# Sequential Conditional Independence Testing with Machine Learning Models

**Angel Reyero Lobo** [*1 2 3] **Michele Meziu** [*3] **Sebastian Arias** [*3] **Peter Grünwald** [3 4]

## Abstract

Conditional Independence Testing is a ubiquitous problem in scientific discovery. The widely employed Model-X assumption shifts the modelling burden from the dependency of the output given the inputs, to the dependencies within inputs. Log-optimal e-variables have been studied in this setting, but it remains unclear how to optimally include machine learning models in these tests. In this work, we exploit the performance drop of a model when a given feature is removed to construct a coin-betting e-value for bounded losses and an exponential e-value that mimics density-based approaches. Surprisingly, in a misspecified setting, we show theoretically and experimentally that neither e-value uniformly outperforms the other. Finally, we provide actionable algorithms, including an antisymmetric kernel density estimation of the loss distribution.

## 1. Introduction

There has been a growing interest in recent years in *conditional independence testing*, notably in the context of scientific discovery. A representative example is identifying which genes are associated with a disease. Performing such tests prevents spurious discoveries caused by high correlations with features that are themselves important. Formally, we test whether a given input $X$ is independent of an output $Y$ given the remaining inputs $Z$, namely

$$X \perp\!\!\!\perp Y \mid Z.$$

This can be interpreted as assessing whether the feature $X$ provides exclusive information about the output that is not already contained in the remaining inputs. If so, the feature can be considered *important*. The set of such important features, namely the Markov blanket, is well-defined and unique whenever no input is an exact function of the others (Candès et al., 2018; Verdinelli & Wasserman, 2024).

Moreover, while traditional statistics has mainly focused on modeling the output as a function of the inputs, as in generalized linear models or generalized additive models, Candès et al. (2018) introduced the *model-X* assumption, which shifts the burden of modeling toward the dependence structure among the inputs themselves. This arguably milder assumption relies on the idea that the relationships among the inputs are often easier to model than their relationship with the output. For example, the inputs may be measurements collected from a common device and linked to a more complex phenomenon, as in the gene–disease setting. Alternatively, labeled data may be expensive to obtain, while large amounts of unlabeled data are available for studying the dependence structure among the inputs.

Furthermore, in related areas such as Interpretable Machine Learning, where the objective is to identify the *important* features contributing to the predictive accuracy of a model, several recent works have empirically and theoretically supported this assumption. Indeed, many standard methods rely on comparing the model performance when using the feature under study against the performance obtained when restricting access to that feature. These works have shown improved results for methods based on modeling the dependence structure among the inputs (e.g. Conditional Feature Importance) rather than retraining a new model without the feature while still focusing on the relationship with the output (e.g. Leave-One-Covariate-Out) (Chamma et al., 2024; Paillard et al., 2025; Reyero-Lobo et al., 2025).

The increasing use of Machine Learning has led to remarkable predictive performance on complex data (Hollmann et al., 2023; Qu et al., 2026). Consequently, *model-agnostic* conditional independence approaches have become particularly relevant, as they can leverage these complex models as surrogates for the underlying data-generating distribution.

At the same time, the process of data collection is no longer static, and scientists increasingly require more flexible inference procedures. For instance, when a $p$-value is close to a desired significance threshold, it is tempting to continue col-

[1]Univ Toulouse, INSA Toulouse, CNRS, IMT, Toulouse, France [2]Université Paris-Saclay, Inria, CEA, Palaiseau, 91120, France [3]Centrum Wiskunde & Informatica, The Netherlands [4]Leiden University, The Netherlands. Correspondence to: Angel Reyero Lobo <angel.reyerol-lobo@inria.fr>, Michele Meziu <Michele.Meziu@cwi.nl>, Sebastian Arias <Sebastian.Arias@cwi.nl>.

*ICML 2026 Workshop on Hypothesis Testing*, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

lecting additional data. Similarly, data may originate from multiple studies, making it unclear how to combine evidence appropriately. Standard $p$-value procedures do not generally maintain type-I error control under such flexible settings. For this reason, there has been growing interest in the study of *e-values* (Grünwald et al., 2024a; Ramdas & Wang, 2025). The e-value is a building block of Safe Anytime-Valid Inference (SAVI): it is a non-negative random variable whose expectation is bounded by one under the null hypothesis. Its sequential counterpart, the e-process, provides an interpretable measure of evidence (Koning, 2024) and allows practitioners to continuously monitor the data and decide whether additional data collection is necessary or sufficient evidence has already been obtained.

Building on Shaer et al. (2023), we aim to further investigate how to sequentially evaluate the importance of a feature within a black-box machine learning model under the model-X assumption, using conditional independence testing. Our approach compares the loss of the original model to that of a modified version where the feature's information is removed and replaced with a conditionally independent counterpart generated under the model-X assumption.

## 2. Related work

Candès et al. (2018) proposed the Conditional Randomization Test (CRT), a procedure based on the model-X assumption, which consists of assuming that the conditional distribution $\mathcal{L}(X \mid Z)$ of $X$ given $Z$ is known. The CRT computes an importance statistic $T(X, Y, Z)$ on the observed data. For instance, this statistic could be the coefficient of a lasso model trained to predict $Y$ from $(X, Z)$, so that the magnitude of the estimated coefficient $\hat{\beta}_X$ can be interpreted as the estimated importance of the feature $X$. If this coefficient is large, it indicates that $X$ is considered important even after accounting for $Z$.

To obtain validity under the null hypothesis, the CRT simulates the null distribution by replacing the original $X$ in the sample with a variable $\tilde{X}$ sampled from $\mathcal{L}(X \mid Z)$. The procedure then computes the importance statistic for many conditional resamplings, namely $T(\tilde{X}, Y, Z)$, and compares these values with the original statistic. This procedure provides $p$-values, and the authors also extended it to the multiple testing setting, leading to the popular knockoffs. The performance of the CRT highly depends on the choice of the statistic $T$, and there is a growing literature proposing different alternatives, ranging from predictive performance measures of machine learning models (Tansey et al., 2022) to coefficients of linear models (Liu et al., 2022).

In the sequential setting, the use of model-X to test conditional independence was also studied in Grünwald et al. (2024b). They proposed the e-value

$$E_{\mathrm{MX}} := \frac{f_{Y|Z,X}}{f_{Y|Z}} = \frac{f_{Y|Z,X}}{\mathbb{E}_{X|Z}[f_{Y|Z,X}]}.$$

Using the tower property and the conditional independence, we can verify that this is indeed a valid e-value, since

$$\begin{aligned}\mathbb{E}\left[\frac{f_{Y|Z,X}}{\mathbb{E}_{X|Z}[f_{Y|Z,X}]}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{f_{Y|Z,X}}{\mathbb{E}_{X|Z}[f_{Y|Z,X}]}\middle| Y, Z\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}_{X|Z}\left[\frac{f_{Y|Z,X}}{\mathbb{E}_{X|Z}[f_{Y|Z,X}]}\right]\right] = 1.\end{aligned} \tag{1}$$

They showed that this e-value is growth-rate optimal against any point alternative which satisfies the model-X assumption and has conditional density $f_{Y|X,Z}$. Nevertheless, this density is generally unknown in practice, and it is not straightforward to estimate it or relate it to machine learning models, which are typically designed for prediction rather than density estimation. We therefore connect this density-based perspective with pseudo-densities that arise naturally from machine learning methods, namely exponentiated losses. This idea is developed in Section 3.1.

