Online Selection of Diverse Committees

Virginie Do, Jamal Atif, Jérôme Lang and Nicolas Usunier

Abstract

Citizens' assemblies need to represent subpopulations according to their proportions in the general population. These large committees are often constructed in an online fashion by contacting people, asking for the demographic features of the volunteers, and deciding to include them or not. This raises a trade-off between the number of people contacted (and the incurring cost) and the representativeness of the committee. We study three methods, theoretically and experimentally: a greedy algorithm that includes volunteers as long as proportionality is not violated; a non-adaptive method that includes a volunteer with a probability depending only on their features, assuming that the joint feature distribution in the volunteer pool is known; and a reinforcement learning based approach when this distribution is not known a priori but learnt online.

1 Introduction

Forming a representative committee consists in selecting a set of individuals, who agree to serve, in such a way that every part of the population, defined by specific features, is represented proportionally to its size. As a paradigmatic example, the Climate Assembly in the UK and the Citizens' Convention for Climate in France brought together 108 and 150 participants respectively, representing sociodemographic categories such as gender, age, education level, professional activity, residency, and location, in proportion to their importance in the wider society. Beyond citizens' deliberative assemblies, proportional representation often has to be respected when forming an evaluation committee, selecting a diverse pool of students or employees, and so on.

Two key criteria for evaluating the committee formation process are the representativeness of the final selection and the number of persons contacted (each of these incurring a cost). The trade-off is that the higher the number of people contacted, the more proportional the resulting committee.

A first possibility is to use an offline strategy (as for the UK assembly): invitations are sent to a large number of people (30,000), and the final group is selected among the pool of volunteers. An alternative setting which is common in hiring is to consider an online process: the decision-maker is given a stream of candidates and has to decide at each timestep whether or not to admit the candidate to the final committee. This work focuses on the latter setting.

A further difficulty is that the distribution of *volunteers* is not necessarily known in advance. For example, although the target is to represent distinct age groups proportionally to their distribution in the wider population, it may be the case that older people are predominant among volunteers.

Multi-attribute proportional representation in committee selection in an off-line setting usually assumes full access to a finite (typically large) database of candidates. This assumption is impractical in a variety of real-world settings: first, the database does not exist beforehand and constructing it would require contacting many more people than necessary; second, in some domains, the decision to hire someone should be made immediately so that people don't change their mind in the meantime (which is typical in professional contexts).

An online strategy must achieve a good trade-off between sample complexity, i.e. the number of timesteps needed to construct a full committee, and the quality of the final committee, as measured by its distance to the target distribution.

We focus on the online setting. We introduce a new model and offer three different strategies, which rely on different assumptions on the input (and the process). The *greedy* strategy selects volunteers as long as their inclusion does not jeopardize the size and representation constraints; it does not assume any prior distribution on the volunteer pool. The *nonadaptive* strategy, based on constrained Markov decision processes, repeatedly chooses a random person, and decides whether to include or not a volonteer with a probability that depends only on their features; it assumes the joint distribution in the volunteer pool is known; it can be parallelised. Finally, the *reinforcement learning* strategy assumes this distribution is not known a priori but can be learnt online.

Which of these strategies are interesting depends on domain specificities. For each, we study bounds for expected quality and sample complexity, and perform experiments using real data from the UK Citizens' Assembly on Brexit.

The outline of the paper is as follows. We discuss related work in Section 2, define the problem in Section 3, define and study our three strategies in Sections 3.2, 4 and 5, analyse our experiments in Section 6 and conclude in Section 7.

2 Related work

Diversity and representation in committee (s)election The problem of selecting a diverse set of candidates from a candidate database, where each candidate is described by a vector of attribute values, has been considered in several places. In [17], the goal is to find a committee of a fixed size whose distribution of attribute values is as close as possible to a given target distribution. In [10, 9], each candidate has a score, obtained from a set of votes, and some constraints on the proportion of selected candidates with a given attribute value are specified; the goal is to find a fixed-size committee of maximal score satisfying the constraints. In the same vein, [3] considers soft constraints, and [7] do not require the size of the committee to be fixed.[1]

Our online setting shifts the difficulty of the multi-attribute representation problem from computational complexity analyses, to the need for probabilistic guarantees on the tradeoffs between sample complexity and achieved proportionality.

Representative and fair sortition Finding a representative committee (typically, a panel of citizens) with respect to a set of attributes, using *sortition*, is the topic of at least two recent papers. [8] show that stratification (random selection from small subgroups defined by attribute values, rather than from the larger group) only helps marginally. [14] go further and consider this three-stage selection process: (1) letters are sent to a large number of random individuals (the *recipients*); (2) these recipients answer whether they agree to participate, and if so, give their features; those individuals constitute the *pool*; (3) a sampling algorithm is used to select the final *panel* from the pool. As the probability of willingness to participate is different across demographic groups, each person is selected with a probability that depends on their features, so as to correct this self-selection bias. This guarantees that the whole process be fair to all individuals of the population, with respect of going from the initial population to the panel.[2]

---

[1] Note that *diversity* and *proportional representation* are often used with a different meaning in multiwinner elections, namely, in the sense that each voter should feel represented in an elected committee, regardless of attributes. A good entry to this literature is the survey [13].

[2] Fairness guarantees are pushed further in following (yet unpublished) work by the authors: see `https://youtu.be/x_1Ce1kT7vc`.

The main differences between this work and ours are: (1) (once again) our process is online; (2) we do not consider individual fairness, only group representativeness; (3) we care about minimizing the number of people contacted. Moreover, unlike off-line processes, our process can be applied in contexts where hiring a person just interviewed cannot be delayed; this may not be crucial for citizens' assemblies (although someone who volunteers at first contact may change their mind if the delay until the final selection is long), but this is definitely so when hiring a diverse team of employees.

Online selection problems Generalized secretary problems [4] are optimal stopping problems where the goal is to hire the best possible subset of persons, assuming that persons arrive one at a time, their value is observed at that time, and the decision to hire or not them must be taken immediately. The problem has been generalized to finding a set of items maximizing a submodular value function [6, 5] While the latter models do not deal with diversity constraints, [24] aims at selecting a group of people arriving in a streaming fashion from a finite pool, with the goal of optimizing their overall quality subject to diversity constraints. The common point with our approach is the online nature of the selection process. The main differences are that they consider only one attribute, the size of the pool is known, and yet more importantly, what is optimized is the intrinsic quality values of the candidates and not the number of persons interviewed. Closer to our setting is [19] who consider diversity along multiple features in online selection of search results, regardless of item quality. They only seek to maximise diversity, and do not consider trade-offs with the number of items observed.

The diverse hiring setting of [22] is very different. At each time step, the decision-maker chooses which candidate to interview and only decides on which subset to hire after multiple rounds, whereas in our setting, candidates arrive one by one and decisions are made immediately.

3 Formal setting

3.1 Problem definition

Let $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_d$ be the product space of d finite domains, each of size $D_i = |\mathcal{X}_i|$, and where we identify $\mathcal{X}_i$ with $[D_i] = \{1, ..., D_i\}$. Each candidate is represented by a *characteristic vector* $x \in \mathcal{X}$ with d *features*. Let $x^i \in \mathcal{X}_i$ denote the value of the i-th feature. For each $i \in [d]$, we consider a *target vector* $\rho^i \in (0,1)^{D_i}$ with $\sum_{j=1}^{D_i} \rho_j^i = 1$.

The candidate database is infinite and the horizon as well. At each timestep $t \geq 1$, the agent observes a candidate x_t drawn i.i.d. from a stationary distribution p over $\mathcal{X}$, i.e. $x_t \sim p$. The decision-maker must immediately decide between two actions: *accept* or *reject* the candidate, which we respectively denote as $a_t = 1$ and $a_t = 0$.

The goal is to select a *committee* C of K candidates that matches the target vectors as closely as possible, while minimizing the number of candidates screened.

For some set C, let $\lambda(C) \in \prod_{i=1}^d [0,1]^{D_i}$ be the *representation profile* of C, where $\lambda_j^i(C) = \frac{|\{x \in C : x^i = j\}|}{|C|}$. We define the *representation loss* as $\|\lambda(C) - \rho\|_\infty = \max_{i \in [d], j \in [D_i]} |\lambda_j^i(C) - \rho_j^i|$. We evaluate how much C matches the target ρ by the ℓ_∞ metric, because it is harsher than ℓ_1, ℓ_2 on committees that are unacceptable in our applications (e.g. committees with no women that achieve perfect representation on all other categories than gender).

