Clustering Items through Bandit Feedback
Finding the Right Feature out of Many

Maximilian Graf [1] & Victor Thuot [2] & Nicolas Verzelen [2]

[1] *Institut für Mathematik, Universität Potsdam, Potsdam, Germany, graf9@uni-potsdam.de*
[2] *INRAE, MISTEA, Institut Agro, Univ Montpellier, Montpellier, France, victor.thuot@inrae.fr,nicolas.verzelen@inrae.fr*

Résumé. Nous étudions un problème de clustering dans un modèle de type bandit. Considérons n objets décrits par des vecteurs de dimension d quelconque. Les objets sont partitionnés en deux groupes inconnus, et les objets d'un même groupe partagent le même vecteur (inconnu). L'algorithme interagit séquentiellement et de façon adaptative avec l'environnement : à chaque étape, il choisit un objet et une coordonnée, puis observe une évaluation bruitée de cette coordonnée. L'objectif est de retrouver la partition sous-jacente avec une probabilité d'erreur au plus δ, tout en utilisant un budget d'observations aussi faible que possible. Nous proposons un algorithme en deux temps : il identifie d'abord une coordonnée fortement discriminante, puis exploite uniquement cette coordonnée pour classifier l'ensemble des objets. La phase de sélection de coordonnée repose sur l'algorithme *Sequential Halving*, un algorithme très polyvalent en exploration pure. Pour cette méthode, nous montrons que, avec probabilité au moins $1-\delta$, la partition reconstruite est correcte, et nous établissons une borne supérieure non asymptotique sur le budget requis. Enfin, nous dérivons une borne inférieure sur le budget de toute procédure δ-correcte, qui coïncide (à des facteurs logarithmiques près) avec notre borne supérieure dans plusieurs régimes parcimonieux d'intérêt.

Mots-clés. bandit, clustering, exploration pure, détection de signal

Abstract. We study the problem of clustering a set of items from noisy bandit feedback. Each of the n items is associated with a d-dimensional feature vector, where the ambient dimension d can be large. The items are partitioned into two unknown groups, and items within the same group share the same (unknown) feature vector. The learner interacts with the environment sequentially and adaptively: at each round, it selects an item and a feature, then observes a noisy evaluation of that feature. The objective is to recover the underlying partition with a fixed error probability δ while keeping the number of observations as small as possible. We propose an algorithm that first identifies a feature that is discriminative for the clustering task and then exploits this feature to classify all items. The feature selection leverages the Sequential Halving algorithm, a classical and versatile method for pure exploration. We prove that, with probability at least $(1-\delta)$, the recovered partition is exact, and we derive a non-asymptotic upper bound on the sample budget. We also establish an instance-dependent lower bound on the budget of any δ-correct algorithm, which matches our upper bound up to constants in several relevant regimes.

Keywords. Bandit learning, clustering, pure exploration, signal detection

1 Introduction

We present a summary of the results from the article (Graf, Thuot, and Verzelen 2025).

We consider a sequential pure exploration problem in which a learner aims to cluster a set of items from noisy bandit feedback. Each of the n items is represented by a feature vector in $\mathbb{R}^d$, and the items are partitioned into two unknown groups such that items within the same group share the same (unknown) feature vector. At each round, the learner selects an item and a feature, then observes a noisy evaluation of the corresponding entry. Given a target error probability δ, the goal is to recover the partition of the items while keeping the number of observations as small as possible.

This setting is motivated by applications such as interactive/crowdsourced labeling systems (Ariu et al. 2024): items (images, experts) are probed through many possible features/questions, but only a few coordinates are truly discriminative. The learner must therefore decide *which item* and *which feature* to sample at each step.

Main contributions. In this work, we characterize the sample complexity of this two-group clustering task in bandit setting. Our main contributions are as follows:

- We introduce `BanditClustering`, a fully adaptive procedure that outputs the correct partition with probability at least $1-\delta$ and adapts to the unknown group means in order to focus sampling on informative features. In **??**, we derive a tight, non-asymptotic upper bound on its sample complexity as a function of n, d, $\log(1/\delta)$, and the separation between the group means.
- Conversely, in Theorem **??** we establish an information-theoretic lower bound on the budget, which shows that `BanditClustering` is instance-optimal up to logarithmic factors in several regimes of interest.