A closely related approach based on the loss of a trained model was proposed by Shaer et al. (2023). In particular, they introduced the *eCRT*, which aims to provide a sequential version of the CRT. Their method relies on the fact that, under the null hypothesis, the quantities

$$q_t = T(X_t, Y_t, Z_t) \quad \text{and} \quad \tilde{q}_t = T(\tilde{X}_t, Y_t, Z_t)$$

are exchangeable for any importance statistic $T$. They take $T(X, Y, Z)$ to be the predictive performance of a model using inputs $(X, Z)$ to predict the output $Y$. They then propose the e-value

$$1 + v\, g(q_t, \tilde{q}_t),$$

where $v$ is a scalar and $g$ is an antisymmetric function, i.e.,

$$g(b, a) = -g(a, b).$$

By antisymmetry and exchangeability, it follows that under the null hypothesis,

$$\mathbb{E}[g(q_t, \tilde{q}_t)] = 0.$$

To ensure that this quantity defines a valid e-value, one additionally requires that $v\, g(q_t, \tilde{q}_t) \geq -1$, guaranteeing positivity. The authors proposed, for instance, using the sign of the difference between $q_t$ and $\tilde{q}_t$. Nevertheless, it is not clear which statistics are optimal in practice. In Section 3.2, we study settings in which no such transformation is needed because the loss itself is already bounded. In Section 3.4 we propose the optimal $g$-function for this exchangeability null hypothesis.

## 3. Proposed e-values

This section presents two e-values based on the preceding ideas and compares them under model misspecification—that is, when the true conditional density $f_{Y|X,Z}$ does not admit a pseudo-density representation based on the exponentiated loss. In such cases, the MX e-value estimate (approximated here by the exponential e-value introduced below) is no longer optimal.

Moreover, in order to use the coin-betting e-value, also presented below, we will work with bounded losses. For the unbounded loss setting, we refer to Appendix B.

The algorithms are given in Appendix A. We denote by $l$ the loss function and, for simplicity, in the analysis we will take the Bayes predictor denoted $m$.

### 3.1. Exponential e-value

This proposal aligns with the density e-value of Grünwald et al. (2024b) by constructing pseudo-probabilities (Grünwald & Mehta, 2020) by exponentiating the loss with a variance-like parameter $\eta$, normalized under the null:

$$E^{\eta}_{\exp} := \frac{\exp(-\eta\, l(m(X,Z),Y))}{\mathbb{E}_{\widetilde{X}|Z}\left[\exp\left(-\eta\, l\left(m(\widetilde{X},Z),Y\right)\right)\right]}. \tag{2}$$

### 3.2. Coin-betting e-value

Coin-betting e-variables have been proven admissible for testing the mean of a bounded random variable (Clerico, 2025; Larsson et al., 2026). Therefore, it is natural to study the following coin-betting version based on these losses:

$$E^{\lambda}_{\mathrm{cb}} = 1 + \lambda\left(-\, l(m(X,Z),Y) + \mathbb{E}_{\widetilde{X}|Z}\left[l\left(m(\widetilde{X},Z),Y\right)\right]\right). \tag{3}$$

We observe that this consists of a marginalization over $\widetilde{X}$ in the e-value of Shaer et al. (2023), where there is no need to apply the sign function since the loss is already bounded.

As a side remark, note that both $\eta$ and $\lambda$ may depend on $Y$ and $Z$ while still preserving the guarantees, by the same tower property argument as in (1). Indeed, we only require them to be $\sigma(Y,Z)$-measurable (see Lemma E.1).

### 3.3. Comparison of both e-values

We first note that none of the available optimality results applies exactly in our setting. Indeed, the exponential e-value is not growth-rate optimal since the alternative density $f_{Y|X,Z}$ is unknown and does not necessarily admit the exponentiated-loss form considered above. Likewise, the admissibility results for coin-betting e-values do not directly apply: although the loss is bounded, our setting additionally involves the Model-X assumption. Consequently, there is no universally preferable option, and we will show that neither method uniformly dominates the other.

We begin by considering the special case of the 0-1 loss. Surprisingly, in this setting, the e-values coincide when evaluated at their respective optimal $(Y,Z)$-dependent choices of $\lambda$ and $\eta$. Define

$$p_1 := \mathbb{P}_{X|Y,Z}\big(m(X,Z) \neq Y \mid Y,Z\big),$$

$$p_0 := \mathbb{P}_{\widetilde{X}|Z}\big(m(\widetilde{X},Z) \neq Y \mid Y,Z\big).$$

These correspond to the conditional probabilities of incorrect prediction when using $X$ and $\widetilde{X}$, respectively.

**Theorem 3.1** ($\ell_{0-1}$ loss)**.** *Assume that the loss is given by the* 0–1 *loss. Then, the optimal value of* $\lambda$ *maximizing the coin-betting e-value* (3) *is*

$$\lambda_\star = \frac{p_1 - p_0}{p_0(1-p_0)}.$$

*Similarly, the optimal* $\eta$ *for the exponential e-value* (2) *is*

$$\eta_\star = \log\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right).$$

*With these choices, both e-values coincide and are given by*

$$E^{\eta_\star}_{\exp} = E^{\lambda_\star}_{\mathrm{cb}} = \begin{cases} \dfrac{1-p_1}{1-p_0}, & m(X,Z) = Y, \\ \dfrac{p_1}{p_0}, & m(X,Z) \neq Y. \end{cases}$$

Now, let us define

$$\Delta = \mathbb{E}\left[l(m(\tilde{X},Z),Y)\right] - \mathbb{E}\left[l(m(X,Z),Y)\right],$$

which we refer to as the *predictiveness signal* of the feature. Based on our assumption that $m$ is the expectation-optimal predictor, we observe that $\Delta$ can be assumed nonnegative. Indeed, under the null hypothesis, we have $\Delta = 0$ by symmetry, whereas under the alternative we expect the feature $X$, which is no longer conditionally independent of $Y$, to provide useful predictive information. Consequently, the predictor using $X$ should achieve a lower loss than the one using the noise feature $\widetilde{X}$.

This quantity can be interpreted as an effect size in the problem: the larger $\Delta$ is, the more predictive the feature is and the easier the testing problem becomes. Conversely, when $\Delta$ is small, distinguishing the feature from noise becomes more difficult.

In the remainder of this section, we restrict attention to the small-signal regime. Indeed, when the signal is large, we expect the e-values to grow rapidly. However, the coin-betting e-value cannot become arbitrarily large since the admissible values of $\lambda$ are constrained to guarantee the nonnegativity of the e-value. By contrast, under standard tail assumptions, the exponential e-value may grow exponentially fast in the effect size.

For instance, assume that the loss based on $\widetilde{X}$ is $\sigma$-sub-Gaussian. We consider the e-power (Vovk & Wang, 2024), a standard notion in the e-values literature to quantify the performance of an e-variable. It satisfies

$$\begin{aligned}
&\mathbb{E}[\log(E_{\text{exp}}^{\eta})] \\
&= \mathbb{E}\left[-\eta l(m(X,Z),Y)\right] - \mathbb{E}\left[\log(E[e^{-\eta l(m(\tilde{X},Z),Y)}])\right] \\
&\geq \mathbb{E}\left[-\eta l(m(X,Z),Y)\right] + \eta\mathbb{E}\left[l(m(\tilde{X},Z),Y)\right] - \frac{\sigma^2\eta^2}{2} \\
&= \eta\Delta - \frac{\sigma^2\eta^2}{2}.
\end{aligned}$$

Optimizing over $\eta$ yields

$$\mathbb{E}\Big[\log(E_{\text{exp}}^{\eta^\star})\Big] \geq \frac{\Delta^2}{2\sigma^2}.$$

Hence, the e-power of the exponential e-value grows quadratically with the signal strength $\Delta$, and therefore diverges as the signal becomes large.