Let $C_t = \{x_{t'} : t' \leq t, a_{t'} = 1\}$ denote the set of all accepted candidates at the end of step t. The agent stops at τ, where τ is the first time when K candidates have been accepted, i.e. the total number of candidates screened. The agent following a (possibly randomized) algorithm ALG must minimize the *sample complexity* $\mathbb{E}^{p,ALG}[\tau]$.

gender \ age	S	J
M	$1/2 - \epsilon'$	$1/4$
F	$1/4$	ϵ'

Table 1: Example candidate distribution p with 2 binary features.

Importantly, we consider two settings: whether the candidate distribution p is *known* or *unknown*.

Remark. In this model, we simply ignore non-volunteers, since the agent only needs to make decisions for volunteers, which from now on we call *candidates*. The joint distribution of characteristic vectors in the population of candidates is p.

3.2 Greedy strategy

We describe a first simple strategy. In `Greedy`, the agent greedily accepts any candidate as long as the number of people in the committee with $x^i = j$ does not exceed the quota $\lceil \rho_j^i K \rceil + \frac{\epsilon K}{(D_i - 1)}$ for any i, j, where $\epsilon > 0$ is some tolerance parameter for the representation quality.

Proposition 1. *The representation loss incurred by* `Greedy` *is bounded as follows:*

$$\|\lambda(C_\tau) - \rho\|_\infty \underset{a.s.}{\leq} \big(\frac{\max_{i \in [d]} D_i - 1}{K} + \epsilon\big).$$

The proof and pseudocode are included in App. A.

This method is simple to interpret and implement, and can even be used when the candidate distribution p is unknown. However, in the following example, we see that `Greedy` may be inefficient because it requires interacting with an arbitrarily large number of candidates to recruit a full committee.

Example 1. *Let $\epsilon' > 0, \ll 1$. There are 2 binary features,* gender *and* age, *with domains $\mathcal{X}_{gender} = \{M, F\}$ and $\mathcal{X}_{age} = \{S, J\}$. The candidates are distributed as p given in Table 1. We want a committee of size $K = 4$ (e.g., a thesis committee) and the target is $\rho^{gender} = (1/2, 1/2)$ and $\rho^{age} = (3/4, 1/4)$.*

Let A be the event that in the first 3 timesteps, the agent observes candidates with characteristic vectors $\{FS, MS, MS\}$ in any order. Then `Greedy` *accepts all of them, i.e. $A = \{C_3 = \{FS, MS, MS\}\}$. We have: $\mathbb{P}[A] = 1/4(1/2 - \epsilon')^2 \times 3! = 3/2(1/2 - \epsilon')^2 \geq 3/2(1/3)^2 = 1/6$.*

Under event A, `Greedy` *can only stop upon finding FJ in order to satisfy the representation constraints. Therefore, $\tau | A$ follows a geometric distribution with success probability ϵ', hence its expectation is $1/\epsilon'$, and $\mathbb{E}^{p, Greedy}[\tau] \geq \mathbb{E}[\tau | A] \times \mathbb{P}[A] = 1/6\epsilon'$. Therefore, the sample complexity of* `Greedy` *in this example is arbitrarily large.*

This example shows the limits of directly applying a naive strategy to our online selection problem, where the difficulty arises from considering multiple features simultaneously, even when there are only 2 binary features. We further discuss the strengths and weaknesses of `Greedy`, and its sensitivity to the tolerance ϵ in our experiments in Section 6.

The greedy strategy is adaptive, in the sense that decisions are made based on the current candidate and candidates accepted in the past. In the following section, we present, with theoretical guarantees, an efficient yet non-adaptive algorithm based on constrained MDPs for the setting in which the candidate distribution is known. We then adapt this approach to the case when this distribution is unknown, using techniques for efficient exploration

/ exploitation in constrained MDPs relying on the principle of optimism in the face of uncertainty.

4 p is known: constrained MDP strategy

In this section, we assume the distribution p is known, and we place ourselves in the limit where we would select a committee of infinite size, and aim to maximize the rate at which candidates are selected, under the constraint that the proportion of accepted candidates per feature value is controlled by ρ. One advantage of this approximation is that the optimal policy is stationary, thus simple to represent. Moreover, as stationary policies can be very well parallelized, in the case where multiple candidates can be interviewed simultaneously. To apply this approach to the finite-size committee selection problem, one needs to interrupt the agent when K candidates have been selected. We showcase a high probability bound of $O(\sqrt{1/K})$ on the representation loss, which guarantees that for large enough values of K, the resulting committee is representative.

From now on, we assume that any feature vector can be observed, i.e., $p(x) > 0$ for all x, so that proportional representation constraints can be satisfied.

4.1 Our model

Fundamentally, our problem could be seen as a contextual bandit with stochastic contexts $x_t \sim p$ and two actions $a_t = 0$ or 1. However, the type of constraints incurred by proportional representation are well studied in constrained MDPs (CMDPs) [2], whereas the contextual bandits literature focused on other constraints (e.g., knapsack constraints [1]). We show how we can efficiently leverage the CMDP framework for our online committee selection problem.

Formally, we introduce an MDP $M = (\mathcal{X}, \mathcal{A}, P, r)$, where the set of states is the d-dimensional candidate space $\mathcal{X}$, the set of actions is $\mathcal{A} = \{0, 1\}$, and the (deterministic) reward is $r(x, a) = \mathbb{1}_{\{a=1\}}$. The transition kernel P, which defines the probability to be in state x' given that the previous state was x and the agent took action a, is very simple in our case: we simply have $P(x'|x, a) = p(x')$ since candidates are drawn i.i.d regardless of the previous actions and candidates.

We consider the *average reward* setting in which the performance of a policy $\pi : \mathcal{X} \times \mathcal{A} \to [0, 1]$ is measured by its *gain* $g^{p,\pi}$, defined as:

$$g^{p,\pi}(x) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}^{p,\pi} \left[\sum_{t=1}^{T} r(x_t, a_t) \Big| x_1 = x \right].$$

We simply write $g^{p,\pi} := g^{\pi}$ when the underlying transition is p without ambiguity.

We include proportional representation constraints following the framework of CMDPs, where the set of allowed policies is restricted by a set of additional constraints specified by reward functions. In our case, for $i \in [d], j \in [D_i]$, we introduce $r_j^i(x, a) = \mathbb{1}_{\{x^i=j, a=1\}}$, and let $\xi_j^i = r_j^i - \rho_j^i r$ be the reward function for the constraint indexed by i, j. Similarly to the gain, we define $h_j^{i\,\pi} = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} \xi_j^i(x_t, a_t) \right]$. The CMDP is defined by:

$$\max_{\pi} \{ g^{\pi} \mid \forall i \in [d], \forall j \in [D_i], h_j^{i\,\pi} = 0 \}. \tag{1}$$

Given the simplicity of the transition kernel, and since the MDP is ergodic by the assumption $p > 0$, the gain is constant, i.e. $\forall x \in \mathcal{X}, g^{\pi}(x) = g^{\pi}$, and problem (1) is well defined. From now on, we only write g^{π} and $\xi_j^{i\,\pi}$. Moreover, the optimal policy for the CMDP (1) is denoted π^* and is *stationary* [2].

Lemma 1. g^π *is the* selection rate *under policy* π:

$$g^\pi = \sum_x p(x)\pi(x,1) = \mathbb{P}^{p,\pi}[a=1]$$

Moreover, if π *is feasible for CMDP* (1)*, then:*

$$\forall i \in [d], \forall j \in [D_i], \mathbb{P}^{p,\pi}[x^i = j | a = 1] = \rho_j^i.$$

Lemma 1 implies that (a) π^* maximises the selection rate of candidates, and (b) the constraints of (1) force candidates x with $x^i = j$ to be accepted in proportions given by ρ_j^i.

The CMDP can be expressed as the linear program:

$$
\begin{aligned}
\max_{\pi \in \mathbb{R}_+^{\mathcal{X} \times \mathcal{A}}} \quad & \sum_{x,a} \pi(x,a)p(x)r(x,a) \\
\text{u.c.} \quad & \forall x \in \mathcal{X}, \sum_a \pi(x,a) = 1 \\
& \forall i,j, \sum_{x,a} \pi(x,a)p(x)\xi_j^i(x,a) = 0.
\end{aligned}
\tag{2}
$$

Notice that problem (2) is feasible by the assumption that $\forall x \in \mathcal{X}, p(x) > 0$. Next we study how well the proportional selection along features is respected when we shift from infinite to finite-sized committee selection.

