Covariance-Adaptive Least-Squares Algorithm for Stochastic Combinatorial Semi-Bandits

Julien Zhou [1] & Pierre Gaillard [2] & Thibaud Rahier [3]
& Houssam Zenati [4] & Julyan Arbel [5]

[1] *Criteo AI Lab, Inria Thoth & Statify, France, julien.zhou@inria.fr;*
[2] *Inria Thoth, France, pierre.gaillard@inria.fr ;*
[3] *Criteo AI Lab, France, t.rahier@criteo.com;*
[4] *Inria Soda & PreMeDICaL, France, houssam.zenati@inria.fr;*
[5] *Inria Statify, France, julyan.arbel@inria.fr*

Abstract. We address stochastic combinatorial semi-bandit problems, where a player can select from P subsets of a set containing d base items. This setting has been studied extensively in prior works which focused on getting acceptable regret upper bounds, despite the potentially exponentially big action space. However, most existing algorithms (e.g. `CombUCB`, `ESCB`, `OLS-UCB`) require prior knowledge on the reward distribution, like an upper bound on a sub-Gaussian proxy-variance, which is hard to estimate tightly. Their regret upper bounds also involve these hypotheses, which can make them slack depending on the instance at hand. We propose a variance-adaptive version of `OLS-UCB`, relying on an online estimation of the covariance coefficients and computing upper confidence bounds for each action available. Estimating the coefficients of a covariance matrix can be manageable in practical settings and leveraging this information with the structure of the problem results in improved regret upper bounds.

Keywords. Bandits; Stochastic Combinatorial Semi-Bandit; Covariance; Confidence Ellipsoid

1 Introduction

In sequential decision-making, the bandit framework has been studied in-depth and was instrumental to several applications. Reference books like Bubeck and Cesa-Bianchi (2012) or Lattimore and Szepesvári (2020) offer a wide perspective on the subject. In this framework, a *decision-maker* or *player* must make choices and receive associated rewards. As the player lacks prior knowledge of its an exploration-exploitation trade-off naturally arises.

We focus on the stochastic combinatorial semi-bandit framework. In this setting, the player chooses a subset of *base items* and receives a feedback for each item chosen. The corresponding action set is unfortunately potentially exponentially big and difficult to explore. However, it presents a structure that could be leveraged. The information collected by choosing different overlapping actions should be shared.

Problem formulation. We consider a set of $d \in \mathbb{N}^*$ *base items*, each item $i \in [d] = \{1, \dots, d\}$ yielding a stochastic reward. A *player* accesses these rewards through a set $\mathcal{A} \subseteq \{0,1\}^d$ of $P = |\mathcal{A}| \in \mathbb{N}^*$ *actions*, each corresponding to a subset of items. The player interacts with the *environment* over a sequence of $T \in \mathbb{N}^*$ *rounds*. At each round $t \in [T]$, the decision-maker chooses an action $A_t \in \mathcal{A}$, the environment samples a reward vector $Y_t \in \mathbb{R}^d$, the decision-maker observes the realization for every item contained in A_t, and receives their sum. The interactions between the player and the environment are summarized in Framework 1.

Framework 1 Stochastic Combinatorial Semi-Bandit

For each $t \in \{1, \dots, T\}$:

- The player chooses an action $A_t \in \mathcal{A}$.
- The environment samples a vector of rewards $Y_t \in \mathbb{R}^d$ from a fixed unknown distribution.
- The player receives the reward $\langle A_t, Y_t \rangle = \sum_i A_{t,i} Y_{t,i}$.
- The player observes $Y_{t,i}$ for all $i \in [d]$ s.t. $A_{t,i} = 1$.

The objective of the decision-maker is to maximize the cumulative expected rewards, or equivalently to minimize the expected cumulative regret defined as:

$$\mathbb{E}[R_T] = T\langle a^*, \mu \rangle - \sum_{t=1}^{T} \mathbb{E}\big[\langle A_t, Y_t \rangle\big] = \sum_{t=1}^{T} \mathbb{E}\big[\Delta_{A_t}\big] \,, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $\mathbb{R}^d$, $a^* \in \arg\max_{a \in \mathcal{A}} \langle a, \mu \rangle$ is an optimal action, and $\Delta_a = \langle a^* - a, \mu \rangle$ is the *sub-optimality gap* for action $a \in \mathcal{A}$.

Assumptions. We make the following assumptions on the rewards $(Y_t)_{t \in [T]}$. For all $t \in [T]$, Y_t is independent of $\mathcal{F}_{t-1} = \sigma(Y_1, \dots, Y_{t-1})$ and A_t is $\mathcal{F}_{t-1}-$measurable. There exists a mean reward vector $\mu \in \mathbb{R}^d$ and a positive semi-definite covariance matrix $\mathbf{\Sigma}^* \in M_d(\mathbb{R})$ such that $\mathbb{E}[Y_t] = \mu \in \mathbb{R}^d$ and $\mathrm{Var}(Y_t) = \mathbf{\Sigma}^*$. There exists a known *bounds* vector $B \in \mathbb{R}_+^d$ such that for all $i \in [d]$, $|Y_{t,i} - \mu_i| \leq B_i$.

