ChaCha for Online AutoML (Supplementary)

A. Evaluation details

A.1. Datasets

The datasets are obtained from OpenML according to the following criterion: (1) the number of instances in the dataset is larger than 10K; (2) no missing value; (3) regression dataset; (4) the dataset is still active and downloadable. The final list of datasets that satisfy the aforementioned criterion is [573, 1201, 1195, 344, 1192, 201, 216, 41065, 42731, 4545, 42688, 1196, 23515, 1206, 1193, 42721, 42571, 42713, 537, 42724, 41540, 4549, 296, 574, 218, 5648, 215, 41539, 1199, 1203, 1191, 564, 1208, 42183, 42225, 42728, 42705, 42729, 42496, 41506]. We converted the original OpenML dataset into VW required format[5] following the instructions. We sequentially group the raw features of each dataset into up to 10 namespaces[6]. Log-transformation on the target variable is performed if the largest value on the target variable is larger than 100 (for datasets whose target variable has negative values, we first shift the value to make it all positive and then do the log-transformation).

A.2. Detailed settings of ChaCha and baselines

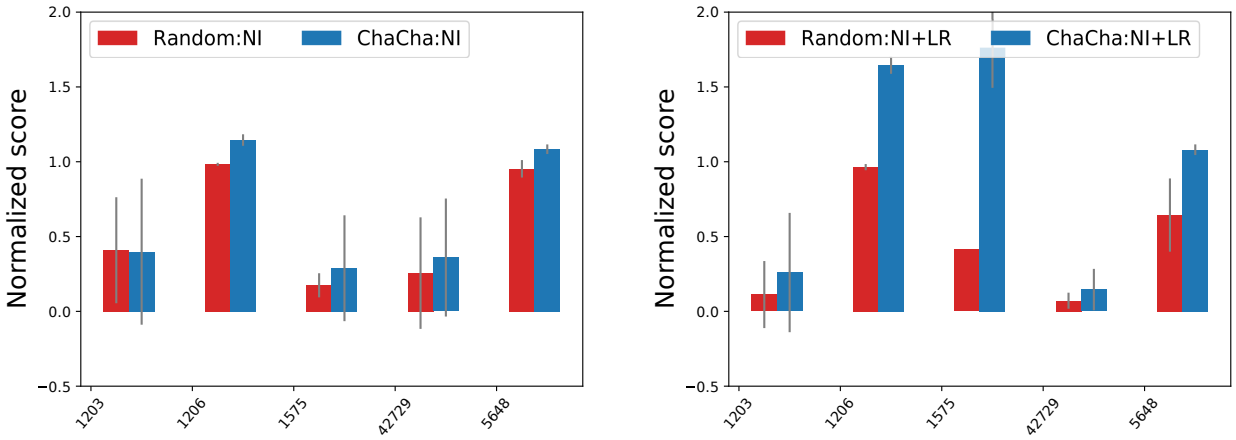
The two compared methods `Random` and `Exhaustive` use the same method as `ChaCha` when selecting one model from the 'live' model pool to make the final prediction at each iteration. The minimum resource lease in `ChaCha` is set to be $5\times$ (dimensionality of the raw features) in all of our experimental evaluations. To ensure the empirical loss of the online learning models is bounded, we use the 'clipped mean absolute error' as the empirical performance proxy in `ChaCha`: we keep track of the minimum value, denoted by $\underline{y}_t$, and the maximum value, denoted by $\bar{y}_t$, of the target variable according to observations received up to time t. When calculating the mean absolute error for models in `ChaCha`, we map our prediction of the target variable $\hat{y}_t$ into this range in the following way: $\min\{\max\{\underline{y}_t, \hat{y}_t\}, \bar{y}_t\}$. By doing so, we ensure the mean absolute error is always bounded by $\bar{y}_t - \underline{y}_t$. Note that this revision is only performed in the update of empirical loss proxy in `ChaCha`, the final output of `ChaCha` is not clipped. For $\epsilon_{c,t}$, we use sample complexity bounds for linear functions. More specifically, $\epsilon_{c,t}$ is set to be $a\sqrt{\frac{d_c \log(|D_{c,t}||S_{m_t}|/\delta)}{|D_{c,t}|}}$, in which d_c is the dimensionality of the feature induced by namespace configuration c, a is constant related to the bound of the loss and is set to be $0.05 * (\bar{y}_t - \underline{y}_t)$, and δ is set as 0.1.

