Ensuring Reliability via Hyperparameter Selection: Review and Advances

Amirmohammad Farzaneh and Osvaldo Simeone
Centre for Intelligent Information Processing Systems,
Department of Engineering,
King's College London,
London, United Kingdom
{amirmohammad.farzaneh, osvaldo.simeone}@kcl.ac.uk

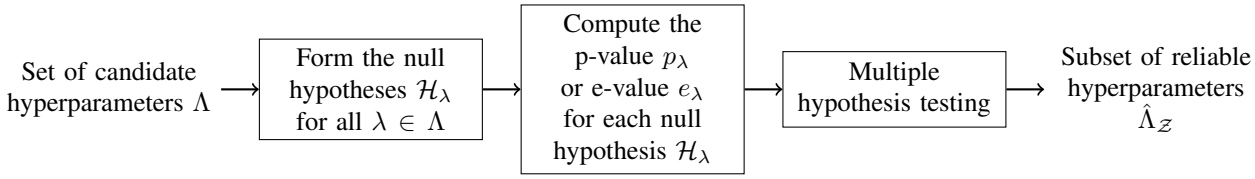

Fig. 1: Block diagram for LTT-based hyperparameter selection methods.

***Abstract*—Hyperparameter selection is a critical step in the deployment of artificial intelligence (AI) models, particularly in the current era of foundational, pre-trained, models. By framing hyperparameter selection as a multiple hypothesis testing problem, recent research has shown that it is possible to provide statistical guarantees on population risk measures attained by the selected hyperparameter. This paper reviews the Learn-Then-Test (LTT) framework, which formalizes this approach, and explores several extensions tailored to engineering-relevant scenarios. These extensions encompass different risk measures and statistical guarantees, multi-objective optimization, the incorporation of prior knowledge and dependency structures into the hyperparameter selection process, as well as adaptivity. The paper also includes illustrative applications for communication systems.**

***Index Terms*—hyperparameter selection, multiple hypothesis testing, reliability**

I. Introduction

The selection of hyperparameters is a fundamental challenge in the optimization of machine learning models, directly impacting efficiency and generalization [7]. Traditional approaches rely on heuristic search strategies, such as grid search and Bayesian optimization, which prioritize empirical performance but fail to provide formal statistical guarantees [8], [9]. The lack of performance guarantees may prevent the use of AI models in sensitive domains, such as healthcare [10] and engineering [11]. Recently, a growing body of research has explored hyperparameter selection from a statistical inference perspective, treating it as a multiple hypothesis testing (MHT) problem. This paradigm provides a principled way to control error rates in the hyperparameter tuning process, making it possible to select hyperparameters in a manner that ensures formal reliability targets. The Learn-Then-Test (LTT) framework [1] formalizes this perspective by structuring hyperparameter selection as a two-stage process. As illustrated in Figure 1, this consists of the following phases: (*i*) Determine a set Λ of candidate hyperparameters using any methodology; (*ii*) Test the performance of the candidate hyperparameters via MHT to determine a subset $\hat{\Lambda}_{\mathcal{Z}} \subseteq \Lambda$ of hyperparameters with formal statistical performance guarantees. LTT has been applied to a number of settings, including language models [12] and vision models [13].

LTT focuses on the control of average risk measures by using standard MHT methods that target the *family-wise error rate* (FWER) criterion. Accordingly, the goal is to ensure that the produced subset $\hat{\Lambda}_{\mathcal{Z}}$ of hyperparameters include no unreliable configurations with sufficiently high probability. Beyond this standard LTT formulation, real-world engineering applications introduce additional complexities, such as the need to account for risk-aware performance measures [2], [14], the freedom to explore multi-objective trade-offs [15], the presence of prior domain knowledge [16], and the cost of acquiring data [17]. As summarized in Table I, addressing these aspects requires extensions to the LTT framework, which are reviewed in this tutorial-style paper.