**Theorem 3.2** (Small-signal regime)**.** *Assume that the signal strength $\Delta$ is sufficiently small. Define*

$$\sigma_0^2 := \mathbb{V}(l(m(\tilde{X},Z),Y)), \qquad \sigma_1^2 := \mathbb{V}(l(m(X,Z),Y)).$$

*Then, up to second order in $\Delta$, the optimal parameter for the exponential e-value satisfies*

$$\eta^\star = \frac{\Delta}{\sigma_0^2} + O(\Delta^2).$$

*Moreover, the corresponding optimal e-power satisfies*

$$\mathcal{P}_{\text{exp}}^\star := \mathbb{E}\Big[\log E_{\text{exp}}^{\eta^\star}\Big] = \frac{\Delta^2}{2\sigma_0^2} + O(\Delta^3).$$

*Similarly, the optimal parameter for the coin-betting e-value satisfies*

$$\lambda^\star = \frac{\Delta}{\sigma_1^2 + \Delta^2} + O(\Delta^2),$$

*and the associated optimal e-power is*

$$\mathcal{P}_{\text{cb}}^\star := \mathbb{E}\Big[\log E_{\text{cb}}^{\lambda^\star}\Big] = \frac{1}{2}\frac{\Delta^2}{\sigma_1^2 + \Delta^2} + O(\Delta^3).$$

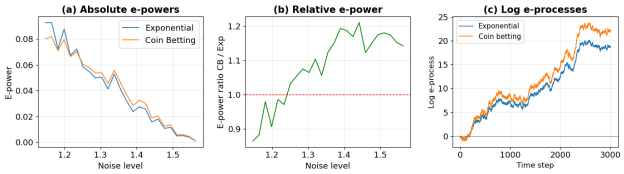


*Figure 1.* **Coin-betting can dominate the exponential:** On the left, the e-powers evaluated at the optimal choices of $\eta$ and $\lambda$. When the noise level is larger, the effect size becomes smaller, and the testing problem is more challenging. In this regime, we observe that $E_{\text{cb}}^{\lambda^\star}$ can outperform $E_{\text{exp}}^{\eta^\star}$. In the middle we plot the ratio of the e-powers of $E_{\text{cb}}^{\lambda^\star}$ over $E_{\text{exp}}^{\eta^\star}$. On the right, we show evidence growth at a fixed noise level, illustrating how e-power translates into faster evidence accumulation in practice. The experimental setting is described in Appendix D.2.

We typically expect $\sigma_0^2$ to be larger than $\sigma_1^2$, since the loss obtained using the informative feature $X$ should typically be smaller and less variable than the loss obtained using the noise feature $\widetilde{X}$. Yet, there exist settings in which $\sigma_1^2+\Delta^2 > \sigma_0^2$, and others in which the reverse holds, as illustrated in Figure 1. Thus, neither e-variable uniformly dominates the other.

### 3.4. KDE: Optimal antisymmetric e-variable

From Theorem 3.2, we observe that in the misspecified setting—where the exponential e-variable is not GRO because the density does not have the form of an exponentiated loss—it is possible to construct more powerful e-variables by using an antisymmetric function of the form

$$1 + g(q_t, \tilde{q}_t).$$

It is therefore natural to ask which choice of the function $g$ is optimal.

To answer this question, we adopt a different perspective. Rather than directly testing the original conditional model-X null hypothesis, we first compute the importance statistics $q_t$ and $\tilde{q}_t$ and then test their exchangeability, as proposed in (Shaer et al., 2023). This viewpoint simplifies the problem: directly characterizing the conditional density $f_{Y|X,Z}$ is generally intractable, whereas the machine learning model allows us to distill the relevant conditional information into the importance statistics. Consequently, we can focus on the testing problem after this reduction.

Moreover, we first observe that any e-variable that depends only on the pair of statistics $(q_t, \tilde{q}_t)$ is dominated by an e-variable of this antisymmetric form.

**Theorem 3.3.** *For any e-value $E = f(q_t, \tilde{q}_t)$ for the exchangeability null hypothesis, there exists another e-value*

$$E' = 1 + g(q_t, \tilde{q}_t) \geq E,$$

*where $g$ is antisymmetric.*

Moreover, the GRO for the exchangeability null hypothesis admits a closed-form expression:

**Theorem 3.4.** *GRO e-value $E^\star$ for this exchangeability null against an alternative $Q$ with density $f_q$ is given by*

$$E^\star := \frac{f_q(q_t, \tilde{q}_t)}{\frac{1}{2}\big(f_q(q_t, \tilde{q}_t) + f_q(\tilde{q}_t, q_t)\big)} = 1 + \frac{f_q(q_t, \tilde{q}_t) - f_q(\tilde{q}_t, q_t)}{f_q(q_t, \tilde{q}_t) + f_q(\tilde{q}_t, q_t)}.$$

Note that, in particular, the GRO is of the form $1 + g$, where $g$ is an antisymmetric function. The proof follows a change-of-measure argument similar to the one used to establish the symmetry result in (Larsson et al., 2025).

At this point, it is important to note that the reduction to the exchangeability problem ignores a substantial amount of information that could potentially be used to construct more powerful e-variables. Although we argue that the ML model extracts all the available information about the $Z$-conditional relationship between $X$ and $Y$ through the importance statistics, these statistics themselves satisfy additional properties that have not yet been incorporated, e.g. positivity of the loss. Furthermore, while the model-X assumption is only used to construct the statistic $\tilde{q}_t$, it can also be exploited for the exchangeability null hypothesis. For instance, it can be used to derandomize the resulting e-variable. Our proposal is then

$$E_{\mathrm{KDE}} := \mathbb{E}_{\tilde{X}|Z}\left[\frac{\hat{f}_q(q_t, \tilde{q}_t)}{\frac{1}{2}\big(\hat{f}_q(q_t, \tilde{q}_t) + \hat{f}_q(\tilde{q}_t, q_t)\big)}\right], \quad (4)$$

where $\hat{f}_q$ is an estimate of the joint distribution of $(q_t, \tilde{q}_t)$.

Although it may appear that we have just replaced the problem of learning one density with that of learning another, it is important to note that the latter is only a two-dimensional density. As a result, it is substantially easier to estimate in practice, for example using a kernel density estimator (KDE), for which standard implementations are readily available (e.g., (Pedregosa et al., 2011)). Consequently, our approach consists of sequentially estimating the joint density of the importance statistics; see Algorithm 2.

## 4. Experiments

In Figure 2 we partially recover the setting of (Shaer et al., 2023), where the predictive signal of $X$ is controlled by the coefficient $\beta$. Under the null hypothesis, none of the e-processes rejects, as expected. For small signals, the antisymmetric e-variable outperforms the exponential one, in agreement with Theorem 3.2. In contrast, for larger signals, the exponential accumulates evidence more rapidly, as it is not subject to the positivity constraint that bounds the antisymmetric by 2. Also, although the KDE approach should theoretically dominate all antisymmetric e-variables, this is not observed in practice. Understanding the practical impact of estimating the joint density of the importance statistics is an important direction for future work.

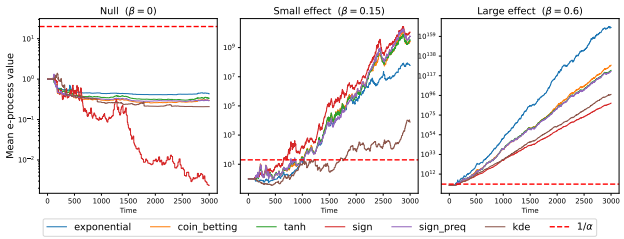


*Figure 2.* **Antisymmetric vs. exponential:** $Y = (Z^\top W)^2 + \beta X + \varepsilon$, with the setup full description in Appendix D.4. The parameter $\beta$ controls the predictive signal of $X$. Under the null ($\beta = 0$), neither e-process rejects. For small signal strengths, most antisymmetric e-processes outperform the exponential, whereas for larger signals, the exponential accumulates evidence faster.

Finally, in Appendix D.5, we apply our framework to HIV mutation data, illustrating its potential for using machine learning in scientific discovery. All code is available at `https://github.com/Antares01/sequential_ml_testing`.