4.2 Theoretical guarantees

We analyze the `CMDP-based` strategy where at each timestep, the agent observes candidates $x_t \sim p$, decides to accept x_t by playing $a_t \sim \pi^*(.|x_t)$ and stops when K candidates have been accepted. We later refer to it as `CMDP` for brevity.

First, we formally relate the gain g^π that we optimize for in (1) to the quantity of interest $\mathbb{E}^{p,\pi}[\tau]$.

Lemma 2. *For any stationary policy* π, $\mathbb{E}^{p,\pi}[\tau] = \frac{K}{g^\pi}$.

Lemma 2 is a direct consequence of the fact that $\tau + K$ follows a negative binomial distribution with parameters K and $1 - g^\pi$, which are respectively the number of successes and the probability of failure, i.e. of rejecting a candidate under π. Note that this is only true because in our case the transition structure of the MDP ensures constant gain. A quick sanity check shows that if the agent systematically accepts all candidates, i.e. $g^\pi = 1$, then $\mathbb{E}^{p,\pi}[\tau] = K$, and that maximizing g^π is equivalent to minimizing $\mathbb{E}^{p,\pi}[\tau]$.

We exhibit a bound on the representation loss of `CMDP` which follows the optimal stationary policy π^* of CMDP (1). Let $\tilde{d} = \sum_{i=1}^d (D_i - 1)$. ($\tilde{d} = d$ when all features are binary.)

Proposition 2. *Let* π^* *be an optimal stationary policy for CMDP* (1)*. Let* $\delta > 0$*. Then,*

$$\mathbb{P}^{p,\pi^*}\left[\|\lambda(C_\tau) - \rho\|_\infty \leq \sqrt{\frac{\log(\frac{2\tilde{d}}{\delta})}{2K}} \right] \geq 1 - \delta.$$

All proofs of this section are available in Appendix B.1.

The upper bound on the representation loss of `CMDP` decreases with the committee size in $\sqrt{1/K}$. This shows that the stationary policy π^* works well for larger committees, although it acts independently from previously accepted candidates. The intuition is that for larger committees, adding a candidate has less impact on the current representation vector.

Algorithm 1: `RL-CMDP` algorithm.

---

    input : confidence δ, committee size K, targets ρ
    output: committee C_τ

1 $t \leftarrow 0$, $C_0 \leftarrow \emptyset$;
2   **while** $|C_t| < K$ **do**
3   **for** *episode $l = 1, 2, ...$* **do**
4 $\tau_l = t + 1$;
5 $\pi_l \leftarrow$ sol. of (4) via the extended LP (5);
6   **while** $n_t(x_t) < 2n_{\tau_l - 1}(x_t)$ **do**
7 $t \leftarrow t + 1$, Execute π_l;
8   **end**
9   **end**
10 **end**
11 **return** C_t

Example 2. *We take the same attributes and same distribution as in Table 1, with $\epsilon' = 1/6$. Here, the target vectors are $\rho^{gender} = (1/2, 1/2)$ and $\rho^{age} = (1/2, 1/2)$: an ideal committee contains as many women as men, as many senior as junior.*

With the optimal policy for LP (2), each time the current volunteer is a senior male, we select him with probability 1/2; all other volunteers are selected with probability 1. The expected final composition of the pool is 30% of junior male, 30% of senior female, 20% of junior female and 20% of senior male. As the policy selects in average 5/6 of the volunteers, the expected time until we select K candidates is $\mathbb{E}^{p,\pi^*}[\tau] = (6/5)K$. More details can be found in App. E.

5 p is unknown: optimistic CMDP strategy

We now tackle the committee selection problem when the candidate distribution p is unknown and must be learned online. Let $g^* = g^{\pi^*}$ be the value of (1), which is the optimal gain of the CMDP when the distribution p is known. We evaluate a learning algorithm by:

1. the performance regret: $R(T) = \sum_{t=1}^{T} (g^* - r(x_t, a_t))$,

2. the cost of constraint violations:
    $R^c(T) = \max_{i,j} \left| \sum_{t=1}^{T} \xi_j^i(x_t, a_t) \right|$.

We propose an algorithm that we call `RL-CMDP` (Reinforcement Learning in CMDP, Alg. 1). It is an adaptation of the *optimistic* algorithm UCRL2 [16], and it also builds on the algorithm OptCMDP proposed by [12] for finite-horizon CMDPs. Learning in average-reward CMDPs involves different challenges, because there is no guarantee that the policy at each episode has constant gain. It does not matter in our case, since as we noted in Sec. 4, the simple structure of the transition kernel ensures constant gain, and does not require to use the Bellman equation. The few works on learning in average-reward CMDPs make unsuitable assumptions for our setting [27, 23].

`RL-CMDP` proceeds in episodes, which end each time the number of observations for some candidate x doubles. During each episode l, observed candidates x_t are accepted on the basis of a single stationary policy π_l.

Let τ_l denote the start time of episode l and $E_l = [\tau_l, \tau_{l+1}]$. Let $n_t(x) = \sum_{t'=1}^{t} \mathbb{1}_{\{x_{t'} = x\}}$ and $N(t) = |C_{t-1}| = \sum_{t'=1}^{t-1} \mathbb{1}_{\{a_{t'}=1\}}$. Let $N_j^i(t) = \sum_{t'=1}^{t-1} \mathbb{1}_{\{x_{t'}^i = j, a_{t'}=1\}}$ be the number of accepted candidates x such that $x^i = j$ before t.

At each episode l, the algorithm estimates the true candidate distribution by the empirical distribution $\hat{p}_l(x) = \frac{n_{\tau_l - 1}(x)}{\tau_l - 1}$ and maintains confidence sets B_l on p. As in UCRL2, these are built using the inequality on the ℓ_1-deviation of p and $\hat{p}_l$ from [25]:

Lemma 3. *With probability $\geq 1 - \frac{\delta}{3}$,*

$$\|\hat{p}_l - p\|_1 \leq \sqrt{\frac{2|\mathcal{X}| \log\left(6|\mathcal{X}|\tau_l(\tau_l - 1)/\delta\right)}{\tau_l - 1}} := \beta_l \tag{3}$$

Let $B_l = \{\tilde{p} \in \Delta(\mathcal{X}) : \|\hat{p}_l - \tilde{p}\|_1 \leq \beta_l\}$ be the confidence set for p at episode l. The associated set of compatible CMDPs is then $\{\tilde{M} = (\mathcal{X}, \mathcal{A}, \tilde{p}, r, \xi) : \tilde{p} \in B_l\}$. At the beginning of each episode, `RL-CMDP` finds the optimum of:

$$\max_{\pi \in \Pi, \tilde{p} \in B_l} \{g^{\tilde{p}, \pi} \,|\, \forall i, j,\, h_j^{i\,\tilde{p}, \pi} = 0\}. \tag{4}$$

Extended LP In order to optimize this problem, we re-write (4) as an extended LP. Following [21] and the CMDP literature, we introduce the state-action occupation measure $\mu(x, a) = \pi(x, a)p(x)$ and variables $\beta(x)$ to linearize the ℓ_1 constraint induced by the confidence set:

$$
\begin{aligned}
\max_{\substack{\mu \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}} \\ \beta \in \mathbb{R}^{\mathcal{X}}}} \quad & \sum_{x,a} \mu(x, a) r(x, a) \\
\text{u.c.} \quad & \mu \geq 0, \sum_{x,a} \mu(x, a) = 1 \\
& \forall x, \sum_a \mu(x, a) \leq \hat{p}_l(x) + \beta(x) \\
& \forall x, \sum_a \mu(x, a) \geq \hat{p}_l(x) - \beta(x) \\
& \forall x, a, \sum_y \beta(y) \leq \mu(x, a)\beta_l \\
& \forall i, j, \sum_{x,a} \mu(x, a)\xi_j^i(x, a) = 0.
\end{aligned}
\tag{5}
$$

The last constraint is the proportional representation constraint. The second to fourth constraints enforce the compatibility of μ with the ℓ_1 confidence set. We retrieve the distribution as $\tilde{p}_l(x) = \sum_a \mu(x, a)$, and the policy as:

$$\pi_l(x, a) = \begin{cases} \frac{\mu(x, a)}{\tilde{p}_l(x)} & \text{if } \tilde{p}_l \neq 0 \\ \frac{1}{2} & \text{otherwise}. \end{cases}$$

Precisely, if some $\tilde{p}_l(x) = 0$, we may set the policy $\pi_l(a|x)$ arbitrarily. Since the MDP induced by $\tilde{p}$ is still weakly communicating, and in particular any policy is unichain, the optimal gain in this CMDP is not affected.