Contributions. We design `OLS-UCBV`, a variance-adaptive algorithm for stochastic combinatorial semi-bandits. Compared to the existing litterature, our algorithm takes the covariance structure into account. It estimates coefficients of a covariance matrix online which can be done tightly, and efficiently. Leveraging these online estimates, `OLS-UCBV` satisfies a poly-log(T) gap-dependant regret bound, with leading terms depending on the variance and the structure of the action space. We solve the limitations of Degenne and Perchet (2016) raised by Perrault et al. (2020), but with different tools yielding a computationally more efficient algorithm, and manage to get a $\sqrt{T}$ gap-free regret upper bounds contrary to the latest.

2 OLS-UCBV

We design OLS-UCBV, a new algorithm that efficiently leverages semi-bandit feedbacks by approximating the coefficients of the covariance matrix $\mathbf{\Sigma}^*$ online, with coefficient-wise upper bounds. OLS-UCBV satisfies the "Optimism in the face of uncertainty" (OFUL) principle of Abbasi-Yadkori et al. (2011), by deriving ellipsoidal confidence regions leveraging Bernstein-like concentration inequalities and a peeling trick.

2.1 Estimators for mean and covariance

Mean estimation. Let $a \in \mathcal{A}$, $t \in [T]$, we denote by $\mathbf{d}_a = \mathrm{diag}(a) \in M_d(\mathbb{R})$ the diagonal matrix where the non-null coefficients are the elements of the action a. The number of times two items $i, j \in [d]$ (with possibly $i = j$) have been chosen together after round t is denoted by $n_{t,(i,j)} = \sum_{s=1}^{t} \mathbb{1}\{\{i,j\} \subseteq A_s\}$. We define $\mathbf{D}_t = \mathrm{diag}((n_{t,(i,i)})_{i\in[d]}) \in M_d(\mathbb{R})$ the diagonal matrix of item counts. Then the least-squares estimator of the mean reward vector μ using all the data from the past rounds after round t is the empirical average

$$\hat{\mu}_t = \mathbf{D}_t^{-1} \sum_{s=1}^{t} \mathbf{d}_{A_s} Y_s = \mu + \mathbf{D}_t^{-1} \sum_{s=1}^{t} \mathbf{d}_{A_s} \eta_s \,, \tag{2}$$

where η_s denotes the deviation of reward Y_s from its mean μ. This average yields the estimator $\langle a, \hat{\mu}_t \rangle$ for the mean reward $\langle a, \mu \rangle$, to which a well-designed optimistic bonus should be added.

Covariance estimation. Let $t \in \mathbb{N}^*$ and $i, j \in [d]$ such that $n_{t,(i,j)} \geq 2$. The coefficients of $\mathbf{\Sigma}^*$ can be estimated online by $\hat{\chi}_t$ with elements as follows

$$\hat{\chi}_{t,(i,j)} = \frac{1}{n_{t,(i,j)}} \sum_{s=1}^{t} A_{s,i} A_{s,j} \mathbb{1}\{n_{s,(i,j)} \geq 2\} (Y_{s,i} - \hat{\mu}_{s-1,i})(Y_{s,j} - \hat{\mu}_{s-1,j}) \,. \tag{3}$$

A major difference with usual batch estimators, like those used in Perrault et al. (2020) and Audibert et al., 2009, is the fact that the observed rewards are "centered" using the sample average of the past rewards at the time of the observation, instead of the whole sample average at time t.

2.2 Confidence bounds for covariance and mean

Covariance coefficients upper confidence bound. The following result controls the error of the online covariance estimator $\hat{\chi}_t$ presented in eq. (3).

Proposition 1. *Let $T \geq 3$, $\delta \in (0,1)$. Then with probability $1-\delta$, for all $t \leq T$ and $(i,j) \in [d]^2$, such that $n_{t,(i,j)} \geq 2$,*

$$|\hat{\chi}_{t,(i,j)} - \mathbf{\Sigma}^*_{i,j}| \leq \mathcal{B}_{t,(i,j)}(\delta) \,,$$

where $\mathcal{B}_{t,(i,j)}(\delta) = 3B_iB_j \left(\frac{h_{T,\delta}}{\sqrt{n_{t,(i,j)}}} + \frac{h^2_{T,\delta}}{n_{t,(i,j)}} \log(T) \right)$ *with* $h_{T,\delta} = \log(5d^2T^2/\delta)$.