A.3. Additional Results

We now provide additional results for the cases where all the methods are run for a larger number of data samples. We compare the results on the three largest datasets with up to 1M data samples in Figure 6. The result shows the consistent advantage of `ChaCha` under large data volumes. In addition, since several bars in Figure 4 and Figure 5 are cropped, we include the actual numbers of the normalized scores in Table 1 and Table 2 for completeness.

Despite the i.i.d assumption in theoretical analysis, we do not exclude the possibility of non-stationary environments in our empirical evaluation (we intentionally do not shuffle the dataset such that potential concept drifts in the original datasets are preserved). In Figure 7, we show the results on an example dataset where concept drift exists. The results indicate a clear existence of concept drift, and `ChaCha` is still maintain its performance advantage. This appealing property is partly because of the base online learning algorithm's capability to adjust to the concept drifts and partly because of `ChaCha`'s progressive way of Champion promotion with the help of the `ConfigOracle`.

[5] https://github.com/VowpalWabbit/vowpal_wabbit/wiki/Input-format
[6] https://github.com/VowpalWabbit/vowpal_wabbit/wiki/Namespaces



(a) Namespace interactions tuning

(b) Namespace interactions and learning rate tuning

Figure 6. Normalized scores after a larger number of data samples (up to 1M). The normalized scores of `ChaCha` on dataset 1206 for both tuning scenarios, and dataset 1575 and 42729 for namespace interactions tuning are better than those reported in Figure 4, which indicates that `ChaCha` may achieve even larger gain as the increase of data samples.

B. Proof details

Proof 1 (Proof of Lemma 1) .

According to the definition of $\epsilon_{c,t}$,

(1) $\forall m \in [M], c \in \mathcal{S}_m$, *with probability at least* $1-\delta$,

$$L^*_{\mathcal{F}_c} - L^*_{\mathcal{F}_{C_m}} + 2\epsilon_{c,t} + 3\epsilon_{C_m,t} \geq L^{PV}_{c,t} + \epsilon_{c,t} - L^{PV}_{C_m,t} + 2\epsilon_{C_m,t} = \overline{L}_{c,t} - \underline{L}_{C_m,t} + \epsilon_{C_m,t} \tag{5}$$

The above inequality indicates that as long as $L^*_{\mathcal{F}_c} - L^*_{\mathcal{F}_{C_m}} + 2\epsilon_{c,t} + 3\epsilon_{C_m,t} < 0$, $\overline{L}_{c,t} - \underline{L}_{C_m,t} + 2\epsilon_{C_m,t} < 0$ *which means that* c *can pass the* `Better` *test when compared with* C_m *at time* t *and concludes the proof for Claim 1.*

(2) For Claim 2.

When the `Better` *test is triggered at time* $t = t_{m+1}$, *we have, with probability at least* $1-\delta$,

$$L^*_{\mathcal{F}_{C_m}} - L^*_{\mathcal{F}_{C_{m+1}}} \geq \underline{L}_{C_m,t} - \overline{L}_{C_{m+1},t} > \underline{L}_{C_m,t} - (\underline{L}_{C_m,t} - \epsilon_{C_m,t}) = \epsilon_{C_m,t} \tag{6}$$

in which the second inequality is guaranteed by the fact that `Better` *test is positive.*

(3) For Claim 3. $\underline{L}_{c,t} - \overline{L}_{C_m,t} \leq L^*_{\mathcal{F}_c} - L^*_{\mathcal{F}_{C_m}}$ *holds with probability at least* $1-\delta'$. *If* $L^*_{\mathcal{F}_c} < L^*_{\mathcal{F}_{C_m}} < 0$, *then with probability at least* $1-\delta$, $\underline{L}_{c,t} - \overline{L}_{C_m,t} < 0$ *(not passing the* `Worse` *test).*

Proof 2 (Proof of Proposition 1) *Without affecting the order of the cumulative regret (w.r.t.* T*), we prove the regret bound assuming* $c^* \in \mathcal{S}_0$ *(for the case* c^* *is added at a particular time point* t'*, we only need to add an additional constant regret term related to* t'*).*

According to Claim 1 of Lemma 1, during a particular phase m, *i.e.* $t_m \leq t \leq t_{m+1}-1$, $\forall \mathcal{F} \in \mathcal{S}_m$, $L^*_{\mathcal{F}_{C_m}} - L^*_{\mathcal{F}_c} \leq 2\epsilon_{c,t} + 3\epsilon_{C_m,t}$ *must hold, otherwise phase* m *would have ended. Since* $c^* \in \mathcal{S}_m$, *we have,* $\forall m \in [M]$, $\sum_{t=t_m}^{t_{m+1}-1}(L^*_{\mathcal{F}_{C_m}} - L^*_{\mathcal{F}_{c^*}}) < \sum_{t=t_m}^{t_{m+1}-1}(2\epsilon_{c^*,t} + 3\epsilon_{C_m,t})$.