We now provide regret and representativeness guarantees.

Theorem 1. *With probability $\geq 1 - \delta$, the regret of `RL-CMDP` satisfies:*

$$R(T) = O\left(\sqrt{|\mathcal{X}|T \log(|\mathcal{X}|T/\delta)}\right)$$
$$R^c(T) = O\left(\sqrt{|\mathcal{X}|T \log(|\mathcal{X}|T/\delta)}\right).$$
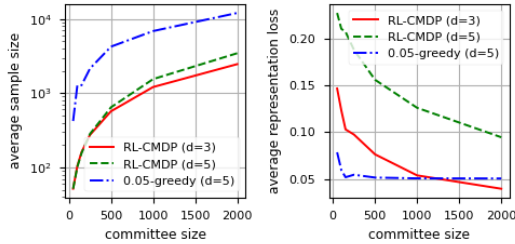
Figure 1: Effect of committee size K on sample complexity and representation loss for `RL-CMDP`, on data simulated from the UK Brexit Assembly, using 3 and 5 features. p is **unknown**.

Moreover, with probability $1 - \delta$, the representation loss of `RL-CMDP` *at horizon T satisfies:*

$$\|\lambda(C_T) - \rho\|_\infty = O\left(\frac{1}{g^*}\sqrt{\frac{|\mathcal{X}|\log\left(|\mathcal{X}|T/\delta\right)}{T}}\right).$$

The full proof is in Appendix B.2. It relies on decomposing regret over episodes, bounding the error on p which decreases over episodes as the confidence sets are refined, and leveraging martingale inequalities on the cumulative rewards.

Since $\frac{R(T)}{T} = g^* - \frac{N(T)}{T}$, it means that with high probability, the difference between the optimal selection rate and the selection rate of `RL-CMDP` decreases in $\sqrt{\log(T)/T}$ w.r.t. the horizon T. The representation loss decreases at the same speed, meaning that the agent should see enough candidates to accurately estimate p, and accept candidates at little cost for representativeness.

Compared to the bound from Proposition 2, the cost of not knowing p on representativeness is a $\sqrt{|\mathcal{X}|\log(|\mathcal{X}|)}$ factor. This is due to the estimation of p in the worst case, which is controlled by Lemma 3. As we show in our experiments (Sec. 6), the impact of $|\mathcal{X}|$ on performance regret (and in turn on sample complexity) is not problematic in our typical citizens' assembly scenario: since there are only a handful of features, our algorithm selects candidates quickly in practice (though representativeness is weakened by not knowing p). For specific structures of p, we obtain bounds with better scaling in $|\mathcal{X}|$, by controlling each entry of p with Bernstein bounds [18], instead the ℓ^1-norm. For completeness, we describe this alternative in Appendix C.

Interestingly, the representation loss is also inversely proportional to g^*, the optimal selection rate in the true CMDP. The reason is that the CMDP constraints do not control the ratios $\lambda_j^i(C_T) = \frac{N_j^i(T)}{N(T)}$, but $N_j^i(T)$ instead (by definition of $R^c(T)$ and ξ_j^i). If $N(T)$ is small, i.e. due to a small selection rate g, then $R_j^i(T) = |N_j^i(T) - \rho_j^i N(T)|$ is small, but not necessarily $|\frac{N_j^i(T)}{N(T)} - \rho_j^i|$: the committee is too small to be representative.

6   Experiments

The goal of these experiments is to answer the following: **(Q1)** In practice, for which range of committee sizes do our strategies achieve satisfying sample complexity and representation loss? **(Q2)** What is the cost of not knowing the distribution p for the sample complexity and representation loss?

Experimental setting   To answer these questions, we use summary data from the 2017 Citizens' Assembly on Brexit. The participants were recruited in an offline manner: vol-

unteers could express interest in a survey, and then 53 citizens were drawn from the pool of volunteers using stratified sampling, in order to construct an assembly that reflects the diversity of the UK electorate. We use summary statistics published in the report [20] to simulate an online recruitment process.

There are $d = 6$ features: the organisers expressed target quotas for 2 ethnicity groups, 2 social classes, 3 age groups, 8 regions, 2 gender groups and 2 Brexit vote groups (remain, leave). The report also includes the number of people contacted per feature group (e.g., women, or people who voted to remain) and the volunteering rate for each feature group, which we use as probability of volunteering given a feature group. We use Bayes' rule to compute the probabilities of feature groups among volunteers, and use them as the marginal distributions $\Pr[x^i = j|\text{volunteers}]$ (since we only consider the population of volunteers). Since we only have access to the marginals, we compute the joint distribution as if the features were independent, although our model is agnostic to the dependence structure of the joint distribution. In Appendix D.2, we present additional experiments with non-independent features, using a real dataset containing demographic attributes. The results are qualitatively similar.

We study `Greedy` with tolerance $\epsilon = 0.02, 0.05$. We run experiments for $K = 50, 100, 150, 250, 500, 1000$, averaged over 50 simulations. More details are found in App. D.1.

(A1) We compare `Greedy` and `CMDP`, when the distribution p is *known*. Figure 2 shows that the greedy strategy with $\epsilon = 0.05$ requires 10 times more samples than `CMDP`, and its representation loss is higher as soon as $K \geq 250$. `Greedy` with lower tolerance $\epsilon = 0.02$ achieves better representation than `CMDP` for smaller committees ($K \leq 100$), but the margin quickly decreases with K. However, even for small committees, it requires about 100 times more samples, which is prohibitively expensive. Figure 2 shows that for `CMDP`, the sample complexity grows linearly in the committee size, with a reasonable slope (we need to find $\tau \approx 500$ volunteers for a committee of size $K \approx 200$).

(A2) To corroborate the previously discussed effect of $|\mathcal{X}|$ when p is *unknown*, we evaluate `RL-CMDP` on different configurations: (1) using only the features ethnicity, social class, and gender ($d = 3, |\mathcal{X}| = 8$), (2) using all features except regions ($d = 5, |\mathcal{X}| = 48$). Fig. 1 shows that unlike `CMDP` which has full knowledge of p, it is for large committee sizes that `RL-CMDP` reaches low representation loss (below 0.05 for $K \geq 1500$ in the configuration(1)). This is because `RL-CMDP` needs to collect more samples to estimate p, as discussed in Th. 1. For known p, the CMDP approach achieves the same representativeness for middle-sized committees (repr. loss ≤ 0.05 for $K \approx 250$). Hence, comparing the cases of known (Fig. 2) and unknown distribution p (Fig. 1), the ignorance of p is not costly for sample complexity, but rather for the representation loss which decreases more slowly.

Consistently with Th. 1, we observe that the representation loss is higher when $\mathcal{X}$ is larger ($d = 5$). For small and middle-sized committees, the loss of `RL-CMDP` is much worse than `Greedy`'s which also works for unknown p. For large committees though, the margin is only 0.05 when $K \gtrsim 2000$ and $\tau \approx 3500$ for `RL-CMDP` (which is $\times 3$ more sample efficient than `Greedy`). In absolute terms, the theoretical regret bounds have a large constant $\sqrt{|\mathcal{X}|}$. This constant is likely unavoidable asymptotically because it comes from Lem. 3, but our experiments suggest that in the non-asymptotic regime, `RL-CMDP` performs better than the bound suggests.
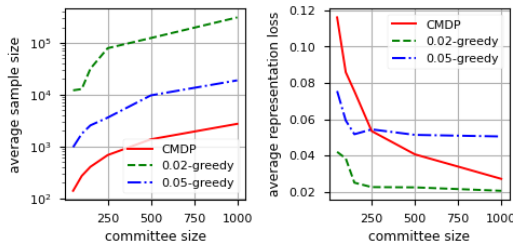
Figure 2: Effect of committee size K on sample complexity and representation loss for different strategies, in the UK Brexit Assembly experiment, using all features. p is **known**.