Specifically, Section II provides a review of the fundamental LTT framework. Each subsequent section explores one of the aspects described by the columns of Table I. Specifically, Section III explores different risk measures; Section IV covers statistical guarantees; Section V discusses multi-objective optimization criteria; the integration of side information is covered in Section VI; and adaptive LTT-based methods are reviewed in Section VII. Conclusions and future research directions are presented in Section VIII.

This work was supported by the European Union's Horizon Europe project CENTRIC (101096379), by the Open Fellowships of the EPSRC (EP/W024101/1), and by the EPSRC project (EP/X011852/1).

TABLE I: Comparison of LTT-based hyperparameter selection methods

Scheme	Risk Measure	Statistical Guarantee	Multi-Objective	Side Information	Adaptive
LTT [1]	Average	FWER	✗	✗	✗
QLTT [2]	Quantile	FWER	✗	✗	✗
IB-MHT [3]	Mutual Information	FWER	✗	✗	✗
PT [4]	Average	FWER	✓	✗	✗
RG-PT [5]	Average	FDR	✓	✓	✗
aLTT [6]	Average	FWER	✗	✗	✓

II. LTT-Based Hyperparameter Selection

The goal of LTT-based hyperparameter selection methods is to control a risk measure $R(\lambda)$ by choosing a vector of hyperparameters λ. As detailed in Section III, the risk measure is a general functional of the true, unknown, distribution P_Z of the data $Z \sim P_Z$. LTT-based schemes require the following inputs:

1) A discrete set Λ of candidate hyperparameter configurations λ.
2) A held-out calibration data set $\mathcal{Z} = \{Z_i\}_{i=1}^{|\mathcal{Z}|}$, with i.i.d. data points $Z_i \sim P_Z$.

LTT-based methods associate with each hyperparameter $\lambda \in \Lambda$ the null hypothesis $\mathcal{H}_\lambda$ that hyperparameter λ is unreliable. Formally, the null hypothesis $\mathcal{H}_\lambda$ is defined as

$$\mathcal{H}_\lambda : \quad R(\lambda) > \alpha, \tag{1}$$

where α is a user-specified threshold. Rejecting the null hypothesis $\mathcal{H}_\lambda$ is equivalent to deeming λ reliable.

Testing is performed by evaluating, based on the calibration data $\mathcal{Z}$, a valid p-value p_λ, or e-value e_λ [18], for each null hypothesis $\mathcal{H}_\lambda$. P-values and e-values are test statistics that provide measures of evidence in favor or against the null hypothesis. While p-values are commonly used for fixed-sample hypothesis testing, e-values allow for more flexible testing strategies. Notably, they support sequential methods whereby one continuously monitors the testing outcomes, stopping whenever sufficient evidence against the null hypothesis $\mathcal{H}_\lambda$ is received [18].

As shown in Figure 1, using the set of p-values $\{p_\lambda\}_{\lambda\in\Lambda}$ or e-values $\{e_\lambda\}_{\lambda\in\Lambda}$ for the set Λ of candidate hyperparameters, LTT-based methods use an MHT algorithm [19] to test the hypotheses $\{\mathcal{H}_\lambda\}_{\lambda\in\Lambda}$. As detailed in Sec. IV, the output of MHT is a subset $\hat{\Lambda}_{\mathcal{Z}} \subseteq \Lambda$ of hyperparameters with the guarantee that it contains no, or few, unreliable hyperparameters λ violating the condition (1), i.e., $R(\lambda) > \alpha$.

III. Risk Measures

Consider an instantaneous risk function $r_\lambda(Z)$, which quantifies the risk or loss associated with a test instance Z when the model is deployed using hyperparameter λ. To capture the overall risk across the entire data distribution, different *risk measures* can be adopted. This section provides an overview of risk measures studied in the literature.