To account for the union over phases, we replace δ *in* $\epsilon_{c,t}$ *from Eq. (1) by* $\delta' := \delta/M$, *and replace* $|\mathcal{S}_t|$ *by* $\max_{m\in M}|\mathcal{S}_m|$ *i.e.,* $\epsilon_{c,t} = \mathtt{comp}_{\mathcal{F}_c}\log(\max_{m\in[M]}\frac{M|D_{t,c}||\mathcal{S}_m|}{\delta})|D_{t,c}|^{p-1})$. *By union bound we have the following inequality holds with*

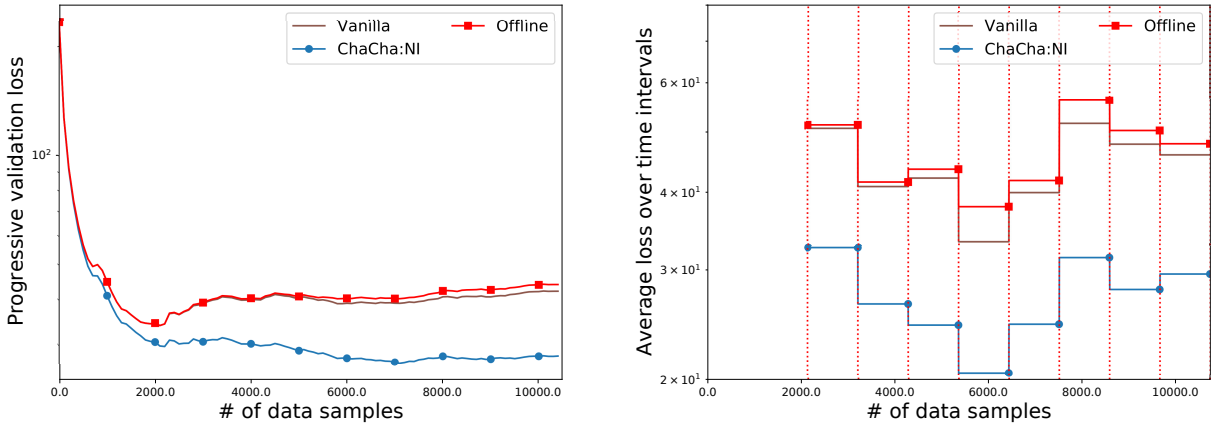


Figure 7. The existence of concept drifts on dataset #42183. The method 'Offline' uses the first 20% of the data to do training and is then tested without further model updates. In the second figure, for each of the methods, we report the average loss over each time intervals, split by vertical lines, starting from 20% data samples. It helps illustrate a clear existence of concept drift in this dataset and the behavior of the compared methods in such a non-stationary environment.

probability at least $1-\delta$,

$$\sum_{m=0}^{M}\sum_{t=t_m}^{t_{m+1}-1}(L^*_{\mathcal{F}_{\mathrm{C}_m}}-L^*_{\mathcal{F}_{c^*}})\le\sum_{m=0}^{M}\sum_{t=t_m}^{t_{m+1}-1}2\epsilon_{c^*,t}+\sum_{m=0}^{M}\sum_{t=t_m}^{t_{m+1}-1}3\epsilon_{\mathrm{C}_m,t}=\sum_{t=1}^{T}2\epsilon_{c^*,t}+\sum_{m=0}^{M}\sum_{t=t_m}^{t_{m+1}-1}3\epsilon_{\mathrm{C}_m,t} \tag{7}$$