7 Conclusion

We formalised the problem of selecting a diverse committee with multi-attribute proportional representation in an online setting. We addressed the case of known candidate distributions with constrained MDPs, and leveraged exploration-exploitation techniques to address unknown distributions.

Among future research directions, we plan to consider (and learn) prior distributions over candidates represented succinctly, using Bayesian networks.

Acknowledgements

References

[1] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems*, pages 3450–3458, 2016.

[2] Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

[3] Haris Aziz. A rule for committee selection with soft diversity constraints. *Group Decision and Negotiation*, 28:1193–1200, 2019.

[4] Moshe Babaioff, Nicole Immorlica, David Kempe, and Robert Kleinberg. Online auctions and generalized secretary problems. *ACM SIGecom Exchanges*, 7(2):1–11, 2008.

[5] Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Streaming submodular maximization: Massive data summarization on the fly. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 671–680, 2014.

[6] MohammadHossein Bateni, Mohammadtaghi Hajiaghayi, and Morteza Zadimoghaddam. Submodular secretary problem and extensions. *ACM Transactions on Algorithms (TALG)*, 9(4):1–23, 2013.

[7] Xiaohui Bei, Shengxin Liu, Chung Keung Poon, and Hongao Wang. Candidate selections with proportional fairness constraints. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, pages 150–158, 2020.

[8] Gerdus Benadè, Paul Gölz, and Ariel D Procaccia. No stratification without representation. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 281–314, 2019.

[9] Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. Multiwinner elections with diversity constraints. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] L. Elisa Celis, Lingxiao Huang, and Nisheeth K. Vishnoi. Multiwinner voting with fairness constraints. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 144–151. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[12] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[13] Piotr Faliszewski, Piotr Skowron, Arkadii Slinko, and Nimrod Talmon. Multiwinner voting: A new challenge for social choice theory. *Trends in computational social choice*, 74:27–47, 2017.

[14] Bailey Flanigan, Paul Gölz, Anupam Gupta, and Ariel D. Procaccia. Neutralizing self-selection bias in sampling for sortition. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[15] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[16] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[17] Jérôme Lang and Piotr Skowron. Multi-attribute proportional representation. *Artificial Intelligence*, 263:74–106, 2018.

[18] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[19] Debmalya Panigrahi, Atish Das Sarma, Gagan Aggarwal, and Andrew Tomkins. Online selection of diverse results. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 263–272, 2012.

[20] A Renwick, S Allan, W Jennings, R McKee, M Russell, and G Smith. A considered public voice on brexit: The report of the citizens' assembly on brexit. 2017.

[21] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. *arXiv preprint arXiv:1905.07773*, 2019.

[22] Candice Schumann, Samsara N Counts, Jeffrey S Foster, and John P Dickerson. The diverse cohort selection problem. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 601–609. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[23] Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

[24] Julia Stoyanovich, Ke Yang, and HV Jagadish. Online set selection with fairness and diversity constraints. In *Proceedings of the EDBT Conference*, 2018.

[25] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

[26] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

[27] Liyuan Zheng and Lillian J Ratliff. Constrained upper confidence reinforcement learning. *arXiv preprint arXiv:2001.09377*, 2020.

Virginie Do
LAMSADE, PSL, Université Paris Dauphine, CNRS, France and Facebook AI Research
Email: `virginie.do@dauphine.eu`

Jamal Atif, Jérôme Lang
LAMSADE, PSL, Université Paris Dauphine, CNRS, France
Email: `first.last@lamsade.dauphine.fr`

Nicolas Usunier
Facebook AI Research
Paris, France
Email: `usunier@fb.com`

Algorithm 2: Greedy algorithm.

    input : tolerance ϵ, committee size K, targets ρ

    output: committee C_τ

1 $t \leftarrow 0$, $C_0 \leftarrow \emptyset$;

2 while $|C_t| < K$ **do**

3 $t \leftarrow t + 1$;

4   Observe $x_t \sim p$;

5   **if** $\forall i, j, N_j^i(t) + \mathbb{1}_{\{x_t^i = j\}} \leq \lceil \rho_j^i K \rceil + \frac{\epsilon K}{D_i - 1}$ **then**

6 $C_t \leftarrow C_{t-1} \cup \{x_t\}$;   // accept x_t

7 $\forall i, j, N_j^i(t-1) \leftarrow N_j^i(t) + \mathbb{1}_{\{x_t^i = j\}}$

8   **end**

9 end

10 return C_t

Algorithm 3: CMDP-based strategy.

    input : optimal policy π^* of (1), committee size K

    output: committee C_τ

1 $t \leftarrow 0$, $C_0 \leftarrow \emptyset$;

2 while $|C_t| < K$ **do**

3 $t \leftarrow t + 1$, observe $x_t \sim p$ and play $a_t \sim \pi^*(.|x_t)$;

4   **if** $a_t = 1$ **then** $C_t \leftarrow C_t \cup \{x_t\}$;

5 end

6 return C_t

A   Details of the algorithms

For precision, we provide the pseudocode of Greedy in Alg. 2, and the CMDP-based strategy in Alg. 3.

We also prove the bound on the representation loss of Greedy from Proposition 1 in Section 3.2.

Proof. For all i, j, we have by the if-condition and the termination condition:

$$\lambda_j^i(C_\tau) = \frac{N_j^i(\tau)}{K} \leq \frac{\lceil \rho_j^i K \rceil}{K} + \frac{\epsilon}{D_i - 1}$$

$$\leq \rho_j^i + \frac{1}{K} + \frac{\epsilon}{D_i - 1} \tag{6}$$

$$\leq \rho_j^i + \frac{D_i - 1}{K} + \epsilon. \tag{7}$$

For $i \in [d]$, for $j_0 \in [D_i]$, we have:

$$\rho_{j_0}^i = 1 - \sum_{j \neq j_0} \rho_j^i \,, \qquad\qquad \lambda_{j_0}^i(C_\tau) = 1 - \sum_{j \neq j_0} \lambda_j^i(C_\tau).$$

Combining these observations with (6):

$$\lambda_{j_0}^i(C_\tau) \geq 1 - \sum_{j \neq j_0} \left(\rho_j^i + \frac{1}{K} + \frac{\epsilon}{D_i - 1}\right)$$

$$= 1 - \sum_{j \neq j_0} \rho_j^i - \frac{D_i - 1}{K} - \epsilon$$

$$= \rho_{j_0}^i - \frac{D_i - 1}{K} - \epsilon.$$

Combining this lower bound with the upper bound (7), we have for all $i \in [d], j_0 \in [D_i]$, $\left|\lambda_{j_0}^i(C_\tau) - \rho_{j_0}^i\right| \leq \frac{D_i - 1}{K} + \epsilon$, which gives the result.

$\square$

B   Proofs

B.1   Proofs of Section 4

Proof of Lemma 1.

Proof. We have:

$$\sum_{x,a} \pi(x,a)p(x)r_j^i(x,a) = \mathbb{E}_{\substack{x \sim p \\ a \sim \pi(\cdot|x)}} \left[r_j^i(x,a)\right]$$

$$= \mathbb{P}^{p,\pi}[a = 1, x^i = j],$$

and $$g^\pi = \sum_{x,a} \pi(x,a)p(x)r(x,a) = \mathbb{E}_{\substack{x \sim p \\ a \sim \pi(.|x)}} [r(x,a)]$$

$$= \mathbb{P}^{p,\pi}[a = 1].$$

The ratio of these two quantities is equal to ρ_j^i by the last constraint of (2). It is also equal to $\mathbb{P}[x^i = j|a = 1]$, which gives the result.

Note that it also holds true for $j = [D_i]$, since

$$\mathbb{P}[x^i = D_i|a = 1] = 1 - \sum_{j' \in [D_i - 1]} \mathbb{P}[x^i = j'|a = 1] \qquad \text{and}$$

$$\rho_{D_i}^i = 1 - \sum_{j' \in [D_i - 1]} \rho_{j'}^i.$$

$\square$

Proof of Proposition 2.