The standard risk measure used in LTT [1] is the *average* risk

$$R_{\text{avg}}(\lambda) = \mathbb{E}_Z\left[r_\lambda(Z)\right], \tag{2}$$

where the expectation is taken with respect to the random variable $Z \sim P_Z$. With this definition, assuming a bounded risk function $r_\lambda(Z) \in (0, 1]$, a valid p-value for the null hypothesis $\mathcal{H}_\lambda$ in (1) can be calculated by using Hoeffding's inequality [1]

$$p_\lambda^{\text{avg}}(\mathcal{Z}) = e^{-2|\mathcal{Z}|(\alpha - \hat{R}_{\text{avg}}(\lambda|\mathcal{Z}))_+^2}, \tag{3}$$

where

$$\hat{R}_{\text{avg}}(\lambda|\mathcal{Z}) = \frac{1}{|\mathcal{Z}|}\sum_{Z\in\mathcal{Z}} r_\lambda(Z) \tag{4}$$

is the empirical estimate for the average risk $R_{\text{avg}}(\lambda)$ obtained using data set $\mathcal{Z}$.

While the average risk $R_{\text{avg}}(\lambda)$ provides a useful baseline for evaluating performance, it may fail to capture the variability of the risk across different instances. In practice, performance metrics beyond averages are often of interest, including uncertainty-aware measures [20] and information-theoretic measures [21].

Among the uncertainty-aware risk measures, *quantiles* play a central role. For example, in cellular wireless systems, one typically wishes to control the lower quantiles of *key performance indicators* (KPIs), such as throughput and latency, to offer performance guarantees also to cell-edge users [14].

The *q-quantile risk* $R_q(\lambda)$ attained by a hyperparameter λ is defined as

$$R_q(\lambda) = \inf_{r\geq 0}\left\{\Pr_Z[r_\lambda(Z) \leq r] \geq 1 - q\right\}. \tag{5}$$

This measure evaluates the smallest risk r such that at least a fraction $1 - q$ of the risk values $r_\lambda(Z)$ lie below it. Thus, controlling the q-quantile risk, i.e., ensuring the inequality $R_q(\lambda) < \alpha$, guarantees that the worst-case risk among the top-q fraction of test instances is no smaller than α.

QLTT [2] applies the LTT procedure described in Section II to the control of the quantile risk $R_q(\lambda)$ via the use of the p-value in [2]. Figure 2 compares the distribution of KPIs attained in a radio resource allocation problem [22] by LTT and QLTT (see details in [2]). The objective here is to ensure that at least 90% of users experience a packet delay below the 10 ms target, i.e., that the $q = 0.9$-quantile of packet delays remains under this threshold. Since LTT controls the average risk, it fails to meet the quantile requirement, whereas QLTT successfully ensures compliance with the target.

An example of an information-theoretic risk measure is *mutual information* (MI). In [3], MI-based reliability requirements were considered in the context of the information

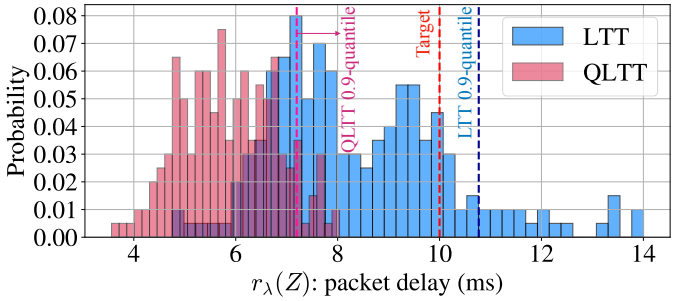

Fig. 2: Packet delay distribution for LTT and QLTT in a resource allocation problem (adapted from [2] by using a different calibration data set $\mathcal{Z}$).

bottleneck problem [23]. In it, the hyperparameter λ of a feature extraction function f_λ that computes features T_λ from covariates X is optimized such that the MI between extracted features T_λ and targets Y, shown with $I(T_\lambda; Y)$, is above a threshold α. [3] calculates a valid p-value for the hypothesis $\mathcal{H}_\lambda : I(T_\lambda; Y) < \alpha$, enabling valid MHT using MI as the risk measure.