## 5. Discussion

In this preliminary work, we compared two classes for sequential conditional independence testing under the Model-X assumption. Surprisingly, we observed that neither method uniformly dominates the other in the misspecified setting, which is arguably the most relevant practical regime, where the alternative is not given by a fixed density but instead must be approximated using a surrogate model.

Also, we presented the optimal function within the antisymmetric class. In our preliminary experiments, it does not yet exhibit the expected performance. Several factors may contribute to making this approach more effective, such as including sequentially forgetting early observations, since the model evolves over time. We will account for the impact of this *simple* density estimation error in the full paper.

Finally, in the full work we will study alternative null hypotheses that may be better aligned with the machine learning tools used in practice. Indeed, conditional independence is not always fully aligned with predictive usefulness. For example, a feature that only affects the conditional variance of the response may fail to improve predictive accuracy while still being conditionally dependent on the outcome. It is therefore important to investigate the behavior of the proposed e-values under such alternative predictive hypotheses.

## References


Barber, R. F. and Candès, E. J. Controlling the false discovery rate via knockoffs. 2015.

Candès, E., Fan, Y., Janson, L., and Lv, J. Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Chamma, A., Engemann, D. A., and Thirion, B. Statistically valid variable importance assessment through conditional permutations. *Advances in Neural Information Processing Systems*, 36, 2024.

Clerico, E. On the optimality of coin-betting for mean estimation. *International Journal of Approximate Reasoning*, 187:109550, 2025. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2025.109550. URL https://www.sciencedirect.com/science/article/pii/S0888613X25001914.

Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Grünwald, P., de Heide, R., and Koolen, W. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 11 2024a. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae011. URL https://doi.org/10.1093/jrsssb/qkae011.

Grünwald, P., Henzi, A., and Lardy, T. Anytime-valid tests of conditional independence under model-x. *Journal of the American Statistical Association*, 119 (546):1554–1565, 2024b. doi: 10.1080/01621459.2023.2205607. URL https://doi.org/10.1080/01621459.2023.2205607.

Grünwald, P. D. and Mehta, N. A. Fast rates for general unbounded loss functions: From erm to generalized bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020. URL http://jmlr.org/papers/v21/18-488.html.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations 2023*, 2023.

Koning, N. W. Continuous testing: Unifying tests and e-values. 2024. URL https://api.semanticscholar.org/CorpusID:272525409.

Larsson, M., Ramdas, A., and Ruf, J. The numeraire e-variable and reverse information projection. *The Annals of Statistics*, 53(3):1015 – 1043, 2025. doi: 10.1214/24-AOS2487. URL https://doi.org/10.1214/24-AOS2487.

Larsson, M., Ramdas, A., and Ruf, J. Testing hypotheses generated by constraints, 2026. URL https://arxiv.org/abs/2504.02974.

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293, June 2022. doi: 10.1093/biomet/asab039.

Lu, Y., Fan, Y., Lv, J., and Stafford Noble, W. Deeppink: reproducible feature selection in deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Paillard, J., Lobo, A. D. R., Kolodyazhniy, V., Thirion, B., and Engemann, D.-A. Measuring variable importance in heterogeneous treatment effects with confidence. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47456–47477. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/paillard25a.html.

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

Qu, J., Holzmüller, D., Varoquaux, G., and Le Morvan, M. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Ramdas, A. and Wang, R. Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1 (1–2):1–390, July 2025. ISSN 2978-4220. doi: 10.1561/3600000002. URL http://dx.doi.org/10.1561/3600000002.

Reyero-Lobo, A., Neuvial, P., and Thirion, B. Conditional feature importance revisited: Double robustness, efficiency and inference, 2025. URL https://arxiv.org/abs/2501.17520.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. Genotypic predictors of human immunodeficiency virus type 1 drug resistance.

*Proceedings of the National Academy of Sciences*, 103 (46):17355–17360, 2006.

Romano, Y., Sesia, M., and Candès, E. Deep knockoffs. *Journal of the American Statistical Association*, 115(532): 1861–1872, 2020.

Shaer, S. and Romano, Y. Learning to increase the power of conditional randomization tests. *Machine Learning*, 112 (7):2317–2357, 2023.

Shaer, S., Maman, G., and Romano, Y. Model-x sequential testing for conditional independence via testing by betting. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2054–2086. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/shaer23a.html.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and and, D. M. B. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162, 2022. doi: 10.1080/10618600.2021.1923520. URL https://doi.org/10.1080/10618600.2021.1923520.

Verdinelli, S. and Wasserman, L. Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4):623–636, November 2024. doi: 10.1214/24-STS911. URL https://projecteuclid.org/journals/statistical-science/volume-39/issue-4/Feature-Importance--A-Closer-Look-at-Shapley-Values-and/10.1214/24-STS911.full.

Vovk, V. and Wang, R. Nonparametric e-tests of symmetry. *The New England Journal of Statistics in Data Science*, 2(2):261–270, 2024. ISSN 2693-7166. doi: 10.51387/24-NEJSDS60.

Zhu, Z., Fan, Y., Kong, Y., Lv, J., and Sun, F. Deeplink: Deep learning inference using knockoffs with applications to genomics. *Proceedings of the National Academy of Sciences*, 118(36):e2104683118, 2021. doi: 10.1073/pnas.2104683118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2104683118.

## A. Algorithms

**Algorithm 1** Sequential Exponential Estimation Procedure

**Input:** Initial estimator $(\widehat{m}_1, \widehat{\eta}_1)$
Set $E_0 \leftarrow 1$

**for** $i = 1, 2, \ldots, \tau$ **do**

1. Sample $(X_i, Z_i, Y_i)$.
2. Compute
$$E_{\exp,i} := \frac{\exp(-\widehat{\eta}_i(Y_i, Z_i)\, l(\widehat{m}_i(X_i, Z_i), Y_i))}{\mathbb{E}_{\widetilde{X}|Z_i}\left[\exp\left(-\widehat{\eta}_i(Y_i, Z_i)\, l\left(\widehat{m}_i(\widetilde{X}, Z_i), Y_i\right)\right)\right]}.$$
3. Update $(\widehat{m}_{i+1}, \widehat{\eta}_{i+1})$ using the data
$$[i] = \{(X_j, Z_j, Y_j)\}_{j=1}^{i}.$$

**end for**

**Return**
$$\prod_{i=1}^{\tau} E_{\exp,i}.$$

Note that, in practice, one simple — although computationally expensive — way to update the model is to retrain it from scratch using the extended training dataset containing the newly observed point. However, this can be prohibitively intensive, and lighter alternatives are available, such as updating the existing model through stochastic gradient descent (SGD).

Regarding the choice of the optimal $\eta$, although it may in principle depend on $(Z, Y)$ and therefore be $\sigma(Z, Y)$-measurable, allowing for such flexibility considerably increases the complexity of the learning problem. In practice, we therefore typically restrict ourselves to constant values of $\eta$. For instance, $\eta$ can be learned in a prequential fashion by maximizing the empirical past log-growth:

$$\widehat{\eta}_i = \operatorname*{argmax}_{\eta \in \mathbb{R}} \sum_{k=1}^{i} \log\left(\frac{e^{-\eta\, l(\widehat{m}_k(X_k, Z_k), Y_k)}}{\mathbb{E}_{\widetilde{X}|Z_k}\left[e^{-\eta\, l(\widehat{m}_k(\widetilde{X}, Z_k), Y_k)}\right]}\right). \tag{5}$$

Note that the coin-betting e-value algorithm consists of replacing Step 2 with

$$E_{\mathrm{cb},i} := 1 + \widehat{\lambda}_i(\mathbb{E}_{\widetilde{X}|Z_i}[l(\widehat{m}_i(\tilde{X}, Z_i), Y_i)] - l(\widehat{m}_i(X_i, Z_i), Y_i)),$$

where $\widehat{\lambda}_i$ can also be learnt prequentially via

$$\widehat{\lambda}_i = \operatorname*{argmax}_{\lambda \in [0, 1/M]} \sum_{k=1}^{i} \log\left(1 + \lambda(\mathbb{E}_{\widetilde{X}|Z_i}[l(\widehat{m}_k(\tilde{X}, Z_k), Y_k)] - l(\widehat{m}_k(X_k, Z_k), Y_k))\right), \tag{6}$$

assuming that the loss is bounded by $M$.

