) normalized transfer error, and (c) transfer deterioration, with the choice of $r = 3$.*

Figure 12 is the analog of Figure 3. 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 73% forecast intervals. The most notable change is that the random forest forecast interval 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 $r = 1$ continue to hold. In particular, as with $r = 1$, 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.

R.8. **More details on worst-case dominance.** Figures 13 and 14 compare the worst case upper bound of the forecast intervals for CPT and RF for our three transfer measures as either γ or τ varies. In each case the dominance relation is clear.
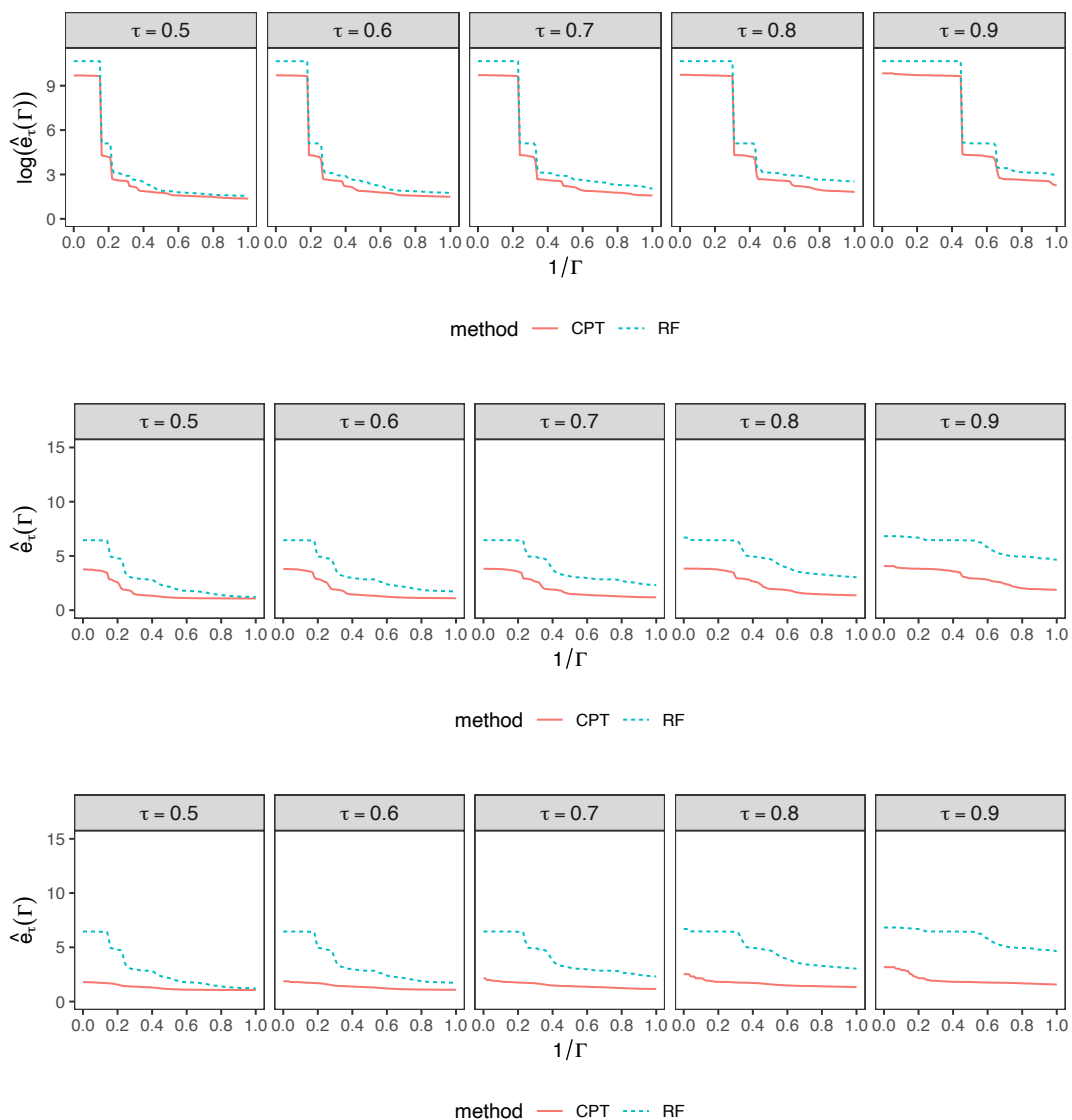