IV. Statistical Guarantees

As discussed in Section II, the goal of LTT-based hyperparameter selection methods is to identify a subset, $\hat{\Lambda}_\mathcal{Z}$, of hyperparameters λ in the candidate set Λ that satisfies the statistical guarantees on the risk measure $R(\lambda)$ (see Sec. III). As reviewed in this section, the statistical guarantees can be formalized in different ways.

LTT [1] employs the FWER as the criterion, which ensures that the probability of including *any* unreliable hyperparameter in subset $\hat{\Lambda}_\mathcal{Z}$ remains below a user-specified threshold δ. Formally, the selected subset $\hat{\Lambda}_\mathcal{Z}$ satisfies the guarantee

$$\Pr_\mathcal{Z}\left[R(\lambda) \leq \alpha \text{ for all } \lambda \in \hat{\Lambda}_\mathcal{Z}\right] \geq 1 - \delta, \tag{6}$$

where the probability is evaluated with respect to the data set $\mathcal{Z}$.

As a simple method for controlling the FWER, Bonferroni correction selects the subset

$$\hat{\Lambda}_\mathcal{Z} = \left\{\lambda \in \Lambda : p_\lambda < \frac{\delta}{|\Lambda|}\right\}. \tag{7}$$

However, as the size of the initial candidate set Λ increases, Bonferroni's method becomes overly conservative, leading to fewer selected hyperparameters. To mitigate this limitation, *fixed sequence testing* (FST) [24] can be applied when prior knowledge is available regarding the relative reliability of different hyperparameters.

With FST, the hyperparameters $\lambda \in \Lambda$ are first arranged in a predefined order, denoted as $(\lambda_1, \ldots, \lambda_{|\Lambda|})$, where λ_1 is presumed to be the most reliable and $\lambda_{|\Lambda|}$ the least reliable. Then, FST sequentially evaluates each hyperparameter in descending order of expected reliability, i.e., $\lambda_1, \lambda_2, \ldots, \lambda_{|\Lambda|}$, halting at the first instance where the inequality $p_{\lambda_j} \leq \delta$ is no longer satisfied. The final subset of reliable hyperparameters is then given by $\hat{\Lambda}_\mathcal{Z} = \{\lambda_1, \ldots, \lambda_j\}$.

As shown in [25], due to the duality between confidence intervals and p-values, FST can also be applied if one has access to an upper confidence bound $R^+_\mathcal{Z}(\lambda)$ at level δ on the risk function $R(\lambda)$, satisfying the property

$$\Pr_\mathcal{Z}[R(\lambda) < R^+_\mathcal{Z}(\lambda)] \geq 1 - \delta. \tag{8}$$

In this case, hyperparameter λ is detected as reliable if the inequality $R^+_\mathcal{Z}(\lambda) < \alpha$ holds.

As summarized in Table I, most existing LTT-based approaches rely on FWER control for statistical guarantees. However, FWER control can be excessively conservative, particularly as the number of candidate hyperparameters in set Λ increases. Rather than restricting the probability of including *any* unreliable hyperparameters in the selected subset $\hat{\Lambda}_\mathcal{Z}$, an alternative approach is to control the expected *proportion* of unreliable hyperparameters within $\hat{\Lambda}_\mathcal{Z}$. This criterion, known as *false discovery rate* (FDR) control, ensures that the selected subset $\hat{\Lambda}_\mathcal{Z} \subseteq \Lambda$ satisfies the inequality [26]

$$\mathbb{E}_Z\left[\frac{\sum_{\lambda \in \hat{\Lambda}_\mathcal{Z}} \mathbf{1}\{R(\lambda) > \alpha\}}{\max(|\hat{\Lambda}_\mathcal{Z}|, 1)}\right] \leq \delta, \tag{9}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, and the expectation is taken over the unknown data distribution P_Z. FDR control is adopted by RG-PT [5], which is briefly described in Section VI.

