A SUPPLEMENTARY MATERIAL

Automatic Termination for Hyperparameter Optimization

A.1 PROOF OF PROPOSITION 1

Proposition 1. *Consider the expected loss f and its estimator $\hat{f}$ defined in Eqs. (4) and (5), respectively, and assume the statistical error of the estimator is bounded as $||\hat{f} - f||_\infty \leq \epsilon_{st}$. Let $\gamma^* = \arg\min_{\gamma\in\Gamma} f(\gamma)$, $\gamma^*_{\mathcal{D}} = \arg\min_{\gamma\in\Gamma} \hat{f}(\gamma)$, and $\hat{\gamma}_{\mathcal{D}}$ be a candidate solution to $\min_{\gamma\in\Gamma} \hat{f}(\gamma)$ such that $\hat{f}(\hat{\gamma}_{\mathcal{D}}) - \hat{f}(\gamma^*_{\mathcal{D}}) \leq \epsilon_{BO}$ for some $\epsilon_{BO} \geq 0$. Then the gap in generalization performance $f(\hat{\gamma}_{\mathcal{D}}) - f(\gamma^*)$ between the proposed solution and the true optimizer can be bounded as follows:*

$$f(\hat{\gamma}_{\mathcal{D}}) - f(\gamma^*) \leq \underbrace{f(\hat{\gamma}_{\mathcal{D}}) - \hat{f}(\hat{\gamma}_{\mathcal{D}})}_{\leq \epsilon_{st}} + \underbrace{\hat{f}(\hat{\gamma}_{\mathcal{D}}) - \hat{f}(\gamma^*_{\mathcal{D}})}_{\leq \epsilon_{BO}} + \underbrace{\hat{f}(\gamma^*_{\mathcal{D}}) - \hat{f}(\gamma^*)}_{\leq 0} + \underbrace{\hat{f}(\gamma^*) - f(\gamma^*)}_{\leq \epsilon_{st}}$$
$$\leq 2\epsilon_{st} + \epsilon_{BO}.$$

Moreover, without further restrictions on f, $\hat{f}$, $\hat{\gamma}_{\mathcal{D}}$ and γ^*, this upper bound is tight. (Proof in

Proof: While the second inequality is due to the definition of $\hat{\gamma}_{\mathcal{D}}$, the others can be proved as follows:

$$f(\hat{\gamma}_{\mathcal{D}}) - \hat{f}(\hat{\gamma}_{\mathcal{D}}) \leq |f(\hat{\gamma}_{\mathcal{D}}) - \hat{f}(\hat{\gamma}_{\mathcal{D}})| \leq \max_{\gamma\in\Gamma} |f(\gamma) - \hat{f}(\gamma)| = ||\hat{f} - f||_\infty \leq \epsilon_{st},$$

$$\gamma^*_{\mathcal{D}} = \arg\min_{\gamma\in\Gamma} \hat{f}(\gamma) \longrightarrow \forall\gamma \in \Gamma : \hat{f}(\gamma^*_{\mathcal{D}}) - \hat{f}(\gamma) \leq 0 \longrightarrow \hat{f}(\gamma^*_{\mathcal{D}}) - \hat{f}(\gamma^*) \leq 0. \quad \blacksquare$$

A.2 EXPERIMENTS SETTING

A.2.1 BO SETTING

We used an internal BO implementation where expected improvement (EI) together with Mat'ern-52 kernel in the GP are used. The hyperparameters of the GP includes output noise, a scalar mean value, bandwidths for every input dimension, 2 input warping parameters and a scalar covariance scale parameter. The closest open-source implementations are GPyOpt using input warped GP [1] or AutoGluon BayesOpt searcher [2]. We maximize type II likelihood to learn the GP hyperparameters in our experiments.

A.2.2 ALGORITHM

A.2.3 SEARCH SPACES FOR CROSS VALIDATION EXPERIMENTS

XGBoost (XGB) and RandomForest (RF) are based on scikit-learn implementations and their search spaces are listed in Table 1.

A.2.4 DATASETS IN CROSS VALIDATION EXPERIMENTS

We list the datasets that are used in our experiments, as well as their characteristics and sources in Table 2. For each dataset, we first randomly draw 20% as test set and for the rest, we use 10-fold cross validations for regression datasets and 10-fold stratified cross validation for classification datasets. The actual data splits depend on the seed controlled in our experiments. For a given experiment, all the hyperparameters trainings use the same data splits for the whole tuning problem. For the experiments without cross-validation, we use 20% dataset as validation set and the rest as training set.