The successive doubling resource allocation strategy ensures $\sum_{t=1}^{T}2\epsilon_{c^*,t}=O\big(comp_{\mathcal{F}^*}\max_{m\in[M]}\frac{|\mathcal{S}_m|}{b}T^p\log(\frac{TM|\mathcal{S}_m|}{\delta})\big)$.
Since we always keep the champion of each phase 'live', $\max_{t_m<t<t_{m+t}}|D_{t,\mathrm{C}_m}|\ge t_{m+1}-t_m=N_m$. *Thus we have,*

$$\begin{aligned}\sum_{t=t_m}^{t_{m+1}-1}3\epsilon_{\mathrm{C}_m,t}&\le 3comp_{\mathcal{F}_{\mathrm{C}_m}}\sum_{t=t_m}^{t_{m+1}-1}|D_{t,\mathrm{C}_m}|^{p-1}\log T \\ &=O\big(comp_{\mathcal{F}_{\mathrm{C}_m}}N_m^p\log(\frac{\max_{m\in[M]}TM|\mathcal{S}_m|}{\delta})\big)\\ &=O\big(comp_{\mathcal{F}_{\mathrm{C}_m}}N_m^p\log T+comp_{\mathcal{F}_{\mathrm{C}_m}}N_m^p\log(\max_{m\in[M]}|\mathcal{S}_m|)\big)\end{aligned} \tag{8}$$

Now we provide an upper bound on the value of $\sum_{m=1}^{M}N_m^p$.

By Claim 2 of Lemma 1,

$$L^*_{\mathcal{F}_{\mathrm{C}_0}}-L^*_{\mathcal{F}_{c^*}}\ge=\sum_{m=0}^{M-1}L^*_{\mathcal{F}_{\mathrm{C}_m}}-L^*_{\mathcal{F}_{\mathrm{C}_{m+1}}}>\sum_{m=0}^{M-1}\epsilon_{\mathrm{C}_m,t_{m+1}}>comp_{\mathcal{F}_{\mathrm{C}_m}}\sum_{m=0}^{M-1}N_m^{p-1}\log(\frac{N_m}{\delta}) \tag{9}$$

Now we discuss the properties of $\{N_m\}_{m\in[M-1]}$. *Since* $\sum_{m=0}^{M-1}N_m^{p-1}\log(\frac{N_m}{\delta})$ *converges,* $\sum_{m=0}^{M-1}\frac{1}{N_m^{1-p}}=\sum_{m=0}^{M-1}N_m^{p-1}$ *converges. Since* $1-p<1$, *we have* $N_m\ge\Omega(m^{\frac{1}{1-p}})$, *otherwise* $\sum_{m=0}^{M-1}\frac{1}{N_m^{1-p}}$ *diverges according to convergence properties of Hyperharmonic series.*

Since $T=\sum_{m=0}^{M}N_m$, *we have* $T>\sum_{m=0}^{M-1}N_m>\sum_{m=0}^{M-1}\Omega(m^{\frac{1}{1-p}})>\Omega(M^{\frac{2-p}{1-p}})$, *which indicates* $M<O(T^{\frac{1-p}{2-p}})$.

$$\sum_{m=0}^{M}N_m^p\le(M+1)^{1-p}T^p=O(T^{\frac{1}{2-p}}) \tag{10}$$

Table 1. Normalized scores (mean ± standard deviation) reported in Figure 4(b) and Figure 5.

Dataset id	Random:NI	ChaCha:NI	ChaCha-w/o-Champion:NI	ChaCha-AggressiveScheduling:NI
1191	0.59 ± 0.35	**0.64** ± 0.22	-335.79 ± 44.34	0.00 ± 0.00
1199	**0.72** ± 0.53	0.10 ± 0.10	-39.37 ± 0.33	0.00 ± 0.00
1203	**0.40** ± 0.35	0.38 ± 0.46	-0.35 ± 0.01	0.00 ± 0.01
1206	0.98 ± 0.03	**1.02** ± 0.09	-2.38 ± 0.04	0.00 ± 0.00
1575	0.15 ± 0.06	0.13 ± 0.08	**1.02** ± 0.05	1.00 ± 0.08
201	**0.12** ± 0.15	0.11 ± 0.13	-6.15 ± 0.02	-0.00 ± 0.00
215	0.38 ± 0.37	**0.41** ± 0.34	-0.30 ± 0.04	0.01 ± 0.02
23515	0.17 ± 0.15	**2.94** ± 0.27	-17.33 ± 0.03	1.04 ± 0.01
344	**0.34** ± 0.25	0.34 ± 0.30	-7.18 ± 1.26	0.00 ± 0.00
41506	0.71 ± 0.18	**0.80** ± 0.18	-6.53 ± 0.73	0.10 ± 0.09
42183	0.11 ± 0.21	**2.04** ± 0.08	0.16 ± 0.05	0.05 ± 0.01
42496	**0.19** ± 0.41	-0.04 ± 0.02	-28.09 ± 2.68	0.00 ± 0.00
42729	**0.46** ± 0.31	0.43 ± 0.25	-54.10 ± 0.04	-0.00 ± 0.00
5648	0.97 ± 0.05	**1.09** ± 0.04	-0.01 ± 0.19	1.06 ± 0.04
564	0.19 ± 0.21	**0.74** ± 0.41	-213.08 ± 14.48	0.00 ± 0.00

Table 2. Normalized scores (mean ± standard deviation) reported in Figure 4(b).