V. Multi-Objective Optimization

In many real-world problems, one is interested in controlling multiple risk measures, say $\{R_l(\lambda)\}_{l=1}^L$, simultaneously. More broadly, the goal is often to address the multi-objective hyperparameter selection problem

$$\begin{aligned} &\min_{\lambda \in \Lambda} \{R_{L_c+1}(\lambda), R_{L_c+2}(\lambda), \ldots, R_L(\lambda)\} \\ &\text{subject to } R_l(\lambda) < \alpha_l \text{ for all } 1 \leq l \leq L_c, \end{aligned} \tag{10}$$

where the first $L_c \leq L$ risk functions are strictly controlled to be below user-defined thresholds $\{\alpha_l\}_{l=1}^{L_c}$.

PT [4] begins by partitioning the calibration dataset $\mathcal{Z}$ into two disjoint subsets $\mathcal{Z}_{\text{OPT}}$ and $\mathcal{Z}_{\text{MHT}}$. The subset $\mathcal{Z}_{\text{MHT}}$ is reserved for the MHT process, while $\mathcal{Z}_{\text{OPT}}$ is used for a preliminary optimization step aimed at estimating the Pareto front for the multi-objective problem (10). Specifically, using the dataset $\mathcal{Z}_{\text{OPT}}$, PT determines the subset $\Lambda_{\text{OPT}} \subseteq \Lambda$ consisting of hyperparameters that form the Pareto front in the space of estimated risk measures $\{\hat{R}_l^{\text{avg}}(\lambda|\mathcal{Z}_{\text{OPT}})\}_{l=1}^L$, calculated for each reliability risk function $R_l(\lambda)$ using (4). Any standard multi-objective optimization algorithm can be employed for this purpose, as discussed in [4]. An illustration of this process is provided in Figure 3(a).

Given the estimated Pareto front, PT forms the new hypotheses

$$\mathcal{H}_\lambda^{\text{multi}} : \text{ there exists } l \in \{1, \ldots, L_c\} \text{ such that } R_l(\lambda) > \alpha_l. \tag{11}$$

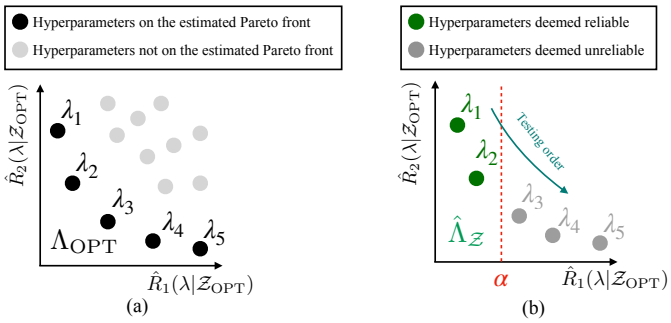

Fig. 3: Illustration of the operation of PT for one reliability risk function $R_1(\lambda)$ to be controlled below the threshold α and one auxiliary risk function $R_2(\lambda)$ to be minimized. (a) Find the subset Λ_{OPT} lying on the estimated Pareto front; (b) Perform MHT via FST to identify the subset $\hat{\Lambda}_{\mathcal{Z}}$ of reliable hyperparameters.

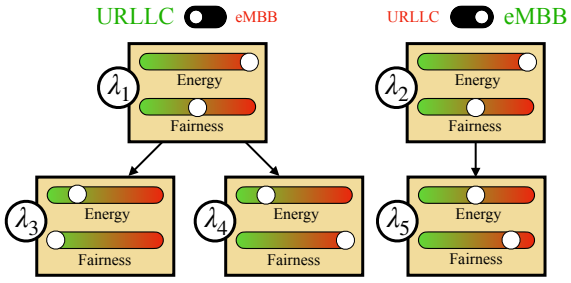

Fig. 4: An example of prior information on the relative reliability of hyperparameters. Hyperparameters corresponding to a higher energy consumption are considered more reliable, and put at a higher level in the RG. Only if these configurations are deemed reliable do we proceed to test hyperparameters with better energy efficiency, which may also improve other performance criteria such as fairness.