This result suggests the following upper confidence bounds for $\mathbf{\Sigma}^*$ to be used into our algorithm:

$$\hat{\mathbf{\Sigma}}_{t,(i,j)} = \hat{\chi}_{t,(i,j)} + \mathcal{B}_{t,(i,j)}(\delta)\,. \tag{4}$$

It also yields $\mathcal{C}$, the high-probability event in which all the $\chi_{t,(i,j)}$ are in our confidence intervals.

$$\mathcal{C} = \left\{\forall t \in [T], (i,j) \in [d]^2 \text{ s.t. } n_{t,(i,j)} \geq 2, \qquad |\hat{\chi}_{t,(i,j)} - \mathbf{\Sigma}^*_{i,j}| \leq \mathcal{B}_{t,(i,j)}(\delta)\right\}. \tag{5}$$

Mean upper confidence bound. We propose an upper confidence bound for the average rewards of all actions $a \in \mathcal{A}$ at any round t.

Denoting by $\hat{\mathbf{\Sigma}}_t$ the covariance matrix upper bound whose coefficients are given by (4), we introduce $\hat{\mathbf{Z}}_t$ a "regularized empirical design matrix" and its "exact" counterpart $\mathbf{Z}_t$,

$$\hat{\mathbf{Z}}_t = \sum_{s=1}^{t} \mathbf{d}_{A_s} \hat{\mathbf{\Sigma}}_t \mathbf{d}_{A_s} + \mathbf{d}_{\hat{\mathbf{\Sigma}}_t} \mathbf{D}_t + d\mathbf{d}_B\,, \tag{6}$$

$$\mathbf{Z}_t = \sum_{s=1}^{t} \mathbf{d}_{A_s} \mathbf{\Sigma}^* \mathbf{d}_{A_s} + \mathbf{d}_{\mathbf{\Sigma}^*} \mathbf{D}_t + d\mathbf{d}_B\,, \tag{7}$$

where $\mathbf{d}_B = \mathrm{diag}((B_i^2)_{i\in[d]})$ and $\mathbf{d}_{\hat{\mathbf{\Sigma}}_t} = \mathrm{diag}(\hat{\mathbf{\Sigma}}_t)$ are matrices in $M_d(\mathbb{R})$.

Let $\delta \in]0,1[$ and $f_{t,\delta} = 6d\log(\log(1+t)) + 3d\log(1+e) + \log(1/\delta)$ be an exploration parameter, we define $\mathcal{G}_t$ the event in which the estimation error remain in an ellipsoid defined by $\mathbf{Z}_t^{-1}$:

$$\mathcal{G}_t = \left\{ \left\| \sum_{s=1}^{t} \mathbf{d}_{A_s} \eta_s \right\|_{\mathbf{Z}_t^{-1}} \leq f_{t,\delta} \right\}. \tag{8}$$

Notably, this event happen at each round with high probability.

Proposition 2. *For events $\{\mathcal{G}_t\}_{t\leq T}$ defined as in (8), we have*

$$\textstyle\sum_{t=d(d+1)}^{T-1} \mathbb{P}\left(\mathcal{G}_t^c\right) \leq \delta T^2. \tag{9}$$

2.3 Algorithm

We now present `OLS-UCBV` written in Algorithm 2. The algorithm begins with an initial exploration phase by sampling every base item $i \in [d]$ and every "reachable" couple $(i,j) \in [d]^2$ at least twice. Then, for all subsequent rounds $t+1$, `OLS-UCBV` picks an action A_{t+1} such that:

$$A_{t+1} \in \arg\max_{a\in\mathcal{A}} \left\{ \langle a, \hat{\mu}_t \rangle + f_t \, \left\| \mathbf{D}_t^{-1} a \right\|_{\hat{\mathbf{Z}}_t} \right\}. \tag{10}$$

Algorithm 2 `OLS-UCBV`

Input $\delta > 0$, $T \geq 1$, $B \in \mathbb{R}_+^d$
for $t = 1, \ldots, T$ **do**
  if $\left\{a \in \mathcal{A} \text{ s.t } \min_{i,j \in a} n_{t,(i,j)} \leq 1\right\} \neq \emptyset$ **then**
    Choose any A_t in the above set
  else
    Choose $A_t \in \mathcal{A}$ from (10) using $\hat{\mu}_{t-1}, \hat{\mathbf{Z}}_{t-1}$
    Environment samples $Y_t \in \mathbb{R}^d$
    Receive reward $\langle A_t, Y_t \rangle = \sum_i A_{t,i} Y_{t,i}$
    Compute $\hat{\mu}_t$ from (2)
    Compute $\hat{\mathbf{\Sigma}}_t$ from (3) and (4)
    Compute $\hat{\mathbf{Z}}_t$ from (6)
  end if
end for

2.4 Regret upper bound

We establish the following regret upper bounds for `OLS-UCBV`.