## B. Coin betting with unbounded losses

Even if the loss is bounded, its bounds may be unknown *a priori*. Therefore, we can introduce a constant $M$ to control the loss:

$$E_{\mathrm{cb}}^{\lambda, M} = 1 + \lambda \frac{1}{M}\left(-l(\widehat{m}(X, Z), Y) \wedge M + \mathbb{E}_{\widetilde{X}|Z}\left[l\left(\widehat{m}(\widetilde{X}, Z), Y\right) \wedge M\right]\right). \tag{7}$$

By doing so, we can apply this coin-betting e-variable even when the loss is not inherently bounded.

Moreover, the parameter $M$, similarly to $\lambda$, can be learned in a prequential manner from past data. Specifically,

$$(\hat{\lambda}_n, \hat{M}_n) := \operatorname*{argmax}_{(\lambda,M)\in[0,1]\times\mathbb{R}_+} \prod_{i=1}^{n-1} \left(1 + \lambda \frac{1}{M}\left(-l(\widehat{m}_i(X_i, Z_i), Y_i) \wedge M + \mathbb{E}_{\widetilde{X}_i|Z_i}\left[l\left(\widehat{m}_i(\widetilde{X}_i, Z_i), Y_i\right) \wedge M\right]\right)\right).$$

Alternatively, since the main requirement is to ensure that the e-value remains nonnegative, we can directly clip the difference both from above and below. This yields

$$E_{\mathrm{cb}} = 1 + \lambda \frac{1}{M}\left(\left(-l(\widehat{m}(X, Z), Y) + l\left(\widehat{m}(\widetilde{X}, Z), Y\right)\right) \wedge M\right) \vee (-M). \tag{8}$$

Finally, we can apply Rao–Blackwellization to remove the randomness due to $\widetilde{X}$ by taking conditional expectations. Notably, this construction generalizes the sign e-value proposed by Shaer et al. (2023).

## C. KDE algorithm

**Algorithm 2** Sequential Kernel Density Estimation Procedure

**Input:** Initial estimator $(\widehat{m}_1, \widehat{\eta}_1, \hat{f}_{q,1})$
Set $E_0 \leftarrow 1$

**for** $i = 1, 2, \ldots, \tau$ **do**

1. Sample $(X_i, Z_i, Y_i)$.
2. Compute $q_{t,i} = l(\widehat{m}_i(X_i, Z_i), Y_i)$ and $\tilde{q}_{t,i} = l\left(\widehat{m}_i(\widetilde{X}, Z_i), Y_i\right)$
3. Compute
$$E_{\mathrm{KDE},i} := \mathbb{E}_{\tilde{X}|Z}\left[\frac{\hat{f}_q(q_t, \tilde{q}_t)}{\frac{1}{2}\left(\hat{f}_{q,i}(q_{t,i}, \tilde{q}_{t,i}) + \hat{f}_{q,i}(\tilde{q}_{t,i}, q_{t_i})\right)}\right]$$
4. Update $(\widehat{m}_{i+1})$ using the data $[i] = \{(X_j, Z_j, Y_j)\}_{j=1}^{i}$ and update $\hat{f}_{q,i+1}$ using $\{(q_{t,j}, \tilde{q}_{t,j})\}_{j=1}^{i}$.

**end for**

**Return**
$$\prod_{i=1}^{\tau} E_{\mathrm{KDE},i}.$$

Note that there are several possible ways to implement this e-process, and exploring these alternatives is an interesting direction for future work. In particular, the ML model is updated sequentially, so the distribution of the importance statistics may evolve over time. Consequently, it may be advantageous for the density estimator to gradually *forget* older observations and place greater emphasis on more recent ones. This can be achieved, for example, by using a sliding-window approach or by assigning exponentially decaying (or otherwise adaptive) weights to past observations, thereby allowing the estimated density to better track changes in the distribution of the importance statistics.

## D. Experiments

### D.1. Estimating the conditional distribution

In practice, we usually do not have access to the conditional distribution of $X \mid Z$, nor can we sample from it directly. Nevertheless, there is substantial ongoing research addressing this issue (see e.g. (Zhu et al., 2021)). When $X$ is discrete, it suffices to model the conditional class probabilities of $X$ given $Z$ and sample accordingly; this can be implemented, for instance, via logistic regression.

When $X$ is continuous, we adopt the same sampling strategy as in (Fisher et al., 2019): we first regress $X$ on $Z$ and then generate samples by adding permuted residuals from the fitted regression model. For further discussion on the guarantees of this conditional sampling procedure, we refer to (Reyero-Lobo et al., 2025).

### D.2. Coin-betting can beat Exponential setting (Figure 1)

In Figure 1 we illustrate the results from Theorem 3.2. To do so, we consider the simple model

$$Y = X + \gamma Z + \epsilon, \qquad X \sim \mathrm{Unif}([0,1]), \qquad \epsilon \sim B(\mathrm{Beta}(0.1, 0.1) - 0.5),$$

for different values of $B$. A $\mathrm{Beta}(0.1, 0.1)$-distributed random variable will assume values in $[0,1]$ concentrated near the boundaries. Shifting by $0.5$ centers the distribution, and multiplying by $B$ has the effect of scaling the noise. In particular, we do not include an additional $Z$ (i.e. we set $\gamma = 0$), and we sample $\widetilde{X} \sim \mathrm{Unif}([0,1])$, i.e., $q(X \mid Z) = \mathrm{Unif}([0,1])$. We take the model to be the oracle predictor $m(X) = X$. To ensure boundedness of the loss, we use the clipped squared loss

$$\ell(m(X), Y) = \min\big((X - Y)^2, 1\big),$$

which implies $\lambda \in [0,1]$.

On the left of Figure 1, we plot the e-power, $\mathbb{E}[\log E_{\mathrm{cb}}^{\lambda^\star}]$ and $\mathbb{E}[\log E_{\mathrm{exp}}^{\eta^\star}]$, evaluated at the optimal $\lambda^\star$ and $\eta^\star$, which are selected over a large grid of possible values. The expectations are approximated using Monte Carlo with 3000 samples. To simulate regimes with different effect sizes $\Delta$, we vary the noise level via the parameter $B$. We observe that when the noise level is small, the effect size is large, and it is beneficial to take larger bets; however, the coin-betting procedure is constrained by the requirement $\lambda \leq 1$, whereas the exponential e-value can still increase through larger values of $\eta$. In contrast, when $B$ is large, the problem becomes harder and the effect size decreases. In this regime, consistent with the theorem, the coin-betting e-value can outperform the exponential e-value. Here we focus on the region in $B$ where $E_{\mathrm{cb}}^{\lambda^\star}$ overtakes $E_{\mathrm{exp}}^{\eta^\star}$ in e-power.

In the middle we highlight the difference of the e-powers shown on the left by plotting their ratio.

On the right, we fix the noise level to a value in the small-signal regime ($B = 1.5$), where the coin-betting e-value has higher e-power than the exponential e-value. In this experiment, at each time step we use the true linear model and update the parameters $\eta$ and $\lambda$ in a prequential manner, as specified in Equations 5 and 6. We observe that the coin-betting evidence grows exponentially faster, indicating a clear advantage consistent with the gap in e-power shown on the left.