As such, rejecting the hypothesis $\mathcal{H}_{\lambda}^{\text{multi}}$ is equivalent to detecting λ as reliable in terms of the constraint in (10). For the null hypothesis (11), PT computes the combined p-value

$$p_{\lambda}(\widetilde{\mathcal{Z}}) = \max_{1 \leq l \leq L_c} p_{\lambda,l}(\widetilde{\mathcal{Z}}), \tag{12}$$

where $\{p_{\lambda,l}(\widetilde{\mathcal{Z}})\}_{l=1}^{L_c}$ are valid p-values for each of the reliability risk functions, and $\widetilde{\mathcal{Z}}$ denotes either $\mathcal{Z}_{\text{OPT}}$ or $\mathcal{Z}_{\text{MHT}}$.

To perform MHT, PT first orders the hyperparameters in Λ_{OPT} based on the p-values $\{p_{\lambda}(\mathcal{Z}_{\text{OPT}})\}_{\lambda \in \Lambda_{\text{OPT}}}$ from low to high to reflect an ordering on their expected reliability, and then uses the p-values $\{p_{\lambda}(\mathcal{Z}_{\text{MHT}})\}_{\lambda \in \Lambda_{\text{OPT}}}$ to perform FST for MHT. This is illustrated in Figure 3(b).

VI. Side Information

In many engineering and real-world applications, we may have access to side information regarding the expected reliability of the hyperparameters in Λ.

As an example of this, consider a telecommunication network with three interacting AI apps: (*i*) a resource allocation app balancing traffic between enhanced Mobile Broadband (eMBB) services and Ultra-Reliable Low-Latency Communications (URLLC) users [27]; (*ii*) a scheduling app enforcing fairness among eMBB users [28]; and (*iii*) an energy management app optimizing a base station's energy efficiency [29]. In this setting, hyperparameters yielding a larger energy consumption are likely to produce more advantageous KPIs in terms of throughput or latency. As depicted in Figure 4, this information can be encoded using a *reliability graph* (RG).

An RG is a *directed acyclic graph* (DAG) that encodes reliability dependencies among the hyperparameters in Λ. In the RG, each hyperparameter in Λ is represented as a node, and a directed edge from λ_1 to λ_2 indicates that λ_1 is expected to be more reliable than λ_2.

RG-PT [5] takes as input pairwise probabilities $0 \leq \eta_{ij} \leq 1$ for each pair of hyperparameters $\lambda_i, \lambda_j \in \Lambda_{\text{OPT}}$. The probability η_{ij} encodes the prior information on how likely it is for hyperparameter λ_i to be more reliable than hyperparameter

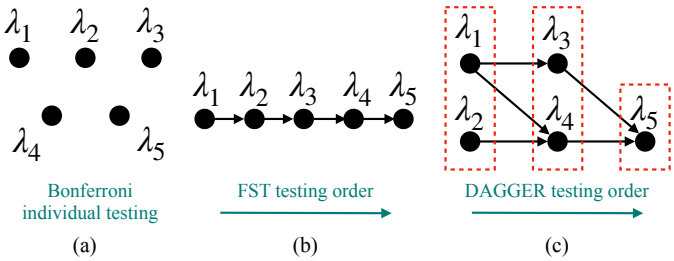

Fig. 5: Comparison of (a) Bonferroni method, (b) FST, and (c) DAGGER testing methods.