Theorem 3. *Let $T \geq 5$, $B \in \mathbb{R}_+^d$, and $\delta = 1/T^2$. Then, `OLS-UCBV` (Alg. 2) satisfies the* gap-dependent *regret upper bound*

$$\mathbb{E}[R_T] = \tilde{O}\left(\sum_{i=1}^{d} \max_{a \in \mathcal{A}/i \in a, \Delta_a > 0} \frac{\sigma_{a,i}^2}{\Delta_a} \right),$$

and the distribution-free *regret upper bound*

$$\mathbb{E}[R_T] = \tilde{O}\left(\sqrt{T \sum_{i=1}^{d} \max_{a \in \mathcal{A}/i \in a} \sigma_{a,i}^2} \right).$$

where $\sigma_{a,i}^2 = \sum_{j \in a} (\mathbf{\Sigma}_{i,j}^*)_+$ for $i \in [d]$ and $a \in \mathcal{A}$, and $(\,\cdot\,)_+ = \max\{\cdot, 0\}$

The logarithmic factors in the dimension d and time T are neglected in the statement. The proof rely on analyzing what happens in the events $\{\mathcal{C} \bigcap \mathcal{G}_t\}$ and is not detailed here. The thorough analysis will be available in a HAL/Arxiv version of this work.

3 Comparisons

We compare asymptotic regret upper bounds of different algorithms in table 1. `CUCB` for (Kveton et al., 2015) uses proxy variances and consider "worst-case" correlations, which is not really satisfactory if we know for example that some base items are independant. In order to account for this possibility, (Degenne and Perchet, 2016) introduces `OLS-UCB` which statisfies a gap-dependent regret upper bound depending on the structure. However, this algorithm needs a proxy-covariance matrix as input, which can be tricky to estimate tightly.

Perrault et al. (2020) manage to replace this by the "true" covariance matrix of the reward distribution, which is estimated online. But their upper bound also include an unsatisfactory $1/\Delta^2$ term which prevent to get a $\sqrt{T}$ gap-free upper bound. `OLS-UCB` manages to satisfies the same kind of gap-dependant upper-bound as `ESCB-C`, but to bypasses the later issue.

Another advantage of `OLS-UCBC` over `ESCB-C` is its computational complexity. The upper bounds of `OLS-UCBC` have a closed form while `ESCB-C` needs to solve linear program over convex sets to compute them.

Algorithm	Info.	Gap-Dependant Asymptotic Regret	Gap-Free Asymptotic Regret
`CUCB`	Γ	$d\sum_i \frac{\Gamma_{i,i}}{\min_{a/i\in a}\Delta_a}$	$\sqrt{dT\sum_i \Gamma_{i,i}}$
`OLS-UCB`	Γ	$(1+\gamma d)\sum_i \frac{\Gamma_{i,i}}{\min_{a/i\in a}\Delta_a} + \frac{d^3 \max_i \Gamma_{i,i}}{\Delta^2_{\min}}$	$(d^3 \max_i \Gamma_{i,i})^{1/3} T^{2/3}$
`ESCB-C`	$\varnothing$	$\frac{1}{\Delta}\sum_i \max_{a/i\in a}\sum_{j\in a}(\mathbf{\Sigma}^*_{i,j})_+ + \frac{d^3 \max_i \Gamma_{i,i}}{\Delta^2_{\min}}$	$(d^3 \max_i \Gamma_{i,i})^{1/3} T^{2/3}$
`OLS-UCBV`	$\varnothing$	$\sum_i \max_{a/i\in a} \frac{\sum_{j\in a}(\mathbf{\Sigma}^*_{i,j})_+}{\Delta_a}$	$\sqrt{T\sum_i \max_{a/i\in a}\sum_{j\in a}(\mathbf{\Sigma}^*_{i,j})_+}$

Table 1: Asymptotic $\tilde{O}(\cdot)$ regret bounds for different types of feedback, up to logarithmic terms, for the following algorithms: `CUCB` Kveton et al. (2015), `OLS-USB` Degenne and Perchet (2016), `ESCB-C` Perrault et al. (2020), and `OLS-UCBV` (ours). *Notations:* a refers to actions; i and j refer to items; Γ is a proxy-covariance matrix; we abbreviate $\max\{x, 0\}$ to $(x)_+$ for any $x \in \mathbb{R}$.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Degenne, R. and Perchet, V. (2016). Combinatorial semi-bandit with known covariance. *Advances in Neural Information Processing Systems*.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. (2015). Tight regret bounds for stochastic combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Perrault, P., Valko, M., and Perchet, V. (2020). Covariance-adapting algorithm for semi-bandits with application to sparse outcomes. In *Conference on Learning Theory*.