Related work. Our problem lies within the line of work on active clustering with bandit feedback (Yang, Zhong, and Tan 2024; Thuot et al. 2025; Yavas et al. 2025; Ariu et al. 2024). In these models, however, each query reveals the full feature vector of a chosen item, whereas in our setting only a single coordinate is observed at each step. This one-coordinate observation scheme allows the learner to concentrate its budget on the most informative features, leading to substantially smaller sample complexity in regimes where only a few coordinates are truly discriminative.

Methodologically, our approach builds on good-arm identification and adaptive sensing (Karnin, Koren, and Somekh 2013; Zhao et al. 2023; Katz-Samuels and Jamieson 2020; Chaudhuri and Kalyanakrishnan 2019; Castro 2014), where signal detection is formulated as a sequential, adaptive testing problem. By jointly exploring over items and features, our algorithm also shares similarities with models of dueling bandits and active ranking (Haddenhorst, Bengs, and Hüllermeier 2021; Ailon, Karnin, and Joachims 2014; Saad, Verzelen, and Carpentier 2023), in which the goal is to recover structure (such as rankings or clusters) from noisy, pairwise or coordinate-wise feedback.

2 Problem formulation and notation

Problem. We study a *clustering* problem where a learner must partition n items into two groups from *noisy bandit feedback*. Each item i has a (latent) feature vector $M_{i,\cdot} \in \mathbb{R}^d$, and we assume a hidden binary-partition model: there exist distinct means $\mu_0, \mu_1 \in \mathbb{R}^d$ and labels[1] $g \in \{0,1\}^n$ such that

$$\forall i \in [n] \ : \quad M_{i,\cdot} = \mu_{g(i)} \ . \tag{1}$$

The learner interacts sequentially and adaptively with this environment: at each round t, it selects an item–feature pair $(I_t, J_t) \in [n] \times [d]$ and observes a noisy sample

$$X_t = M_{I_t,J_t} + \epsilon_t,$$

where $(\epsilon_t)_t$ is a sequence of i.i.d. 1-subGaussian noise (e.g., bounded or Gaussian), in the sense that $\mathbb{E}[\exp(x\epsilon_t)] \leq \exp(x^2/2) \quad \forall x \in \mathbb{R}$. Given a confidence level $\delta \in (0,1)$, the objective is to design a *fixed-confidence* (i.e., δ-correct) algorithm that recovers g with probability at least $1-\delta$, while minimizing the *budget* T (the total number of observations).

Algorithm 1 Generic bandit clustering sampling protocol

Require: Confidence level $\delta \in (0,1)$
1: Initialize $t \leftarrow 1$, history $\mathcal{H}_0 \leftarrow \varnothing$
2: **while** stopping rule not satisfied **do**
3: Select an item $I_t \in \{1,\dots,n\}$ based on $\mathcal{H}_{t-1}$ {Sampling rule}
4: Select a feature $J_t \in \{1,\dots,d\}$ based on $\mathcal{H}_{t-1}$
5: Observe $X_t = M_{I_t,J_t} + \epsilon_t$ {Observation}
6: Update statistics from (I_t, J_t, X_t) and $\mathcal{H}_{t-1}$
7: Set $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(I_t, J_t, X_t)\}$
8: Check stopping condition (e.g., current estimate $\hat{g}$ is δ-confident) {Stopping rule}
9: $t \leftarrow t+1$
10: **end while**
11: **Output**: estimated partition $\hat{g} \in \{0,1\}^n$ based on $\mathcal{H}_T$ {Recommendation rule}

Model difficulty. The intrinsic difficulty of the task is governed by:

- the *gap vector* $\Delta := \mu_1 - \mu_0 \in \mathbb{R}^d \setminus \{0\}$ (smaller $\|\Delta\|$ makes clustering harder),
- its *effective sparsity* (only a few coordinates of Δ may be large),
- the *balancedness* of the partition,

$$\theta := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(i)=0} \wedge \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(i)=1} \in [1/n, 1/2].$$

We work in the pure exploration framework and aim to characterize the optimal instance-dependent sample complexity $T(\Delta, \theta, n, d, \delta)$ required for this clustering task.

[1] Unique up to a global flip of all labels.