Dataset id	Random:NI+LR	ChaCha:NI+LR
1191	**1.00** ± 0.00	1.00 ± 0.00
1199	1.00 ± 0.00	**1.01** ± 0.03
1203	**0.11** ± 0.22	-0.00 ± 0.01
1206	0.95 ± 0.06	**1.25** ± 0.03
1575	0.62 ± 0.00	**2.08** ± 0.54
201	**1.00** ± 0.00	**1.00** ± 0.00
215	0.09 ± 0.19	**0.29** ± 0.39
23515	1.00 ± 0.00	**1.06** ± 0.17
344	1.00 ± 0.00	**1.15** ± 0.08
41506	0.86 ± 0.00	**0.95** ± 0.10
42183	0.08 ± 0.00	**1.86** ± 0.40
42496	0.24 ± 0.00	**0.39** ± 0.30
42729	**0.27** ± 0.10	**0.27** ± 0.10
5648	0.56 ± 0.31	**1.11** ± 0.03
564	**1.00** ± 0.00	**1.00** ± 0.00

in which the second inequality is based on Jensen's inequality. Substituting Eq. (10) *into Eq.* (8) *and Eq.* (7) *concludes the proof.*

Proof 3 (Proof of Theorem 1)

$$\sum_{t=1}^{T}(L_{\hat{c}_t,t} - L^*_{\mathcal{F}_{c^*}}) \sum_{m=0}^{M} \sum_{t=t_m}^{t_{m+1}-1} (L_{\hat{c}_t,t} - L^*_{\mathrm{C}_m}) + \sum_{m=0}^{M} \sum_{t=t_m}^{t_{m+1}-1} (L^*_{\mathrm{C}_m} - L^*_{\mathcal{F}_{c^*}}) \tag{11}$$

The second term in the right-hand side of Eq. (11) *can be upper bounded by Proposition 1. Now we upper bound the first term of the right-hand side of Eq.* (11)*.* $\forall m \in \{0, 1, \cdots, M\}$,

$$\begin{aligned}\sum_{t=t_m}^{t_{m+1}-1}(L_{\hat{c}_t,t}-L^*_{\mathrm{C}_m}) &\leq \sum_{t=t_m}^{t_{m+1}-1}\epsilon_{\hat{c}_t,t}+L^{PV}_{\hat{c}_t,t}-L^{PV}_{\mathrm{C}_m,t}+\epsilon_{\mathrm{C}_m,t} \\ &\leq \sum_{t=t_m}^{t_{m+1}-1}\epsilon_{\mathrm{C}_m,t}+L^{PV}_{\mathrm{C}_m,t}-L^{PV}_{\mathrm{C}_m,t}+\epsilon_{\mathrm{C}_m,t} \\ &= \sum_{t=t_m}^{t_{m+1}-1}2\epsilon_{\mathrm{C}_m,t}=O(\texttt{comp}_{\mathcal{F}_{\mathrm{C}_m}}N_m^p\log T)\end{aligned} \tag{12}$$

in which the last inequality is based on the fact that $\hat{c}_t=\arg\min_{c\in\mathcal{B}_t}(L^{PV}_{c,t}+\epsilon_{c,t})$. *Substituting conclusion in Eq* (10) *which is proved in Proposition* (1)*, into the above inequality, we have*

$$\sum_{m=0}^{M}\sum_{t=t_m}^{t_{m+1}-1}(L_{\hat{c}_t,t}-L^*_{\mathrm{C}_m})=O(\max_{m\in[M]}\texttt{comp}_{\mathcal{F}_{\mathrm{C}_m}}T^{\frac{1}{2-p}}\log T) \tag{13}$$

Substituting Eq. (13) *and the conclusion in Proposition 1 into Eq.* (11) *finishes the proof.*