λ_j. Additionally, the strength of the prior information is determined via another input n_p, which can be seen as a pseudocount variable as in the standard categorical-Dirichlet model [30]. Using held-out data and this prior knowledge, RG-PT estimates an RG, which is then given as an input to DAGGER [31], an FDR-controlling algorithm for performing MHT on DAGs. As illustrated in Figure 5, DAGGER leverages the RG to guide the MHT procedure, and identifies the subset $\hat{\Lambda}_{\mathcal{Z}}$ of reliable hyperparameters, ensuring compliance with the FDR constraint in (9).

VII. Adaptivity

The hyperparameter selection schemes considered so far are static, in the sense that they operate with a given fixed held-out data set $\mathcal{Z}$. In contrast, adaptive hyperparameter selection schemes can collect and use data dynamically based on accumulated evidence, potentially reducing the required data while maintaining statistical reliability. This is particularly useful when evaluating risk functions is expensive or risky.

Formally, a test is adaptive if (*i*) it operates sequentially, selecting the next subset of hyperparameters to test based on previous observations; and (*ii*) the testing process can conclude as soon as sufficient evidence has been gathered.

aLTT [6] is an adaptive LTT-based hyperparameter selection method that improves upon LTT by leveraging sequential data-dependent testing. Building on [18], aLTT dynamically selects

which hyperparameters to test, and determines when to stop testing. Importantly, aLTT relies on e-values and e-processes, given that p-values are not well suited to support optimal continuation and monitoring [32].

In particular, in aLTT, each null hypothesis $\mathcal{H}_{\lambda_i} : R(\lambda_i) > \alpha$ is tested using the e-value sequence

$$E_{i,t} = \prod_{t' \leq t} (1 + \mu_{t'}(\alpha - r(\lambda_i, t'))), \tag{13}$$

where μ_t is a betting strategy that adapts to new observations [18], and $r(\lambda_i, t)$ represents the observed risk for hyperparameter λ_i at time t. The use of e-values allows for anytime-valid inference, meaning aLTT can stop early without compromising statistical guarantees [6].

VIII. Conclusion

In this paper, we have provided a review of LTT-based hyperparameter selection methods, which leverage MHT to ensure the reliability of selected hyperparameters with provable statistical guarantees. We described the fundamental LTT framework and explored several extensions that address critical aspects of hyperparameter selection in various engineering and real-world applications. These include alternative risk measures, multi-objective optimization, the incorporation of prior information, and adaptive testing mechanisms.

The field of hyperparameter selection via MHT is growing quickly, and interesting directions for research include the development of contextual MHT methods and the extension to semi-supervised testing techniques.

References

[1] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei, "Learn then test: Calibrating predictive algorithms to achieve risk control," *arXiv preprint arXiv:2110.01052*, 2021.

[2] A. Farzaneh, S. Park, and O. Simeone, "Quantile learn-then-test: Quantile-based risk control for hyperparameter optimization," *IEEE Signal Processing Letters*, 2024.

[3] A. Farzaneh and O. Simeone, "Statistically valid information bottleneck via multiple hypothesis testing," *arXiv preprint arXiv:2409.07325*, 2024.

[4] B. Laufer-Goldshtein, A. Fisch, R. Barzilay, and T. S. Jaakkola, "Efficiently controlling multiple risks with Pareto testing," in *Proc. International Conference on Learning Representations*, 2023.

[5] A. Farzaneh and O. Simeone, "Multi-objective hyperparameter selection via hypothesis testing on reliability graphs," *arXiv preprint arXiv:2501.13018*, 2025.

[6] M. Zecchin and O. Simeone, "Adaptive learn-then-test: Statistically valid and efficient hyperparameter selection," *arXiv preprint arXiv:2409.15844*, 2024.

[7] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[8] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[9] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.

[10] V. J. Dzau, M. H. Laitner, A. Temple, and T. H. Nguyen, "Achieving the promise of artificial intelligence in health and medicine: Building a foundation for the future," 2023.

