7 Appendix

7.1 Notations

In order to facilitate the reading of the paper, we provide below (see Table 2) the list of notations.

Table 2: Notations

notation	description
K	number of actions
M	number of contextual variables
D_θ	maximum depth of the tree θ
L	number of trees
T	time horizon
$\mathcal{A}$	set of actions
$\mathcal{V}$	set of variables
$\mathcal{S}$	set of remaining variables
$\mathbf{x}$	context vector $\mathbf{x} = (x_1, \ldots, x_M)$
$\mathbf{y}$	reward vector $\mathbf{y} = (y_1, \ldots, y_K)$
k_t	action chosen at time t
c_θ	context path the tree θ, $c_\theta = (x_{i_1}, v_{i_1}), ..., (x_{i_{d_\theta}}, v_{i_{d_\theta}})$
d_θ	current depth of the context path c_θ
μ_k	expected reward of action k, $\mu_k = \mathbb{E}_{D_y}[y_k]$
$\mu_k^i \vert v$	expected reward of action k conditioned to $x_i = v$, $\mu_k^i \vert v = \mathbb{E}_{D_y}[y_k \cdot \mathbb{1}_{x_i = v}]$
$\mu_{k,v}^i$	expected reward of action k and $x_i = v$, $\mu_{k,v}^i = \mathbb{E}_{D_{x,y}}[y_k \cdot \mathbb{1}_{x_i = v}]$
μ^i	expected reward for the use of the variable x_i to select the best actions
δ	probability of error
ϵ	approximation error
Δ_1	minimum of difference with the expected reward of the best action μ_{k^*} and the expected reward of a given action k: $\Delta_1 = \min_{k \neq k^*}(\mu_{k^*} - \mu_k)$
Δ_2	minimum of difference with the best variable expected reward μ^{i^*} and the expected reward for other variables: $\Delta_2 = \min_{i \neq i^*}(\mu^{i^*} - \mu^i)$
t^*	sample complexity of the decision stump

7.2 Lemma 1

Proof. We cannot use directly Hoeffding inequality (see Hoeffding (1963)) to bound the estimated gains of the use of variables. The proof of Lemma 1 overcomes this difficulty by bounding each estimated gain μ^i by the sum of the bounds over the values of the expected reward of the best action $\mu_{k^*,v}^i$ (inequality 7). From Hoeffding's inequality, at time t we have:

$$\mathbb{P}\left(\left|\hat{\mu}_{k,v}^i - \mu_{k,v}^i\right| \geq \alpha_{t_k}\right) \leq 2\exp(-2\alpha_{t_k}^2 t_k) = \frac{\delta}{2KMt_k^2} \, ,$$

where $\alpha_{t_k} = \sqrt{\frac{1}{2t_k}\log\frac{4KMt_k^2}{\delta}}$.

Using Hoeffding's inequality on each time t_k, applying the union bound and then $\sum 1/t_k^2 = \pi^2/6$, the following inequality holds for any time t with a probability $1 - \frac{\delta\pi^2}{12KM}$:

$$\hat{\mu}_{k,v}^i - \alpha_{t_k} \leq \mu_{k,v}^i \leq \hat{\mu}_{k,v}^i + \alpha_{t_k} \tag{5}$$

If the inequality (5) holds for the actions $k' = \arg\max_k \hat{\mu}^i_{k,v}$, and $k^* = \arg\max_k \mu^i_{k,v}$, we have:

$$\begin{aligned} &\hat{\mu}^i_{k',v} - \alpha_{t_k} \leq \mu^i_{k',v} \leq \mu^i_{k^*,v} \leq \hat{\mu}^i_{k^*,v} + \alpha_{t_k} \leq \hat{\mu}^i_{k',v} + \alpha_{t_k} \\ &\Rightarrow \hat{\mu}^i_{k',v} - \alpha_{t_k} \leq \mu^i_{k^*,v} \leq \hat{\mu}^i_{k',v} + \alpha_{t_k} \end{aligned} \tag{6}$$