### D.3. E-values considered in the experiments

In the experiments shown in Figures 2 and 3, `exponential` denotes the exponential e-variable defined in (2). All the remaining methods are based on antisymmetric e-variables. In particular, `coin_betting` denotes the capped coin-betting e-variable with the learnable hyperparameter $M$ from (8). The method `tanh` uses the antisymmetric function

$$g_{\tanh}(q_t, \tilde{q}_t) = \tanh\big(20(q_t - \tilde{q}_t)\big),$$

while both `sign` and `sign_preq` use

$$g_{\mathrm{sign}}(q_t, \tilde{q}_t) = \mathrm{sign}(q_t - \tilde{q}_t).$$

Finally, `kde` corresponds to the proposed kernel density estimation approach defined in (4). Except for `sign`, all hyperparameters are learned prequentially. The `sign` method corresponds exactly to the procedure of (Shaer et al., 2023), which averages over all e-processes instead of selecting the maximum.

### D.4. Simulation setting (Figure 2)

We closely follow the simulation setting described by (Shaer et al., 2023) in section 5. The data is generated according to:

$$Y = (Z^\top W)^2 + \beta X + \varepsilon,\ X \mid Z \sim \mathcal{N}(U^\top Z, 1),\ Z \sim \mathcal{N}(0, I_d),$$

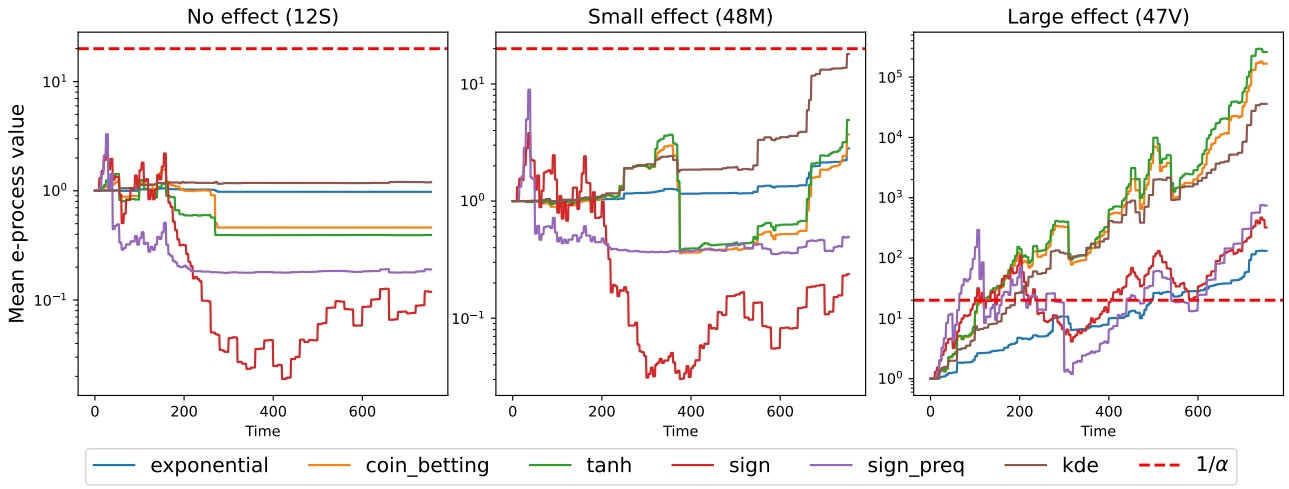


*Figure 3.* **HIV dataset:** Evidence collected against the null for three different mutations.

with $d = 19$. The only difference w.r.t. the original experiment is that here $U$ and $W$ are uniformly-sampled probability vectors of length $d$, as opposed to being sampled from independent Gaussian distributions. We introduce this change in order to decrease the variance across experiments, and to ensure that $\beta$ alone is a good representation of effect size in the experiment.
Although the true conditional distribution $X|Z$ is known exactly, this information is not fed to the e-processes, instead the Model-X kernel is learned over time in order to verify practical safety guarantees. The employed model $\widehat{m}$ is Gradient Boosting. In order to avoid retraining the model after each new datapoint, a batch size $b$ is fixed, so both model retraining and e-process adaptation happens after $b$ new datapoints are acquired.

We first note on the left that under the null hypothesis, no e-process rejects, thus confirming type-1 error guarantees. The middle figure confirms our theoretical results, showing that antisymmetric e-variables can outperform the exponential when the signal is small. The figure of the right-hand side shows that with large effects, the exponential e-variable dominates.

The experiment was performed independently on batches of size: [2, 5, 10, 20], and over 20 random seeds. Figure 2 represents the average outcome over all seeds and batch sizes.

### D.5. Real data: HIV dataset (Figure 3)

In this section we present a further experiment, based on the HIV drug resistance dataset [1] used by Shaer et al. (2023) to showcase their e-values, and previously in the knockoff literature by Barber & Candès (2015), Lu et al. (2018), Romano et al. (2020), Shaer & Romano (2023). The task is to detect whether a mutation in the human immunodeficiency virus (HIV) is associated to an increase in drug resistance Rhee et al. (2006). The dataset contains in $d = 150$ binary features and $n = 1555$ real-valued responses. In the experiment we focus on three of these features that were deemed representative by Shaer et al. (2023): one mutation that has not been shown to affect drug resistance (12S), one mutation that has a major effect on drug resistance but for which it is difficult to collect evidence (48M), and one that has a major effect on drug resistance (47V).

In Figure 3 we show for three different mutations the evidence accumulated by different e-variables against the conditional independence null. The plotted evidence is obtained as the average evidence collected by batch-wise e-processes with batch sizes $[5, 10, 20]$, as explained in Section D.4. On the left we observe that all e-processes are unable to collect evidence against the null, since the mutation '12S' has no proven effect on drug resistance. In the middle plot the KDE-based method shows more evidence against the null than all the other methods we consider, despite being unable to reject the null for a mutation affecting drug resistance (48M). On the right all the methods we consider rech an evidence of at least $1/\alpha$ and thus reject the null hypothesis for mutation '47V', which has been reported to have a major effect. We observe that the KDE-based e-variable is competitive with the other antisymmetric betting strategies.

[1]Dataset available at https://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/

# E. Proofs

## E.1. $\sigma(Y, Z)$-measurability of $\lambda$ and $\eta$

**Lemma E.1.** *Any $\hat{\eta}_n$ that is $\sigma\big(Y_n, Z_n, \{(X_i, Y_i, Z_i)\}_{i=1}^{n-1}\big)$ -measurable provides a valid exponential supermartingale. Similarly, any $\hat{\lambda}_n$ that is $\sigma\big(Y_n, Z_n, \{(X_i, Y_i, Z_i)\}_{i=1}^{n-1}\big)$ -measurable provides a valid coin-betting supermartingale.*

*Proof.* We denote by $\mathcal{D}_{n-1} := \{(X_i, Y_i, Z_i)\}_{i=1}^{n-1}$ the previous observed data. We observe that

$$\mathbb{E}\left[\frac{\exp(-\widehat{\eta}\, l(\widehat{m}(X,Z),Y))}{\mathbb{E}_{\widetilde{X}|Z}\Big[\exp\Big(-\widehat{\eta}\, l\Big(\widehat{m}(\widetilde{X},Z),Y\Big)\Big)\Big]}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\exp(-\widehat{\eta}\, l(\widehat{m}(X,Z),Y))}{\mathbb{E}_{\widetilde{X}|Z}\Big[\exp\Big(-\widehat{\eta}\, l\Big(\widehat{m}(\widetilde{X},Z),Y\Big)\Big)\Big]} \mid Y, Z, \mathcal{D}_{n-1}\right]\right]$$
$$= \mathbb{E}\left[\mathbb{E}_{\widetilde{X}|Z}\left[\frac{\exp(-\widehat{\eta}\, l(\widehat{m}(X,Z),Y))}{\mathbb{E}_{\widetilde{X}|Z}\Big[\exp\Big(-\widehat{\eta}\, l\Big(\widehat{m}(\widetilde{X},Z),Y\Big)\Big)\Big]}\right]\right] = 1.$$

The term $\mathbb{E}_{\widetilde{X}|Z}\left[\frac{\exp(-\widehat{\eta}\, l(\widehat{m}(X,Z),Y))}{\mathbb{E}_{\widetilde{X}|Z}\Big[\exp\Big(-\widehat{\eta}\, l\Big(\widehat{m}(\widetilde{X},Z),Y\Big)\Big)\Big]}\right]$ simplifies to 1 for every choice of $\widehat{\eta}(Y, Z)$ dependent on $Y, Z, \mathcal{D}_{n-1}$. □

Therefore, applying Ville's inequality, we obtain valid-$\alpha$-type-I error control for the sequential threshold at $1/\alpha$.