3 Algorithms and upper bounds

BanditClustering. Our δ-correct algorithm `BanditClustering` exploits bandit feedback to concentrate samples on informative features. It proceeds in two stages:

1. **Representative identification.** Fix $r_0 = 1$ as a representative of the first group. An adaptive signal-detection routine, based on a variant of Sequential Halving, explores item–feature pairs to find $r_1 \in [n]$ such that $M_{r_1,\cdot} \neq M_{r_0,\cdot}$.
2. **Feature selection and classification.** Given representatives (r_0, r_1), the algorithm focuses on $\Delta = M_{r_1,\cdot} - M_{r_0,\cdot}$. A second Sequential-Halving–type procedure over the d coordinates identifies a feature j with large $|\Delta_j|$, then uses this discriminative feature to classify all n items.

Sequential Halving subroutine. Both steps rely on a bandit subroutine, `CSH` (`Compare Sequential Halving`), adapted from the Sequential Halving algorithm and combined with a sub-sampling scheme. Given a reference row r_0, a set of candidate rows $I \subset [n] \setminus \{r_0\}$, a number of halving rounds L and a budget T, `CSH` first draws a subset I_0 of 2^L pairs $(i, j) \in I \times [d]$. Then, at each epoch $l \in [L]$, it allocates a budget of order T/L uniformly over the active set I_l, computes empirical gaps $|M_{i,j} - M_{r_0,j}|$, and discards roughly half of the least promising pairs with the smallest gap so that $|I_{l+1}| = |I_l|/2$. After L rounds, it returns a pair (i, j) with a large empirical gap, using a total budget T.

In the first step (`CandidateRow`), we run `CSH` with $I = [n]\setminus\{r_0\}$ and an increasing sequence of budgets, and then test whether the selected pair (i, j) satisfies $|M_{i,j} - M_{r_0,j}| > 0$ with high confidence. Up to logarithmic factors, this costs

$$T_{\mathrm{rep}}(\delta) \asymp \frac{d}{\theta \, \|\Delta\|_2^2} \log\Big(\frac{1}{\delta}\Big).$$

In the second step (`ClusterByCandidates`), we fix (r_0, r_1) and use `CSH` with $I = \{r_1\}$ to identify a feature j with large $|\Delta_j|$. Once such a feature is found, all items are classified by sampling only coordinate j, with a per-item cost of order $\Delta_j^{-2} \log(n/\delta)$. Our stopping rule balances the cost of selecting a discriminative feature with the cost of classifying all items. This yields a total complexity

$$T_{\mathrm{feat}}(\delta) \asymp \min_{s \in [d]} \Big(\tfrac{d}{s} + n\Big) \tfrac{1}{\Delta_{(s)}^2} \log\Big(\tfrac{1}{\delta}\Big),$$

where $|\Delta_{(1)}| \geq \cdots \geq |\Delta_{(d)}|$.

Theorem 3.1. *The algorithm* ***`BanditClustering`*** *is δ-correct. Moreover, its budget T satisfies*

$$T \lesssim \frac{d}{\theta\|\Delta\|_2^2} \log\Big(\tfrac{1}{\delta}\Big) + \min_{s \in [d]} \Big(\frac{d}{s} + n\Big)\Big(\frac{1}{\Delta_{(s)}^2} + 1\Big) \log\Big(\tfrac{1}{\delta}\Big), \tag{2}$$

up to logarithmic factors in d, n and $1/\delta$.

Our complexity can be interpreted through an *effective sparsity* s and *effective gap magnitude* $\Delta_{(s)}$: very large coordinates of Δ are rare (costly to find), very small ones are frequent but too weak to classify on, and the optimal s balances these two effects. Importantly, the algorithm adapts to this unknown level of sparsity, reaching best trade-off in (**??**). Our lower bound shows that this trade-off is intrinsic to the problem.

4 Lower bounds

We complement our upper bound with an instance-dependent lower bound showing that the two stages of `BanditClustering` are essentially unavoidable. We consider Gaussian environments (which satisfy our noise assumption) obtained from M by permuting rows and columns. Writing $\mathcal{E}_{\mathrm{per}}(M)$ for this family, any algorithm must perform well on at least one such permutation.