Proof. For any $t > 0$, we have

$$\lambda_j^i(C_t) = \frac{\sum_{s=1}^t \mathbb{1}_{\{x_s^i = j, a_s = 1\}}}{\sum_{s=1}^t \mathbb{1}_{\{a_s = 1\}}}.$$

and by Lemma 1, we have:

$$\mathbb{E}\left[\mathbb{1}_{\{x^i = j\}}|a = 1\right] = \rho_j^i.$$

Let $\delta' > 0$. Conditionally on any $T \geq K, (a_1, ..., a_T) \in \{0, 1\}^T$ s.t. $a_1 + ... + a_T = K$ and $a_T = 1$, the draws of $x_t^i | a_t = 1$ are independent and thus, by Hoeffding's inequality [15], we have:

$$\mathbb{P}\left[|\lambda_j^i(C_T) - \rho_j^i| \geq \sqrt{\frac{\log(\frac{2}{\delta'})}{2N(T)}} \middle| a_1, ..., a_T\right] \geq 1 - \delta'$$

$$= \mathbb{P}\left[|\lambda_j^i(C_T) - \rho_j^i| \geq \sqrt{\frac{\log(\frac{2}{\delta'})}{2K}} \middle| a_1, ..., a_T\right].$$

Summing up over all such sequences $(a_1, ..., a_T)$, we obtain that:

$$\mathbb{P}\left[|\lambda_j^i(C_\tau) - \rho_j^i| \geq \sqrt{\frac{\log(\frac{2}{\delta'})}{2K}}\right] \geq 1 - \delta'.$$

The result follows from applying a union bound over all $i \in [d], j \in [D_i - 1]$ (there are $\tilde{d}$ such (i, j) pairs) and choosing $\delta' = \delta/\tilde{d}$.

$\square$

B.2 Proof of Theorem 1

The following lemma states a standard and useful inequality, which is similar to Lem. 19 in [16].

Lemma 4. *Recall that L is the random number of episodes ran by* `RL-CMDP` *up until horizon T. We have:*

$$\sum_{l=1}^{L} \frac{|E_l|}{\sqrt{\tau_l - 1}} \leq 2\sqrt{T}.$$

Proof. The proof is similar to that of Lem. 13 in [26]: we see E_l as the "derivative" of τ_l. Formally, let us define:

$$F(x) = \sum_{l=1}^{\lfloor x \rfloor} |E_l| + |E_{\lceil x \rceil}|(x - \lfloor x \rfloor)$$

$$f(x) := F'(x) = |E_{\lceil x \rceil}|.$$

We first observe that for any integer $l \in \mathbb{N}$, $f(l) = |E_l|$ and $F(l) = \tau_l$. Secondly, we have

$$F(x) \leq \sum_{l=1}^{\lfloor x \rfloor} |E_l| + |E_{\lceil x \rceil}| = \sum_{l=1}^{\lceil x \rceil} |E_l| = F(\lceil x \rceil),$$

and thus:

$$\frac{f(\lceil x \rceil)}{\sqrt{F(\lceil x \rceil) - 1}} \leq \frac{f(x)}{\sqrt{F(x) - 1}}.$$

We derive our bound as follows:

$$\sum_{l=1}^{L} \frac{|E_l|}{\sqrt{\tau_l - 1}} = \sum_{l=1}^{L} \frac{f(l)}{\sqrt{F(l) - 1}} = \int_{1}^{L} \frac{f(\lceil x \rceil)}{\sqrt{F(\lceil x \rceil) - 1}} \, dx$$

$$\leq \int_{1}^{L} \frac{f(x)}{\sqrt{F(x) - 1}} \, dx = 2(\sqrt{F(L) - 1})$$

$$= 2(\sqrt{\tau_L - 1}) \leq 2\sqrt{T}.$$

□

We introduce the following notation: for $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, let $f^{\pi}(x) := \sum_a f(x, a)\pi(x, a)$. For all $t > 0$, let l_t denote the episode number at time t. The following useful lemma is based on a martingale argument.

Lemma 5. *Let* $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. *Let* $\delta' > 0$. *We have:*

$$\mathbb{P}\left[\sum_{t=1}^{T}(\langle f^{\pi_{l_t}}, p \rangle - f(x_t, a_t)) \leq \sqrt{2T \log(1/\delta')}\right] \geq 1 - \delta'$$

$$\mathbb{P}\left[\sum_{t=1}^{T}\left|\langle f^{\pi_{l_t}}, p \rangle - f(x_t, a_t)\right| \leq \sqrt{2T \log(2/\delta')}\right] \geq 1 - \delta'.$$

Proof. We define the filtration $\mathcal{F}_t = \sigma(x_1, a_1, ..., x_t, a_t)$ and we first show that the sequence defined by $M_t = \langle f^{\pi_{l_t}}, p \rangle - f(x_t, a_t)$ is a martingale difference sequence w.r.t. $\mathcal{F}_t$. $\mathbb{E}[M_t] < \infty$ since the rewards are bounded. Next, the proof that $\mathbb{E}[M_t | \mathcal{F}_{t-1}] = 0$ relies on the fact that l_t, and in turn the stationary policy π_{l_t}, are $\mathcal{F}_{t-1}$−measurable.

Therefore,

$$\mathbb{E}\left[\langle f^{\pi_{l_t}}, p \rangle \big| \mathcal{F}_{t-1}\right] = \langle f^{\pi_{l_t}}, p \rangle.$$

We also have:

$$\mathbb{E}\left[f(x_t, a_t) \big| \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\sum_{x,a} f(x, a)\mathbb{1}_{\{(x_t, a_t)=(x,a)\}} \big| \mathcal{F}_{t-1}\right]$$

$$= \sum_{x,a} f(x, a)\pi_{l_t}(x, a) = \langle f^{\pi_{l_t}}, p \rangle.$$

Subtracting the two expressions above, we get $\mathbb{E}[M_t | \mathcal{F}_{t-1}] = 0$. $(M_t)_t$ is thus a Martingale difference sequence, such that $-1 \leq M_t \leq 1$. The result follows from Azuma-Hoeffding's inequality.

□

We now prove Theorem 1.

Proof. We define $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ to be the "good event", with:

$$\mathcal{E}_1 = \{\forall l \geq 1, \tilde{p}_l \in B_l\},$$

$$\mathcal{E}_2 = \{\sum_{t=1}^{T}(\langle r^{\pi_{l_t}}, p \rangle - r(x_t, a_t)) \leq \sqrt{2T \log(3/\delta)}\},$$

$$\mathcal{E}_3 = \left\{\forall i, j, \quad \sum_{t=1}^{T}\left|\langle \xi_j^{i \, \pi_{l_t}}, p \rangle - \xi_j^i(x_t, a_t)\right| \leq \sqrt{2T \log\left(\frac{6\tilde{d}}{\delta}\right)}\right\}.$$

By Lemma 3, we have

$$\mathbb{P}\left[\exists l \geq 1, \tilde{p}_l \in B_l\right] \geq 1 - \frac{\delta}{3}. \tag{8}$$

Combining (8) with Lemma 5 and using union bounds, $\mathbb{P}\left[\mathcal{E}\right] \geq 1 - \delta$. From now on, we assume that the good event $\mathcal{E}$ holds true.

Performance regret   We start by upper bounding the performance regret $R(T)$. Let $\Delta_l = \sum_{t \in E_l}(g^* - r(x_t, a_t))$ be the regret of episode l. Let $(\pi_l, \tilde{p}_l)$ be the solution of the optimistic CMDP (4) at episode l. Since (π^*, p) is feasible for (4), then $g^* \leq g^{\tilde{p}_l, \pi_l}$. We also note that

$$g^{\tilde{p}_l, \pi_l} = \sum_{x,a} r(x, a)\tilde{p}_l(x)\pi_l(x, a) = \sum_x r^{\pi_l}(x)\tilde{p}_l(x).$$

Therefore, we have:

$$\begin{aligned}
\Delta_l &\leq \sum_{t \in E_l}(g^{\tilde{p}_l, \pi_l} - r(x_t, a_t)) \\
&= \sum_{t \in E_l}\left(\sum_x r^{\pi_l}(x)\tilde{p}_l(x) - r(x_t, a_t)\right) \\
&= \sum_{t \in E_l}\sum_x r^{\pi_l}(x)(\tilde{p}_l(x) - p(x)) \\
&\quad + \sum_{t \in E_l}\left(\sum_x r^{\pi_l}(x)p(x) - r(x_t, a_t)\right)
\end{aligned} \tag{9}$$

Using Hölder's inequality and the fact that $\|r\|_\infty = 1$, the first term can be bounded by $|E_l|\|\tilde{p}_l - p\|_1$. By validity of the confidence intervals under event $\mathcal{E}$:

$$\|\tilde{p}_l - p\|_1 \leq 2\beta_l \leq \frac{2\sqrt{2|\mathcal{X}|\log\left(6|\mathcal{X}|T(T-1)/\delta\right)}}{\sqrt{\tau_l - 1}} \tag{10}$$