If the previous inequality (6) holds for all values v of the variable x_i, we have:

$$\begin{aligned} &\sum_{v\in\{0,1\}} \left(\hat{\mu}^i_{k',v} - \alpha_{t_k}\right) \leq \sum_{v\in\{0,1\}} \mu^i_{k^*,v} \leq \sum_{v\in\{0,1\}} \left(\hat{\mu}^i_{k',v} + \alpha_{t_k}\right) \\ &\Leftrightarrow \hat{\mu}^i - 2\alpha_{t_k} \leq \mu^i \leq \hat{\mu}^i + 2\alpha_{t_k} \end{aligned} \tag{7}$$

If the previous inequality holds for $i' = \arg\max_i \hat{\mu}^i$, we have:

$$\hat{\mu}^{i'} - 2\alpha_{t_k} \leq \mu^{i'} \leq \mu^{i^*} \tag{8}$$

As a consequence, the variable x_i cannot be the best one when:

$$\hat{\mu}^i + 2\alpha_{t_k} \leq \hat{\mu}^{i'} - 2\alpha_{t_k} \ , \tag{9}$$

Using the union bound, the probability of making an error about the selection on the next variable by eliminating each variable x_i when the inequality (9) holds is bounded by the sum for each variable x_i and each value v that the inequality 6 does not hold for k' and k^*:

$$\mathbb{P}\left(i^* \neq i'\right) \leq \sum_{i\in\mathcal{V}} \frac{K\delta\pi^2}{12KM} \leq \sum_{i\in\mathcal{V}} \frac{\delta}{M} \leq \delta \tag{10}$$

Now, we have to consider t^*_k, the number of steps needed to select the optimal variable. If the best variable has not been eliminated (probability $1-\delta$), the last variable x_i is eliminated when:

$$\hat{\mu}^{i^*} - \hat{\mu}^i \geq 4\alpha_{t^*_k}$$

The difference between the expected reward of a variable x_i and the best next variable is defined by:

$$\Delta_i = \mu^{i^*} - \mu^i$$

Assume that:

$$\Delta_i \geq 4\alpha_{t_k} \tag{11}$$

The following inequality holds for the variable x_i with a probability $1 - \frac{\delta}{KM}$:

$$\hat{\mu}^i - 2\alpha_{t_k} \leq \mu^i \leq \hat{\mu}^i + 2\alpha_{t_k}$$

Then, using the previous inequality in the inequality (11), we obtain:

$$(\hat{\mu}^{i^*} + 2\alpha_{t_k}) - (\hat{\mu}^i + 2\alpha_{t_k}) \geq \mu^{i^*} - \mu^i \geq 4\alpha_{t_k}$$

Hence, we have:

$$\hat{\mu}^{i^*} - \hat{\mu}^i \geq 4\alpha_{t_k}$$

The condition $\Delta_i \geq 4\alpha_{t_k}$ implies the elimination of the variable x_i. Then, we have:

$$\Delta_i \geq 4\alpha_{t_k}$$

$$\Rightarrow \Delta_i^2 \geq \frac{8}{t_k} \log \frac{4KMt_k^2}{\delta} \tag{12}$$

The time t_k^*, where all non optimal variables have been eliminated, is reached when the variable corresponding to the minimum of Δ_i^2 is eliminated.

$$\Rightarrow \Delta^2 \geq \frac{8}{t_k^*} \log \frac{4KM{t_k^*}^2}{\delta} \; , \tag{13}$$

where $\Delta = \min_{i \neq i^*} \Delta_i$.