Theorem 4.1. *Fix $\delta \in (0, 1/4)$ and let $\mathcal{A}$ be δ-PAC for the clustering task. Then there exists $\tilde{\nu} \in \mathcal{E}_{\mathrm{per}}(M)$ such that*

$$\mathbb{P}_{\mathcal{A},\tilde{\nu}}\left(T \;\geq\; \frac{2d}{\theta\|\Delta\|_2^2}\log\Big(\tfrac{1}{6\delta}\Big) \;\vee\; \frac{2(n-2)}{\Delta_{(1)}^2}\log\Big(\tfrac{1}{4.8\delta}\Big)\right) \;\geq\; \delta. \tag{3}$$

The first term, of order $\frac{d}{\theta\|\Delta\|_2^2}\log(1/\delta)$, corresponds to the cost of finding two items from different groups, and matches (up to logarithmic factors) the representative-identification stage of `BanditClustering`. The second term, of order $\frac{n}{\Delta_{(1)}^2}\log(1/\delta)$, is the cost of classifying all n items once a maximally discriminative feature is known.

In the sparse, constant-magnitude case where Δ only takes two values, this lower bound matches our upper bound from Theorem **??** up to polylogarithmic factors. Indeed, consider the regime where Δ is s-sparse with constant amplitude, i.e.,

$$\Delta_j \in \{0, h\}, \qquad \#\{j : \Delta_j = h\} = s,$$

`BanditClustering` is optimal, and our upper bound (**??**) matches the lower bound (**??**), up to polylogarithmic factors,

$$T \;\lesssim\; \log\Big(\frac{1}{\delta}\Big)\Big(\frac{d}{\theta\,\|\Delta\|_2^2} + \frac{n}{h^2}\Big) \;\asymp\; \log\Big(\frac{1}{\delta}\Big)\Big(\frac{d}{\theta s\,h^2} + \frac{n}{h^2}\Big).$$

5 Experiments

We report numerical results on synthetic data supporting our theoretical guarantees. The code is available online.[2]

[2] `https://github.com/grafmaxi/bandit_two_clusters`

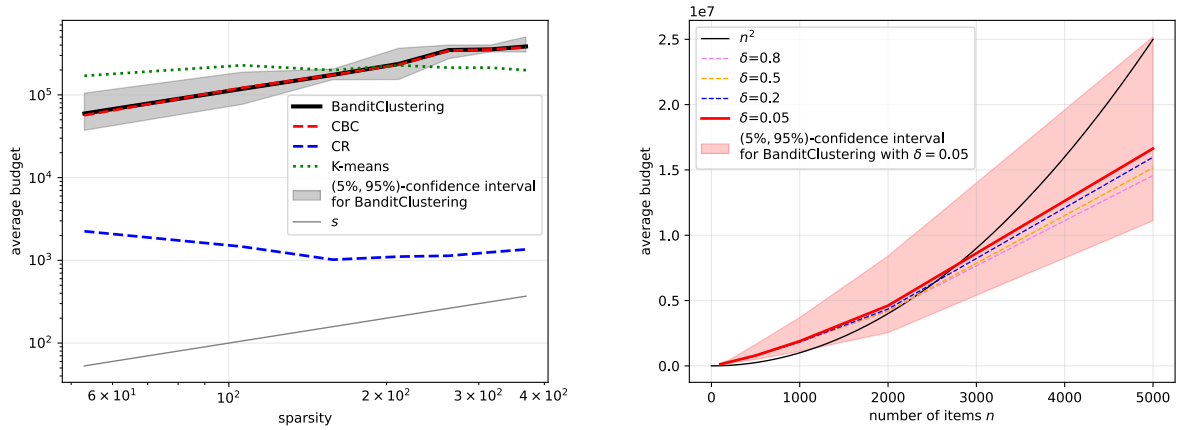

Figure 1: Left: budget vs. sparsity s (fixed $\|\Delta\|_2$). Right: budget vs. n with $d = 10n$.

In the sparse regime (left) with balanced groups ($\theta = 1/2$), we consider Gaussian noise, $n = 20$, $d = 1000$, and a gap vector of fixed norm $\|\Delta\|_2 = 15$, parametrized as $\Delta^s = (\underbrace{h_s, \dots, h_s}_{s \text{ times}}, 0, \dots, 0)$ with $h_s = 15/\sqrt{s}$ for $s \in \{1, \dots, d\}$. In this setting, `BanditClustering` requires significantly fewer samples than a uniform-sampling-plus-K-means baseline, and its cost is dominated by the classification phase, growing approximately linearly in s, as predicted by the complexity term $n/\Delta^2_{(s)}$.