## E.2. Optimal e-values for the $0$-$1$ loss

In this $0$–$1$ loss setting we can provide an explicit expression for the e-values that optimize the e-power across the coin-betting and model-X classes, which turn out to coincide.

We begin by optimizing $\eta$ for $E^{\eta}_{\mathrm{ML}}$. Observe that

$$\mathbb{E}_{X|Y,Z}\left[\log\left(E^{\eta}_{\mathrm{ML}}\right)\right] = \mathbb{E}_{X|Y,Z}\left[-\eta \mathbb{1}_{\{m(X,Z)\neq Y\}} - \log \mathbb{E}_{\widetilde{X}|Z}\left[e^{-\eta \mathbb{1}_{\{m(\widetilde{X},Z)\neq Y\}}}\right]\right].$$

Differentiating with respect to $\eta$ gives

$$-\mathbb{E}_{X|Y,Z}\left[\mathbb{1}_{\{m(X,Z)\neq Y\}}\right] + \frac{\mathbb{E}_{\widetilde{X}|Z}\left[\mathbb{1}_{\{m(\widetilde{X},Z)\neq Y\}} e^{-\eta \mathbb{1}_{\{m(\widetilde{X},Z)\neq Y\}}}\right]}{\mathbb{E}_{\widetilde{X}|Z}\left[e^{-\eta \mathbb{1}_{\{m(\widetilde{X},Z)\neq Y\}}}\right]}.$$

Let

$$p_1 = \mathbb{P}_{X|Y,Z}(m(X,Z) \neq Y \mid Y, Z), \qquad p_0 = \mathbb{P}_{\widetilde{X}|Z}(m(\widetilde{X},Z) \neq Y \mid Y, Z).$$

Note that

$$\mathbb{E}_{\widetilde{X}|Z}\left[\mathbb{1}_{\{m(\widetilde{X},Z)\neq Y\}} e^{-\eta \mathbb{1}_{\{m(\widetilde{X},Z)\neq Y\}}}\right] = e^{-\eta} p_0.$$

Thus the first-order condition becomes

$$-p_1 + \frac{e^{-\eta_\star} p_0}{e^{-\eta_\star} p_0 + 1 - p_0} = 0.$$

Solving yields

$$\eta_\star = \log \frac{p_0(1-p_1)}{p_1(1-p_0)}.$$

With this optimal $\eta_\star$, the e-variable $E_{\mathrm{ML}}^{\eta_\star}$ is

$$\begin{aligned}
E_{\mathrm{ML}}^{\eta_\star} &= \frac{e^{-\eta_\star l(m(X,Z),Y)}}{\mathbb{E}_{\widetilde{X}|Z}\left[e^{-\eta_\star l(m(\widetilde{X},Z),Y)}\right]} \\
&= \frac{(e^{-\eta_\star})^{\mathbb{1}_{\{m(X,Z)\neq Y\}}}}{p_0 e^{-\eta_\star} + 1 - p_0} \\
&= \frac{\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{\mathbb{1}_{\{m(X,Z)\neq Y\}}}}{p_0 \frac{p_1(1-p_0)}{p_0(1-p_1)} + 1 - p_0} \\
&= \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{\mathbb{1}_{\{m(X,Z)\neq Y\}}} \frac{1-p_1}{1-p_0}.
\end{aligned}$$

Therefore, there are two cases:

$$E = \begin{cases} \dfrac{1-p_1}{1-p_0}, & m(X,Z) = Y, \\ \dfrac{p_1}{p_0}, & m(X,Z) \neq Y. \end{cases}$$

We can proceed similarly for the coin-betting e-value. Optimizing $\lambda$ yields

$$\lambda_\star = \frac{p_1 - p_0}{p_0(p_0 - 1)}.$$

Substituting this value into the e-variable gives two cases.

If the prediction is correct, $Y = m(X,Z)$, then

$$E_{\mathrm{CB}} = 1 + \lambda_\star p_0 = 1 + \frac{p_1 - p_0}{p_0 - 1} = \frac{1-p_1}{1-p_0}.$$

If the prediction is incorrect,

$$E_{\mathrm{CB}} = 1 + \frac{p_1 - p_0}{p_0(p_0-1)}(p_0 - 1) = \frac{p_1}{p_0}.$$

Thus the coin-betting and model-X optimal e-variables coincide.

### E.3. E-power in the small-signal regime

**Exponential e-power expansion:**

*Proof.* First, note that the e-power of the exponential e-value is given by

$$\mathcal{P}_{\mathrm{exp}}^{\eta} = \mathbb{E}\big[\log\big(E_{\mathrm{exp}}^{\eta}\big)\big] = \mathbb{E}\Big[-\eta\, l(m(X,Z),Y) - \log \mathbb{E}_{\widetilde{X}|Z}\Big[e^{-\eta l(m(\widetilde{X},Z),Y)}\Big]\Big].$$

Recall that the inner expectation is formally taken with respect to $P_{\widetilde{X}|X,Y,Z}$. However, since $\widetilde{X} \perp\!\!\!\perp (X,Y) \mid Z$, we simply write $P_{\widetilde{X}|Z}$ to avoid cumbersome notation.

Define

$$\psi(\eta) := \log \mathbb{E}_{\widetilde{X}|Z}\Big[e^{-\eta l(m(\widetilde{X},Z),Y)}\Big].$$

Using a Taylor expansion around $\eta = 0$,

$$\psi(\eta) = \psi(0) + \eta\psi'(0) + \frac{\eta^2}{2}\psi''(0) + O(\eta^3).$$

Since

$$\psi(0) = \log \mathbb{E}_{\widetilde{X}|Z}[1] = 0,$$

we have

$$\psi(0) = 0.$$

Differentiating,

$$\psi'(\eta) = -\frac{\mathbb{E}_{\widetilde{X}|Z}\Big[l(m(\widetilde{X}, Z), Y)e^{-\eta l(m(\widetilde{X},Z),Y)}\Big]}{\mathbb{E}_{\widetilde{X}|Z}\Big[e^{-\eta l(m(\widetilde{X},Z),Y)}\Big]},$$

and therefore

$$\psi'(0) = -\mathbb{E}_{\widetilde{X}|Z}\Big[l(m(\widetilde{X}, Z), Y)\Big].$$

Differentiating once more yields

$$\psi''(\eta) = \frac{\mathbb{E}_{\widetilde{X}|Z}\Big[l(m(\widetilde{X}, Z), Y)^2 e^{-\eta l(m(\widetilde{X},Z),Y)}\Big]\ \mathbb{E}_{\widetilde{X}|Z}\Big[e^{-\eta l(m(\widetilde{X},Z),Y)}\Big]}{\Big(\mathbb{E}_{\widetilde{X}|Z}\Big[e^{-\eta l(m(\widetilde{X},Z),Y)}\Big]\Big)^2} - \frac{\Big(\mathbb{E}_{\widetilde{X}|Z}\Big[l(m(\widetilde{X}, Z), Y)e^{-\eta l(m(\widetilde{X},Z),Y)}\Big]\Big)^2}{\Big(\mathbb{E}_{\widetilde{X}|Z}\Big[e^{-\eta l(m(\widetilde{X},Z),Y)}\Big]\Big)^2}.$$