The inequality (13) holds for all variables x_i with a probability $1 - \delta$ for:

$$t_k^* = \frac{64}{\Delta^2} \log \frac{4KM}{\delta.\Delta} \tag{14}$$

Indeed, if we replace the value of t_k^* in the right term of the inequality (12), we obtain:

$$\begin{aligned}
&\frac{\Delta^2}{8\log\frac{4KM}{\delta.\Delta}} \left(\log \frac{4KM}{\delta} + 2\log\frac{64}{\Delta^2} + 2\log\log\frac{4KM}{\delta.\Delta} \right) = \\
&\frac{\Delta^2}{8\log\frac{4KM}{\delta.\Delta}} \left(\log \frac{4KM}{\delta} - 4\log\Delta + 12\log 2 + 2\log\log\frac{4KM}{\delta.\Delta} \right) \leq \\
&\frac{\Delta^2}{8\log\frac{4KM}{\delta.\Delta}} \left(4\log \frac{4KM}{\delta.\Delta} + 12\log 2 + 2\log\log\frac{4KM}{\delta.\Delta} \right)
\end{aligned}$$

For $x \geq 13$, we have:

$$12\log 2 + 2\log\log x < 4\log x$$

Hence, for $8KM \geq 13$, we have:

$$\begin{aligned}
&\frac{\Delta^2}{8\log\frac{4KM}{\delta.\Delta}} \left(4\log \frac{4KM}{\delta.\Delta} + 12\log 2 + 2\log\log\frac{4KM}{\delta.\Delta} \right) \leq \\
&\frac{\Delta^2}{8\log\frac{4KM}{\delta.\Delta}} 8\log\frac{4KM}{\delta.\Delta} = \Delta^2
\end{aligned}$$

Hence, we obtain:

$$t_k^* = \frac{64}{\Delta^2} \log \frac{4KM}{\delta} \text{ , with a probability } 1 - \delta$$

As the actions are chosen the same number of times (Round-robin function), we have $t = Kt_k$, and thus:

$$t^* = \frac{64K}{\Delta^2} \log \frac{4KM}{\delta} \text{ , with a probability } 1 - \delta$$

□

7.3 Lemma 2

Proof. Let $y_{k,v}^i$ be a bounded random variable corresponding to the reward of the action k when the value v of the variable i is observed. Let y^i be a random variable such that:

$$y^i = \max_k y_{k,v}^i$$

We have:

$$\mathbb{E}_{D_{x,y}}[y^i] = \mu^i$$

Each y^i is updated each step t_k when each action has been played once. Let Θ be the sum of the binary random variables $\theta_1, ..., \theta_{t_k}, ..., \theta_{t_k^*}$ such that $\theta_{t_k} = \mathbb{1}_{y^i(t_k) \geq y^j(t_k)}$. Let p_{ij} be the probability that the use of variable i leads to more rewards than the use of variable j. We have:

$$p_{ij} = \frac{1}{2} - \Delta_{ij} \text{ , where } \Delta_{ij} = \mu^i - \mu^j.$$

Slud's inequality (see Slud (1977)) states that when $p \leq 1/2$ and $t_k^* \leq x \leq t_k^*.(1-p)$, we have:

$$P(\Theta \geq x) \geq P\left(Z \geq \frac{x - t_k^*.p}{\sqrt{t_k^*.p(1-p)}}\right) , \tag{15}$$

where Z is a normal $\mathcal{N}(0,1)$ random variable.

To choose the best variable between i and j, one needs to find the time t_k^* where $P(\Theta \geq t_k^*/2) \geq \delta$. To state the number of trials t_k^* needed to estimate Δ_{ij}, we recall and adapt the arguments developed in Mousavi (2010). Using Slud's inequality (see equation 15), we have:

$$P(\Theta \geq t_k^*/2) \geq P\left(Z \geq \frac{t_k^*.\Delta_{ij}}{\sqrt{t_k^*.p_{ij}(1-p_{ij})}}\right) , \tag{16}$$

Then, we use the lower bound of the error function (see Chu (1955)):

$$P(Z \geq z) \geq 1 - \sqrt{1 - \exp\left(-\frac{z^2}{2}\right)}$$

Therefore, we have:

$$\begin{aligned} P(\Theta \geq t_k^*/2) &\geq 1 - \sqrt{1 - \exp\left(-\frac{t_k^*.\Delta_{ij}^2}{2p_{ij}(1-p_{ij})}\right)} \\ &\geq 1 - \sqrt{1 - \exp\left(-\frac{t_k^*.\Delta_{ij}^2}{p_{ij}}\right)} \\ &\geq \frac{1}{2}\exp\left(-\frac{t_k^*.\Delta_{ij}^2}{p_{ij}}\right) \end{aligned}$$