A.3 DETAILED RESULTS

We first show the scatter plots of RTC and RYC scores for different automatic termination methods on HPO-Bench-datasets in Fig. 6 and the results on NAS-Bench-201 in Fig. 7.

[1] https://github.com/SheffieldML/GPyOpt

[2] https://github.com/awslabs/autogluon

Algorithm 1 BO for HPO with cross-validation and automatic termination

Require: Model $\mathcal{M}_\gamma$ parametrized by $\gamma \in \Gamma$, data $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$ for k-fold cross-validation, acquisition function $\alpha(\gamma)$

Initialize $y_t^* = +\infty$ and $G_t = \{\}$
for $t = 1, 2, \ldots$ **do**
  Sample $\gamma_t \in \arg\max_{\gamma \in \Gamma} \alpha(\gamma)$
  for $i = 1, 2, \ldots, k$ **do**
    Fit the model $\mathcal{M}_\gamma(\cdot; \mathcal{D}_{-i})$, where $\mathcal{D}_{-i} = \cup_{j \neq i} \mathcal{D}_i$
    Evaluate the fitted model $y_t^i = \frac{1}{|\mathcal{D}_i|} \sum_{\mathrm{x}_i, \mathrm{y}_i \in \mathcal{D}_i} \ell(\mathrm{y}_i, \mathcal{M}_\gamma(\mathrm{x}_i, \mathcal{D}_{-i}))$
  end for
  Calculate the sample mean $y_t = \frac{1}{k} \sum_k y_t^i$,
  if $y_t \leq y_t^*$ **then**
    Update $y_t^* = y_t$ and the current best $\gamma_t^* = \gamma_t$
    Calculate the sample variance $s_{\mathrm{cv}}^2 = \frac{1}{k} \sum_i (y_t - y_t^i)^2$
    Calculate the variance estimate $\mathrm{Var}\hat{f}(\gamma_t^*) \approx \left(\frac{1}{k} + \frac{|\mathcal{D}_i|}{|\mathcal{D}_{-i}|}\right) s_{\mathrm{cv}}^2$ from Eq. (8)
  end if
  Update $G_t = G_{t-1} \cup \gamma_t$ and $y_{1:t} = y_{1:t-1} \cup y_t$
  Update σ_t, μ_t with Eqs. (1) and (2)
  Calculate upper bound $\bar{r}_t := \min_{\gamma \in G_t} \mathrm{ucb}_t(\gamma) - \min_{\gamma \in \Gamma} \mathrm{lcb}_t(\gamma)$ for simple regret from Eq. (7)
  if the condition $\bar{r}_t \leq \sqrt{\mathrm{Var}\hat{f}(\gamma_t^*)}$ holds **then**
    terminate BO loop
  end if
end for
Output: γ_t^*

Table 1: Search spaces description for each algorithm.

tasks	hyperparameter	search space	scale
XGBoost	n_estimators	$[2, 2^9]$	log
	learning_rate	$[10^{-6}, 1]$	log
	gamma	$[10^{-6}, 2^6]$	log
	min_child_weight	$[10^{-6}, 2^5]$	log
	max_depth	$[2, 2^5]$	log
	subsample	$[0.5, 1]$	linear
	colsample_bytree	$[0.3, 1]$	linear
	reg_lambda	$[10^{-6}, 2]$	log
	reg_alpha	$[10^{-6}, 2]$	log
RandomForest	n_estimators	$[1, 2^8]$	log
	min_samples_split	$[0.01, 0.5]$	log
	max_depth	$[1, 5]$	log

A.3.1 Detailed numbers of RYC and RTC scores

We report detailed RYC scores and RTC scores of different HPO automatic termination methods for the experiments in the main text in Table 3, Table 4 and Table 5.