Evaluating at $\eta = 0$, we obtain

$$\begin{aligned}\psi''(0) &= \mathbb{E}_{\widetilde{X}|Z}\Big[l(m(\widetilde{X}, Z), Y)^2\Big] - \Big(\mathbb{E}_{\widetilde{X}|Z}\Big[l(m(\widetilde{X}, Z), Y)\Big]\Big)^2 \\ &= \mathrm{Var}_{\widetilde{X}|Z}\Big(l(m(\widetilde{X}, Z), Y)\Big) \\ &=: \sigma_0^2.\end{aligned}$$

Substituting the Taylor expansion into the e-power expression gives

$$\begin{aligned}\mathcal{P}^{\eta}_{\exp} &= -\eta\, \mathbb{E}[l(m(X, Z), Y)] \\ &\quad - \mathbb{E}\Big[-\eta\, \mathbb{E}_{\widetilde{X}|Z}\Big[l(m(\widetilde{X}, Z), Y)\Big] + \frac{\eta^2}{2}\ \mathrm{Var}_{\widetilde{X}|Z}\Big(l(m(\widetilde{X}, Z), Y)\Big) + O(\eta^3)\Big] \\ &= -\eta\Big(\mathbb{E}[l(m(X, Z), Y)] - \mathbb{E}\big[\mathbb{E}_{\widetilde{X}|Z}[l(m(\widetilde{X}, Z), Y)]\big]\Big) \\ &\quad - \frac{\eta^2}{2}\, \mathbb{E}\Big[\mathrm{Var}_{\widetilde{X}|Z}\Big(l(m(\widetilde{X}, Z), Y)\Big)\Big] + O(\eta^3).\end{aligned}$$

Recall that

$$\Delta := \mathbb{E}\big[l(m(\widetilde{X}, Z), Y)\big] - \mathbb{E}[l(m(X, Z), Y)]$$

and

$$\sigma_0^2 := \mathbb{E}\Big[\mathrm{Var}_{\widetilde{X}|Z}\Big(l(m(\widetilde{X}, Z), Y)\Big)\Big].$$

Then

$$\mathcal{P}^{\eta}_{\exp} = \eta\Delta - \frac{\eta^2}{2}\sigma_0^2 + O(\eta^3).$$

Ignoring higher-order terms and maximizing the quadratic approximation yields

$$\eta^* = \frac{\Delta}{\sigma_0^2},$$

and

$$\mathcal{P}^*_{\exp} = \frac{\Delta^2}{2\sigma_0^2}.$$

□

**Coin-betting e-power expansion:**

*Proof.* We recall that the coin-betting e-value is given by

$$E^\lambda_{\mathrm{cb}} = 1 - \lambda\Big(l(\widehat{m}(X,Z),Y) - \mathbb{E}_{\widetilde{X}|Z}\Big[l\Big(\widehat{m}(\widetilde{X},Z),Y\Big)\Big]\Big)\,,$$

where the expectation is taken only over $\widetilde{X}$ condition both on $Z$ and $Y$, but for simplicity, and to underline that $\widetilde{X}$ is independent of $Y$, we hide this conditioning on $Y$. Its e-power is

$$\mathcal{P}^\lambda_{\mathrm{cb}} = \mathbb{E}\big[\log\big(E^\lambda_{\mathrm{cb}}\big)\big] = \mathbb{E}\Big[\log\Big(1 - \lambda\Big(l(\widehat{m}(X,Z),Y) - \mathbb{E}_{\widetilde{X}|Z}\Big[l\Big(\widehat{m}(\widetilde{X},Z),Y\Big)\Big]\Big)\Big)\Big]\,.$$

Using the Taylor expansion

$$\log(1+x) = x - \frac{x^2}{2} + O(x^3),$$

we obtain

$$\mathcal{P}^\lambda_{\mathrm{cb}} = \mathbb{E}\Big[\lambda D + \frac{1}{2}\lambda^2 D^2\Big] + O(\lambda^3),$$

where

$$D := \mathbb{E}_{\widetilde{X}|Z}\Big[l\Big(\widehat{m}(\widetilde{X},Z),Y\Big)\Big] - l(\widehat{m}(X,Z),Y)\,.$$

Recall that the predictiveness signal is

$$\Delta := \mathbb{E}[D] = \mathbb{E}\Big[\mathbb{E}_{\widetilde{X}|Z}\Big[l\Big(\widehat{m}(\widetilde{X},Z),Y\Big)\Big]\Big] - \mathbb{E}[l(\widehat{m}(X,Z),Y)]\,,$$

and define

$$\sigma_1'^2 := \mathbb{E}[D^2].$$

Then

$$\mathcal{P}^\lambda_{\mathrm{cb}} = \lambda\Delta - \frac{1}{2}\lambda^2\sigma_1'^2 + O(\lambda^3).$$

Ignoring higher-order terms, the optimal value of $\lambda$ is obtained by maximizing the quadratic approximation, yielding

$$\lambda^* = \frac{\Delta}{\sigma_1'^2}.$$

Substituting this value back gives

$$\mathcal{P}^*_{\mathrm{cb}} = \frac{\Delta^2}{2\sigma_1'^2}.$$

We conclude by noting that $\sigma_1'^2 = \sigma_1^2 + \Delta^2$. □

### E.4. Antisymmetric e-value's class

#### E.4.1. Optimality of the 1 + Antisymmetric for exchangeability

*Proof.* Indeed, for any e-variable $E = f(t, \widetilde{t})$, we have

$$f(t, \widetilde{t}) = \frac{1}{2}\big(f(t, \widetilde{t}) + f(\widetilde{t}, t)\big) + \frac{1}{2}\big(f(t, \widetilde{t}) - f(\widetilde{t}, t)\big).$$

The second term is antisymmetric. Moreover, for all $t$ and $\widetilde{t}$ the following is in the null hypothesis,

$$P = \frac{1}{2}\big(\delta_{t,\widetilde{t}} + \delta_{\widetilde{t},t}\big),$$

and therefore

$$\frac{1}{2}\big(f(t, \widetilde{t}) + f(\widetilde{t}, t)\big) \leq 1.$$

Verifying that $1 + \frac{1}{2}\big(f(t, \widetilde{t}) - f(\widetilde{t}, t)\big)$ is an e-variable proves the claim. □

#### E.4.2. KDE optimality

*Proof.* The proof consists of a change-of-measure argument similar to that in Larsson et al. (2025). Indeed, denoting by $f_q$ the density under the alternative $Q$ and by $\tilde{P}$ the measure with density

$$p(q_t, \tilde{q}_t) = \frac{1}{2}\big(f_q(q_t, \tilde{q}_t) + f_q(\tilde{q}_t, q_t)\big)\, \mathbf{1}_{\{f_q(q_t, \tilde{q}_t) > 0\}},$$

we obtain

$$\begin{aligned}
\mathbb{E}_Q\left[\frac{E}{E^\star}\right] &= \mathbb{E}_Q\left[\frac{1 + g(q_t, \widetilde{q}_t)}{\frac{f_q(q_t, \widetilde{q}_t)}{\frac{1}{2}\left(f_q(q_t, \widetilde{q}_t) + f_q(\widetilde{q}_t, q_t)\right)}}\right] \\
&= \mathbb{E}_{\tilde{P}}\left[(1 + g(q_t, \widetilde{q}_t))\, \mathbf{1}_{\{f_q(q_t, \widetilde{q}_t) > 0\}}\right] \\
&\leq \mathbb{E}_{\tilde{P}}[1 + g(q_t, \widetilde{q}_t)] = 1.
\end{aligned}$$

□