Summing up over episodes $l = 1, ..., L$:

$$\begin{aligned}
R(T) &\leq 2\sqrt{2|\mathcal{X}|\log\left(\frac{6|\mathcal{X}|T(T-1)}{\delta}\right)}\sum_{l=1}^{L}\frac{|E_l|}{\sqrt{\tau_l - 1}} \\
&\quad + \sum_{t=1}^{T}\left(\sum_x r^{\pi_{l_t}}(x)p(x) - r(x_t, a_t)\right).
\end{aligned} \tag{11}$$

We bound the first sum using Lemma 4. The second term can be bounded as in Lemma 5 because $\mathcal{E}_2$ holds true. This gives us the resulting bound which holds under $\mathcal{E}$:

$$R(T) \leq 4\sqrt{|\mathcal{X}|\log\left(\frac{6|\mathcal{X}|T(T-1)}{\delta}\right)T} + \sqrt{2T\log\left(\frac{3}{\delta}\right)}.$$

Cost of constraint violations   The proof for the cost of constraint violations is very similar. Let us bound $R_j^i(T) := \sum_{t=1}^{T}|\xi_j^i(x_t, a_t)|$ for all i, j. We briefly drop the sub/superscripts i, j.

At each episode l, since $(\pi_l, \tilde{p}_l)$ is a solution of (4), we have $h^{\tilde{p}_l, \pi_l} = 0$, and thus $\sum_{x,a} \xi(x, a)\pi_l(x, a)\tilde{p}_l(x) = \sum_x \xi^{\pi_l}(x)\tilde{p}_l(x) = 0$. Therefore, we have:

$$\left| \sum_{t=1}^{T} \xi(x_t, a_t) \right| = \left| \sum_{l=1}^{L} \big(\sum_{t \in E_l} \xi(x_t, a_t) - \sum_{x} \xi^{\pi_l}(x) \tilde{p}_l(x) \big) \right|$$

$$\leq \left| \sum_{l=1}^{L} \sum_{t \in E_l} \sum_{x} \xi^{\pi_l}(x)(p(x) - \tilde{p}_l(x)) \right.$$

$$+ \sum_{l=1}^{L} \big(\sum_{t \in E_l} \xi(x_t, a_t) - \sum_{x} \xi^{\pi_l}(x) p(x) \big) \Big|$$

$$\leq \sum_{l=1}^{L} \sum_{t \in E_l} \left| \sum_{x} \xi^{\pi_l}(x)(p(x) - \tilde{p}_l(x)) \right|$$

$$+ \left| \sum_{l=1}^{L} \big(\sum_{t \in E_l} \xi(x_t, a_t) - \sum_{x} \xi^{\pi_l}(x) p(x) \big) \right|$$

$$\leq \sum_{l=1}^{L} |E_l| \|\xi^{\pi_l}\|_{\infty} \|p - \tilde{p}_l\|_1$$

$$+ \left| \sum_{t=1}^{T} \Big(\xi(x_t, a_t) - \sum_{x} \xi^{\pi_{l_t}}(x) p(x) \Big) \right|,$$

where the first part of the last inequality is again by Hölder's inequality. Similarly to the performance regret, the first term is bounded using the validity of confidence intervals under the good event $\mathcal{E}$ and Lemma 4, and the second term is bounded by the martingale argument using Lemma 5. Hence, under $\mathcal{E}$ we have for any i, j:

$$R_j^i(T) \leq 4\sqrt{|\mathcal{X}| \log\big(\frac{6|\mathcal{X}|T(T-1)}{\delta}\big)T} + \sqrt{2T \log\big(\frac{6\tilde{d}}{\delta}\big)}.$$

And thus the same bounds holds for $R^c(T) = \max_{i,j} R_j^i(T)$.

Representation loss   We may now derive the bound on representation loss.

Let $f(T) = O\big(\sqrt{|\mathcal{X}| \log(|\mathcal{X}|T/\delta)}\big)$. The regret bounds imply that with $1 - \delta$:

$$R(T) = g^* T - N(T) \leq f(T) \Rightarrow N(T) \geq g^* T - f(T)$$

$$\frac{R^c(T)}{N(T)} = \max_{i,j} \left| \frac{N_j^i(T)}{N(T)} - \rho_j^i \frac{N(T)}{N(T)} \right| \leq \frac{f(T)}{N(T)}$$

$$\text{i.e., } \|\lambda(C_T) - \rho\|_{\infty} \leq \frac{f(T)}{N(T)}.$$

Therefore, using $N(T) \geq 1$, we have:

$$\|\lambda(C_T) - \rho\|_{\infty} \leq \frac{f(T)}{\max(1, g^* T - f(T))}$$

$$= O\left(\sqrt{\frac{|\mathcal{X}| \log(|\mathcal{X}|T/\delta)}{g^{*2} T}} \right).$$

$\square$

C Alternative to `RL-CMDP` with Bernstein bounds

We present `RL-CMDP-B`, an alternative to `RL-CMDP` which uses Bernstein empirical bounds [18].

At each episode l, the algorithm estimates the distributions by $\hat{p}_l(x) = \frac{n_{\tau_l-1}(x)}{\tau_l-1}$ and maintains confidence intervals $[\underline{p}_l(x), \overline{p}_l(x)]$. These are built using Bernstein's empirical inequality [18], which implies that there exists constants B_1, B_2 such that with probability $\geq 1 - \frac{\delta}{3}$, for each $l \geq 1$ and $x \in \mathcal{X}$,

$$|p(x) - \hat{p}_l(x)| \leq B_1 \sqrt{\frac{\hat{\sigma}_l^2(x) \log(\frac{6|\mathcal{X}|\tau_l}{\delta})}{1 \wedge (\tau_l - 1)}} + B_2 \frac{\log(\frac{6|\mathcal{X}|\tau_l}{\delta})}{1 \wedge (\tau_l - 1)}, \tag{12}$$

where $\hat{\sigma}_l(x) = \sqrt{\hat{p}_l(x)(1 - \hat{p}_l(x))}$.

Following e.g. [12], we re-write (4) as an extended LP by introducing the state-action occupation measure $\mu(x,a) = \pi(x,a)p(x)$.

$$\max_{\mu \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}} \quad \sum_{x,a} \mu(x,a)r(x,a)$$

$$\text{u.c.} \quad \mu \geq 0, \sum_{x,a} \mu(x,a) = 1$$

$$\forall x, \sum_a \mu(x,a) \leq \overline{p}_l(x) \tag{13}$$

$$\forall x, \sum_a \mu(x,a) \geq \underline{p}_l(x)$$

$$\forall i,j, \sum_{x,a} \mu(x,a)\xi_j^i(x,a) = 0.$$

The second to fourth constraints enforce the compatibility of μ with the confidence intervals. Controlling each entry of p with Bernstein bounds instead of the ℓ^1-norm allows for a simpler optimization problem than the extended LP (5). We get the following regret bound:

Theorem 2 (Regret guarantees). *With probability $\geq 1 - \delta$, the regret of `RL-CMDP-B` satisfies:*

$$R(T) = O\big(\sqrt{|\mathcal{X}|T \log(|\mathcal{X}|T/\delta)} + |\mathcal{X}| \log(|\mathcal{X}|T/\delta)^2\big)$$
$$R^c(T) = O\big(\sqrt{|\mathcal{X}|T \log(|\mathcal{X}|T/\delta)} + |\mathcal{X}| \log(|\mathcal{X}|T/\delta)^2\big).$$

With probability $\geq 1 - \delta$, the representation loss satisfies:

$$\|\lambda(C_T) - \rho\|_\infty$$
$$= O\left(\frac{1}{g^*}\sqrt{\frac{|\mathcal{X}| \log\left(|\mathcal{X}|T/\delta\right)}{T}} + \frac{|\mathcal{X}| \log(|\mathcal{X}|T/\delta)^2}{g^*T}\right).$$

When using Bernstein bounds, the representation loss carries $O(|\mathcal{X}| \log(|\mathcal{X}|T/\delta)^2)$. This factor but has a bigger scaling with $|\mathcal{X}|$, but decreases rapidly in $\frac{\log(T)^2}{T}$.