A.3.2 Correlation between validation and test metrics

In Fig. 8, we show the correlation between validation and test metrics of hyperparameters when tuning XGB and RF on tst-census dataset in Fig. 8

dataset	problem_type	n_rows	n_cols	n_classes	source
openml14	classification	1999	76	10	openml
openml20	classification	1999	240	10	openml
tst-hate-crimes	classification	2024	43	63	data.gov
openml-9910	classification	3751	1776	2	openml
farmads	classification	4142	4	2	uci
openml-3892	classification	4229	1617	2	openml
sylvine	classification	5124	21	2	openml
op100-9952	classification	5404	5	2	openml
openml28	classification	5619	64	10	openml
philippine	classification	5832	309	2	data.gov
fabert	classification	8237	801	2	openml
openml32	classification	10991	16	10	openml
openml34538	regression	1744	43	-	openml
tst-census	regression	2000	44	-	data.gov
openml405	regression	4449	202	-	openml
tmdb-movie-metadata	regression	4809	22	-	kaggle
openml503	regression	6573	14	-	openml
openml558	regression	8191	32	-	openml
openml308	regression	8191	32	-	openml

Table 2: Datasets used in our experiments including their characteristics and sources.

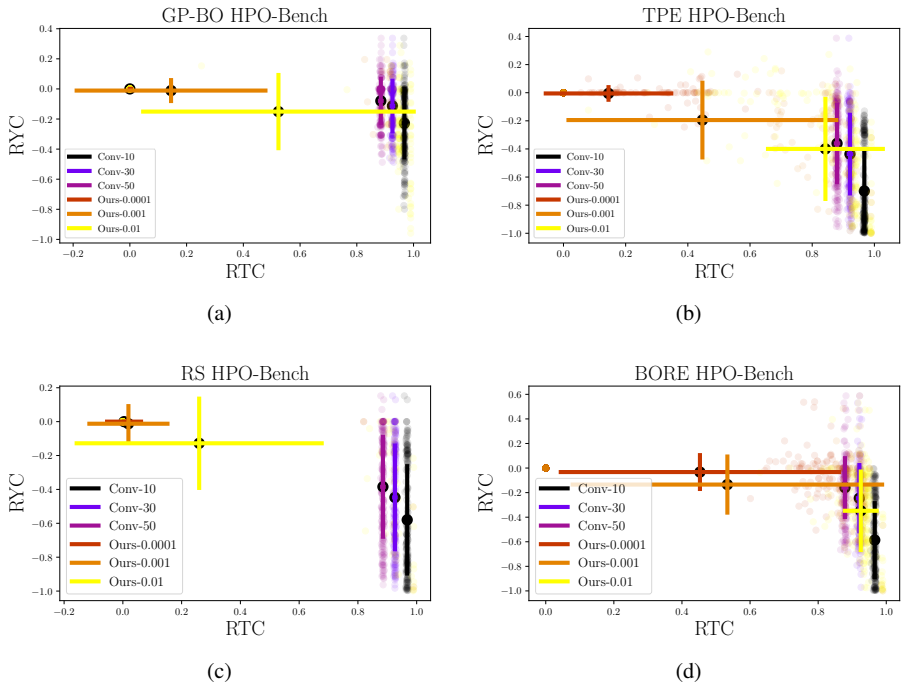


Figure 6: Fig. (a) - (d), the mean and standard deviation of RYC and RTC scores for considered automatic termination methods on HPO-Bench datasets using GP based BO (GP-BO), Random Search (RS), TPE and BORE optimizers. The mean value is shown in the big dot and the standard deviation is shown as error bar in both dimensions.