FIGURE 13. *The worst case upper prediction bound $\hat{e}_\tau(\Gamma)$ (as defined in (10)) for (a) transfer errors, (b) normalized transfer errors, and (c) transfer deterioration of CPT and RF as a function of $\Gamma \in [1, \infty)$, discretized at $100/i(i = 0, 1, \ldots, 100)$, at different quantiles $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.*
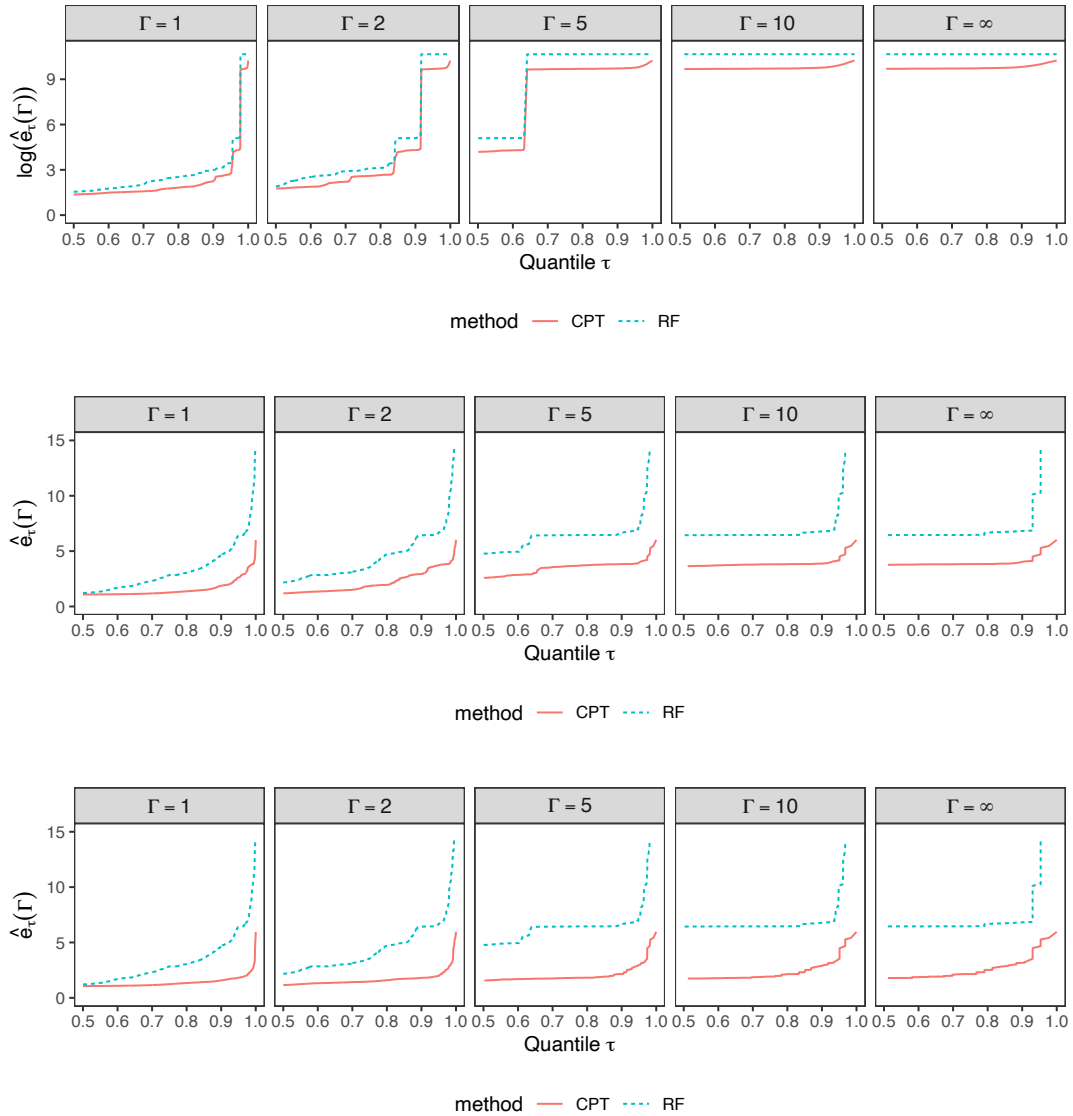
FIGURE 14. *The worst case upper prediction bound $\hat{e}_\tau(\Gamma)$ (as defined in (10)) for (a) transfer errors, (b) normalized transfer errors, and (c) transfer deterioration of CPT and RF as a function of $\tau \in [0.5, 1]$ without discretization for $\Gamma \in \{1, 2, 5, 10, \infty\}$.*