[11] S. Bensalem, C.-H. Cheng, W. Huang, X. Huang, C. Wu, and X. Zhao, "What, indeed, is an achievable provable guarantee for learning-enabled safety-critical systems," in *International Conference on Bridging the Gap between AI and Reality*, pp. 55–76, Springer, 2023.

[12] T. Schuster, A. Fisch, J. Gupta, M. Dehghani, D. Bahri, V. Tran, Y. Tay, and D. Metzler, "Confident adaptive language modeling," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17456–17472, 2022.

[13] A. N. Angelopoulos, A. P. Kohli, S. Bates, M. Jordan, J. Malik, T. Alshaabi, S. Upadhyayula, and Y. Romano, "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging," in *International Conference on Machine Learning*, pp. 717–730, PMLR, 2022.

[14] R. Karasik, O. Simeone, H. Jang, and S. S. Shitz, "Learning to broadcast for ultra-reliable communication with differential quality of service via the conditional value at risk," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 8060–8074, 2022.

[15] Y. Sun, D. W. K. Ng, and R. Schober, "Multi-objective optimization for power efficient full-duplex wireless communication systems," in *2015 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2015.

[16] T. Raviv, S. Park, O. Simeone, Y. C. Eldar, and N. Shlezinger, "Adaptive and flexible model-based ai for deep receivers in dynamic channels," *IEEE Wireless Communications*, 2024.

[17] Q. Hou, S. Park, M. Zecchin, Y. Cai, G. Yu, and O. Simeone, "What if we had used a different app? reliable counterfactual kpi analysis in wireless systems," *IEEE Transactions on Cognitive Communications and Networking*, 2025.

[18] A. Ramdas and R. Wang, "Hypothesis testing with e-values," *arXiv preprint arXiv:2410.23614*, 2024.

[19] J. A. Rice, *Mathematical Statistics and Data Analysis.* Belmont, CA: Duxbury Press., third ed., 2006.

[20] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information fusion*, vol. 76, pp. 243–297, 2021.

[21] B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, "An information-theoretic approach to generalization theory," *arXiv preprint arXiv:2408.13275*, 2024.

[22] A. Valcarce, "Wireless Suite: A collection of problems in wireless telecommunications," *https://github.com/nokia/wireless-suite*, 2020.

[23] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Allerton Conference on Communication, Control and Computation*, 2001.

[24] P. Bauer, "Multiple testing in clinical trials," *Statistics in Medicine*, vol. 10, no. 6, pp. 871–890, 1991.

[25] B.-S. Einbinder, L. Ringel, and Y. Romano, "Semi-supervised risk control via prediction-powered inference," *arXiv preprint arXiv:2412.11174*, 2024.

[26] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[27] M. Elsayed and M. Erol-Kantarci, "AI-Enabled Radio Resource Allocation in 5G for URLLC and eMBB Users," in *2019 IEEE 2nd 5G World Forum (5GWF)*, pp. 590–595, 2019.

[28] X. Han, K. Xiao, R. Liu, X. Liu, G. C. Alexandropoulos, and S. Jin, "Dynamic resource allocation schemes for embb and urllc services in 5g wireless networks," *Intelligent and Converged Networks*, vol. 3, no. 2, pp. 145–160, 2022.

[29] T. Rumeng, W. Tong, S. Ying, and H. Yanpu, "Intelligent energy saving solution of 5g base station based on artificial intelligence technologies," in *2021 IEEE International Joint EMC/SI/PI and EMC Europe Symposium*, pp. 739–742, 2021.

[30] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[31] A. Ramdas, J. Chen, M. J. Wainwright, and M. I. Jordan, "A sequential algorithm for false discovery rate control on directed acyclic graphs," *Biometrika*, vol. 106, no. 1, pp. 69–86, 2019.

[32] Z. Xu and A. Ramdas, "Online multiple testing with e-values," in *International Conference on Artificial Intelligence and Statistics*, pp. 3997–4005, PMLR, 2024.