As $p_{ij} = 1/2 - \Delta_{ij}$, we have:

$$\log\delta = \log\frac{1}{2} - \frac{t_k^*.\Delta_{ij}^2}{1/2 - \Delta_{ij}} \geq \log\frac{1}{2} - 2t_k^*.\Delta_{ij}^2$$

Hence, we have:

$$t_k^* = \Omega\left(\frac{1}{\Delta_{ij}^2}\log\frac{1}{\delta}\right)$$

Then, we need to use the fact that as all the values of all the variables are observed when one action is played: the $M(M-1)/2$ estimations of bias are solved in parallel. In worst case, $\min_{ij} \Delta_{ij} = \min_j \Delta_{i^*j} = \Delta$. Thus any algorithm needs at least a sample complexity t^*, where:

$$t^* = K.t^*_k = \Omega\left(\frac{K}{\Delta^2}\log\frac{1}{\delta}\right)$$

□

7.4 Theorem 1

Proof. Lemma 1 states that the sample complexity needed to find the best variable is:

$$t_1^* = \frac{64K}{\Delta_1^2}\log\frac{4KM}{\delta\Delta_1}, \text{ where } \quad \Delta_1 = \min_{i\neq i^*}(\mu^{i^*} - \mu^i)$$

Lemma 3 states that the sample complexity needed to find the optimal action for a value v of the best variable is:

$$t_{2,v}^* = \frac{64K}{\Delta_{2,v}^2}\log\frac{4K}{\delta\Delta_{2,v}}, \text{ where } \quad \Delta_{2,v} = \min_{k\neq k^*}(\mu^{i^*}_{k^*,v} - \mu^{i^*}_{k,v}).$$

The sample complexity of decision stump algorithm is bounded by the sum of the sample complexities of variable selection and action elimination algorithms:

$$t^* = t_1^* + t_2^* \text{ , where } t_2^* = \max_v t_{2,v}^*.$$

□

7.5 Theorem 2

Proof. In worst case, all the values of variables have different best actions, and $K = 2M$. If an action is suppressed before the best variable is selected, the estimation of the mean reward of one variable is underestimated. In worst case this variable is the best one, and a sub-optimal variable is selected. Thus, the best variable has to be selected before an action be eliminated. The lower bound of the decision stump problem is the sum of variable selection and best arm identification lower bounds, stated respectively in Lemma 2 and Lemma 4.

□

7.6 Theorem 3

Proof. The proof of Theorem 3 uses Lemma 1 and Lemma 3. Using the slight modification of the variable elimination inequality proposed in section 3, Lemma 1 states that for each decision stump, we have:

$$\mathbb{P}\left(i^*_{d_\theta}|c_\theta \neq i'_{d_\theta}|c_\theta\right) \leq \frac{\delta_1}{D_\theta L}$$

Then, from the union bound, we have for each path c_θ:

$$\begin{aligned}\mathbb{P}(c_\theta \neq c_\theta^*) \leq & \mathbb{P}\left(i_1^* \neq i_1'\right) + \mathbb{P}\left(i_2^*|x_{i_1}(t) \neq i_2'|x_{i_1}(t)\right) + ... \\ & + \mathbb{P}\left(i^*_{D_\theta}|(x_{i_1}(t), ..., x_{i_{D_\theta - 1}}(t)) \neq i'_{D_\theta}|(x_{i_1}(t), ..., x_{i_{D_\theta - 1}}(t))\right) \leq \frac{\delta_1}{L}\end{aligned}$$