When n and d grow proportionally (right), with $d = 10n$ and a fixed sparse Δ, the empirical budget of `BanditClustering` grows essentially linearly in n, whereas naive uniform sampling would need at least quadratic order to see enough informative coordinates. This linear scaling is consistent with our instance-dependent upper bound, which reduces to order-n behavior when the sparsity, gap magnitude, and balance parameter θ are kept fixed.

6 Discussion

Compared to active clustering models where each pull reveals a full feature vector (Yang, Zhong, and Tan 2024; Thuot et al. 2025; Yavas et al. 2025; Ariu et al. 2024), our one-coordinate feedback avoids high-dimensional terms of order $d^{3/2}\sqrt{n}$ and is particularly advantageous when only a few features are strongly discriminative. In the sparse, constant-gap case, the resulting gains can be of order $n \wedge (d/s)$.

Extensions to more than two groups and to weakly heterogeneous clusters (where within-group variation remains small relative to between-group gaps) appear feasible by reusing our two-group procedure as a building block, but obtaining sharp rates in these more general settings is challenging and left for future work.

Bibliographie

Ailon, Nir, Zohar Karnin, and Thorsten Joachims (2014). "Reducing dueling bandits to cardinal bandits". In: *International Conference on Machine Learning*. PMLR, pp. 856–864.

Ariu, Kaito, Jungseul Ok, Alexandre Proutiere, and Seyoung Yun (2024). "Optimal clustering from noisy binary feedback". In: *Machine Learning* 113.5, pp. 2733–2764.

Castro, Rui M (2014). "Adaptive sensing performance lower bounds for sparse signal detection and support estimation". In: *Bernoulli* 20.4, pp. 2217–2246.

Chaudhuri, Arghya Roy and Shivaram Kalyanakrishnan (2019). "PAC identification of many good arms in stochastic multi-armed bandits". In: *International Conference on Machine Learning*. PMLR, pp. 991–1000.

Graf, Maximilian, Victor Thuot, and Nicolas Verzelen (13–19 Jul 2025). "Clustering Items through Bandit Feedback: Finding the Right Feature out of Many". In: *Proceedings of the 42nd International Conference on Machine Learning*. Ed. by Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu. Vol. 267. Proceedings of Machine Learning Research. PMLR, pp. 20296–20325. URL: https://proceedings.mlr.press/v267/graf25a.html.

Haddenhorst, Björn, Viktor Bengs, and Eyke Hüllermeier (2021). "Identification of the generalized Condorcet winner in multi-dueling bandits". In: *Advances in Neural Information Processing Systems* 34, pp. 25904–25916.

Karnin, Zohar, Tomer Koren, and Oren Somekh (17–19 Jun 2013). "Almost Optimal Exploration in Multi-Armed Bandits". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Atlanta, Georgia, USA: PMLR, pp. 1238–1246.

Katz-Samuels, Julian and Kevin Jamieson (26–28 Aug 2020). "The True Sample Complexity of Identifying Good Arms". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1781–1791.

Saad, El Mehdi, Nicolas Verzelen, and Alexandra Carpentier (2023). "Active ranking of experts based on their performances in many tasks". In: *International Conference on Machine Learning*. PMLR, pp. 29490–29513.

Thuot, Victor, Alexandra Carpentier, Christophe Giraud, and Nicolas Verzelen (24–27 Feb 2025). "Clustering with bandit feedback: breaking down the computation/information gap". In: *Proceedings of The 36th International Conference on Algorithmic Learning Theory*. Ed. by Gautam Kamath and Po-Ling Loh. Vol. 272. Proceedings of Machine Learning Research. PMLR, pp. 1221–1284. URL: https://proceedings.mlr.press/v272/thuot25a.html.

Yang, Junwen, Zixin Zhong, and Vincent YF Tan (2024). "Optimal clustering with bandit feedback". In: *Journal of Machine Learning Research* 25.186, pp. 1–54.

Yavas, Recep Can, Yuqi Huang, Vincent Y. F. Tan, and Jonathan Scarlett (2025). "A General Framework for Clustering and Distribution Matching with Bandit Feedback". In: *IEEE Transactions on Information Theory*, pp. 1–1. DOI: 10.1109/TIT.2025.3528655.

Zhao, Yao, Connor Stephens, Csaba Szepesvári, and Kwang-Sung Jun (2023). "Revisiting simple regret: Fast rates for returning a good arm". In: *International Conference on Machine Learning*. PMLR, pp. 42110–42158.