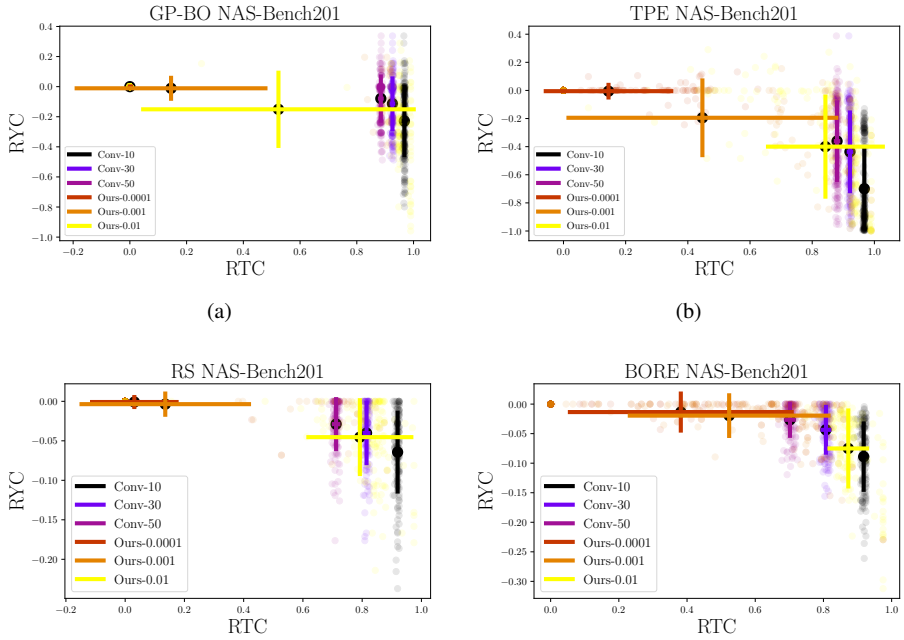


Figure 7: Fig. (a) - (d), the mean and standard deviation of RYC and RTC scores for considered automatic termination methods on NAS-Bench-201 datasets using GP based BO (GP-BO), Random Search (RS), TPE and BORE optimizers. The mean value is shown in the big dot and the standard deviation is shown as error bar in both dimensions.

	RTC		RYC	
algo	RF	XGB	RF	XGB
Conv_10	0.840	0.841	-0.031	-0.051
Conv_30	0.686	0.666	-0.022	-0.026
Conv_50	0.498	0.504	-0.015	-0.021
EI_1e-08	0.896	0.850	-0.057	-0.052
EI_1e-12	0.895	0.779	-0.055	-0.047
EI_1e-16	0.893	0.718	-0.052	-0.045
PI_0.0001	0.898	0.875	-0.059	-0.059
PI_1e-08	0.895	0.814	-0.055	-0.052
PI_1e-12	0.894	0.739	-0.055	-0.044
Ours_0.21	0.318	0.144	-0.004	-0.003
Ours_0.5	0.580	0.224	-0.013	-0.006

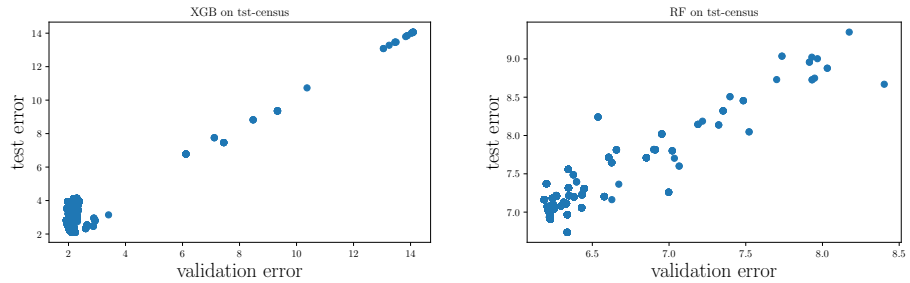
Table 3: RTC and RYC scores for early stopping methods in cross validation benchmarks.

	RTC				RYC			
dataset	naval	parkinsons	protein	slice	naval	parkinsons	protein	slice
Conv_10	0.943	0.947	0.946	0.942	-0.605	-0.582	-0.117	-0.432
Conv_30	0.826	0.837	0.837	0.840	-0.064	-0.235	-0.021	-0.119
Conv_50	0.748	0.729	0.734	0.747	-0.038	-0.107	-0.008	-0.058
Ours_0.0001	0.790	0.018	0.198	0.822	-0.041	-0.012	-0.005	-0.072
Ours_0.001	0.910	0.038	0.271	0.934	-0.220	-0.031	-0.018	-0.281
Ours_0.01	0.941	0.901	0.906	0.953	-0.498	-0.378	-0.071	-0.466

Table 4: RTC and RYC scores for early stopping methods in HPO-Bench.

	RTC			RYC		
dataset	ImageNet	cifar10	cifar100	ImageNet	cifar10	cifar100
Conv_10	0.880	0.889	0.888	-0.034	-0.098	-0.097
Conv_30	0.612	0.611	0.606	-0.010	-0.019	-0.036
Conv_50	0.372	0.361	0.372	-0.004	-0.006	-0.014
Ours_0.0001	0.274	0.311	0.519	-0.002	-0.008	-0.026
Ours_0.001	0.377	0.622	0.582	-0.005	-0.023	-0.033
Ours_0.01	0.837	0.902	0.879	-0.022	-0.106	-0.099

Table 5: RTC and RYC scores for early stopping methods in NAS-Bench-201.



(a) Training XGB on test-census dataset

(b) Training RF on test-census dataset.

Figure 8: We show validation error for training XGB (a) and RF (b) on `tst-census` dataset on the x-axis and test error on the y-axis. In the *low* error region, the validation metrics are not well correlated with the test metrics.