For the action corresponding to the path c_θ, Lemma 3 states that:

$$\mathbb{P}(k \neq k^*) \leq \frac{\delta_1}{L}$$

From the union bound, we have for any path c_θ:

$$\mathbb{P}(c_\theta \neq c^*_\theta) \leq \delta_1 \text{ and } \mathbb{P}(k \neq k^*) \leq \delta_1$$

Using Lemma 1 and Lemma 3, and summing the sample complexity of each 2^{D_θ} variable selection tasks and the sample complexity of each 2^{D_θ} action selection tasks, we bound the sample complexity of any tree θ by:

$$t^* \leq 2^D \frac{64K}{\Delta_1^2} \log \frac{4KMDL}{\delta \Delta_1} + 2^D \frac{64K}{\Delta_2^2} \log \frac{4KL}{\delta \Delta_2} \ ,$$

where $\delta = 2\delta_1$, and $D = \max D_\theta$.

□

7.7 Theorem 4

Proof. To build a decision tree of depth D_θ, any greedy algorithm needs to solve $\sum_{d<D_\theta} 2^d = 2^{D_\theta}$ variable selection problems (one per node), and 2^{D_θ} action selection problems (one per leaf). Then, using Lemma 2 and Lemma 4, any greedy algorithm needs a sample complexity of at least:

$$t^* \geq \Omega \left(2^D \left[\frac{1}{\Delta_1^2} + \frac{1}{\Delta_2^2} \right] K \log \frac{1}{\delta} \right)$$

□

7.8 Additional experimental results

We provide below (see Table 3) the classification rates to compare the asymptotical performances of each algorithm, and the processing times.

References

Chu, J. T.: On bounds for the normal integral. Biometrika 42:263-265, 1955.

Mousavi, N.: How tight is Chernoff bound ?
https://ece.uwaterloo.ca/ nmousavi/Papers/Chernoff-Tightness.pdf

Slud, E. V.: Distribution Inequalities for the Binomial Law. Ann. Probab.,5(3):404-412, 1977.

Table 3: Summary of results on the datasets played in a loop. The regret against the optimal *random forest* is evaluated on ten trials. Each trial corresponds to a random starting point in the dataset. The confidence interval is given with a probability 95%. The classification rate is evaluated on the last 100000 contexts. The mean running time was evaluated on a simple computer with a quad core processor and 6 GB of RAM.

Algorithm	Regret	Classification rate	Running time
Forest Cover Type, action: Cover Type (7 types)			
BANDITRON	$1.99\ 10^6 \pm 10^5$	49.1%	10 min
LINUCB	$1.23\ 10^6 \pm 10^3$	60%	360 min
NEURALBANDIT	$0.567\ 10^6 \pm 2.10^4$	68%	150 min
BANDIT TREE $D = 8$	$0.843\ 10^6 \pm 2.10^5$	64.2%	5 min
BANDIT FOREST			
$D10 - 18$	$0.742\ 10^6 \pm 5.10^4$	65.8%	500 min
Adult, action: occupation (14 types)			
BANDITRON	$1.94\ 10^6 \pm 3.10^4$	21.1%	20 min
LINUCB	$1.51\ 10^6 \pm 4.10^4$	25.7%	400 min
NEURALBANDIT	$1.2\ 10^6 \pm 10^5$	29.6%	140 min
BANDIT TREE $D = 8$	$1.33\ 10^6 \pm 10^5$	27.9%	4 min
BANDIT FOREST			
$D10 - 18$	$1.12\ 10^6 \pm 7.10^4$	31%	400 min
Census1990, action: Yearsch (18 types)			
BANDITRON	$2.07\ 10^6 \pm 2.10^5$	27%	26 min
LINUCB	$0.77\ 10^6 \pm 5.10^4$	40.3%	1080 min
NEURALBANDIT	$0.838\ 10^6 \pm 10^5$	41.7%	300 min
BANDIT TREE $D = 8$	$0.78\ 10^6 \pm 2.10^5$	41%	10 min
BANDIT FOREST			
$D10 - 18$	$0.686\ 10^6 \pm 5.10^4$	43.2%	1000 min