The Bernstein version of `RL-CMDP` may be advantageous for some candidate distributions p. For example, if the support $\mathcal{S}$ of p is very small compared to $\mathcal{X}$, the first term in the

Bernstein empirical inequality (12) is equal to zero for all x outside the support. Therefore, the representation loss scales as:

$$\|\lambda(C_T) - \rho\|_\infty$$

$$= O\left(\frac{1}{g^*}\sqrt{\frac{|\mathcal{S}|\log\left(|\mathcal{S}|T/\delta\right)}{T}} + \frac{|\mathcal{X}|\log(|\mathcal{X}|T/\delta)^2}{g^*T}\right),$$

where $|\mathcal{S}| \ll |\mathcal{X}|$. Thus, the second term with fast decrease in $\frac{\log(T)^2}{T}$ controls the bound on representation loss.

C.1 Proofs

The following lemma states a useful inequality akin to Lemma 4.

Lemma 6. *We have:*

$$\sum_{l=1}^{L} \frac{|E_l|}{\tau_l - 1} \le \log(T)$$

Proof. The proof is similar to Lem. 13 in [26]. Using the same notation as in the proof of Lemma 4,

$$\sum_{l=1}^{L} \frac{|E_l|}{\tau_l - 1} = \sum_{l=1}^{L} \frac{f(l)}{F(l) - 1} = \int_1^L \frac{f(\lceil x \rceil)}{F(\lceil x \rceil) - 1}\, dx$$

$$\le \int_1^L \frac{f(x)}{F(x) - 1}\, dx = \log(F(L) - 1)$$

$$= \log(\tau_L - 1) \le \log T.$$

$\square$

We now prove Theorem 2.

Proof. We re-use the same steps and notation as for the proof of Theorem 1.

Here instead, $\mathcal{E}_1$ is the event such that the confidence intervals are valid (12). Under the high-probability good event $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we thus have:

$|\pi_l(x) - p(x)| \lesssim \sqrt{\frac{\hat{p}_l(x)(1-\hat{p}_l(x))b_{\delta,T}}{\tau_l - 1}} + \frac{b_{\delta,T}}{\tau_l - 1}$

where $b_{\delta,T} = \log(\frac{6|\mathcal{X}|T}{\delta})$.

In the following, the first inequality is by validity of the Bernstein confidence intervals

under $\mathcal{E}$, and the second inequality is by Cauchy-Schwarz's inequality:

$$
\sum_{t \in E_l} \sum_x r^{\pi_l}(x)(\tilde{p}_l(x) - p(x))
$$

$$
\leq \sum_{t \in E_l} \sum_x r^{\pi_l}(x) \sqrt{\frac{\hat{p}_l(x)(1 - \hat{p}_l(x))b_{\delta,T}}{\tau_l - 1}}
$$

$$
+ \sum_{t \in E_l} \frac{b_{\delta,T}}{\tau_l - 1} \underbrace{\sum_x r^{\pi_l}(x)}_{\leq |\mathcal{X}|}
\tag{14}
$$

$$
\leq \sum_{t \in E_l} \sqrt{\underbrace{\left(\sum_x 1 - \hat{p}_l(x)\right)}_{\leq |\mathcal{X}|} \underbrace{\left(\sum_x \hat{p}_l(x) r^{\pi_l}(x) b_{\delta,T}\right)}_{\leq b_{\delta,T}}} \sqrt{\frac{1}{\tau_l - 1}}
$$

$$
+ \sum_{t \in E_l} \frac{|\mathcal{X}| b_{\delta,T}}{\tau_l - 1}
$$

By Lemmas 4 and 6, we have:

$$
\sqrt{|\mathcal{X}| b_{\delta,T}} \sum_{l=1}^{L} \frac{|E_l|}{\sqrt{\tau_l - 1}} \leq 2\sqrt{|\mathcal{X}| b_{\delta,T} T}
$$

$$
|\mathcal{X}| b_{\delta,T} \sum_{l=1}^{L} \frac{|E_l|}{\tau_l - 1} \leq |\mathcal{X}| b_{\delta,T} \log(T).
$$

Summing up over episodes in inequality (14) and plugging in the above inequalities gives the desired bound by following the steps of the proof of Theorem 1. $\qquad\square$

D Experiments

D.1 Details on the Brexit experiments

We provide in Table 2 the target vectors $(\rho_j^i)_{i,j}$ and marginal distributions $(\mathbb{P}^p[x^i = j])_{i,j}$ extracted from the Citizens' Assembly on Brexit report [20].[3] The report includes the volunteering rates for each feature group, i.e. $\Pr[\text{volunteer}|x^i = j]$. To compute the marginal distributions $(\Pr[x^i = j|\text{volunteer}])_{i,j}$, we thus use Bayes' rule to compute the probability of each feature group among the volunteer population[4], that is:

$$
\mathbb{P}^p[x^i = j] = \Pr[x^i = j|\text{volunteer}]
$$
$$
= \frac{\Pr[\text{volunteer}|x^i = j]\Pr[x^i = j]}{\Pr[\text{volunteer}]}.
$$

We often have $\rho_j^i \neq \mathbb{P}^p[x^i = j]$. For example, compared to the age target, we are less likely to find younger people (≤ 34 years old) among volunteers. For gender, while the target was gender parity, we are much less likely to find women than men in the volunteer population.

---

[3] https://citizensassembly.co.uk/wp-content/uploads/2017/12/Citizens-Assembly-on-Brexit-Report.pdf, pages 28-32.

[4] In doing so, we notice that the probability of finding non-voter volunteers is almost zero, hence we only consider "remain" and "leave" for the feature Brexit vote. Indeed, the report states "The only target that proved impossible to meet was that for non-voters in the 2016 referendum." p.28.

	Targets	Marginals
Ethnicity	0.860 / 0.140	0.863 / 0.136
Social class	0.550 / 0.450	0.556 / 0.444
Age	0.288 / 0.344 / 0.367	0.154 / 0.432 / 0.414
Region	0.233 / 0.160 / 0.093 / 0.134 / 0.222 / 0.047 / 0.082 / 0.028	0.179 / 0.155 / 0.090 / 0.117 / 0.211 / 0.073 / 0.154 / 0.021
Gender	0.507 / 0.493	0.384 / 0.616
Brexit vote	0.481 / 0.519	0.565 / 0.434

Table 2: Target quotas (from the report) and marginal distribution (computed using Bayes' rule) for the Brexit experiment.

For our experiments presented in Section 6, we used Python and the CPLEX LP solver, and a machine with Intel Xeon Gold 6230 CPUs, 2.10 GHz, 1.3 MiB of cache.

D.2 Experiments with dependent features

The goal of these experiments is to answer the following: what is the impact of the dependence structure of the joint feature distribution p on the sample complexity and representation loss of our algorithms? Since we may only retrieve marginal distributions from the Citizen's Assembly on Brexit report, we keep the target quotas on each feature but simulate joint feature distributions from another dataset with demographic attributes, the standard Adult Census Income dataset [11].

The Adult dataset consists of approximately 49.000 entries of subjects in the US, each with 14 demographic features and a binary label indicating whether a subject's income is above or below 50K USD. We only keep features that can be mapped to our Brexit Citizen's Assembly example: gender, age, ethnicity and income, which we use in lieu of social class. We do not consider proportional representation for region and Brexit vote since there are no such features in the Adult dataset. In our preprocessing of the Adult dataset, we create the same three age categories (<35, 35-54, >54), the same two ethnicity groups (white / non-white) and we use the binary income variable as a proxy for social class, by assigning $> 50K$ to upper class and $\leq 50K$ to lower class. This leaves us with 4 features with $2, 2, 2, 3$ possible values.

To create dependencies between features, we consider two graphical structures shown in Figure 3, and for each we fit a Bayesian network to the dataset to generate a model of the joint distribution $p(x)$. We consider one structure with little dependence, and one structure with strong dependence between features.

Figure 4 shows that both when p is known, the sample complexity is higher when there is more dependence (Bayesian network (2)) between features, but the representation loss is the same. When p is unknown, the representation loss is lower for structure (2) with more dependence, than structure (1) with little dependence, but the sample size is higher for (2). For structure (2), the representation loss is low (≈ 0.07) for $K = 1000$. Importantly, it implies that in practice, the representation loss is much lower than the worst case bound given by Theorem 1.

E   Detailed example for Section 4

We take the same attributes and same distribution as in Table 1, with $\epsilon' = {}^1\!/{}_6$:

(a) Structure 1: weak dependence. (b) Structure 2: strong dependence.

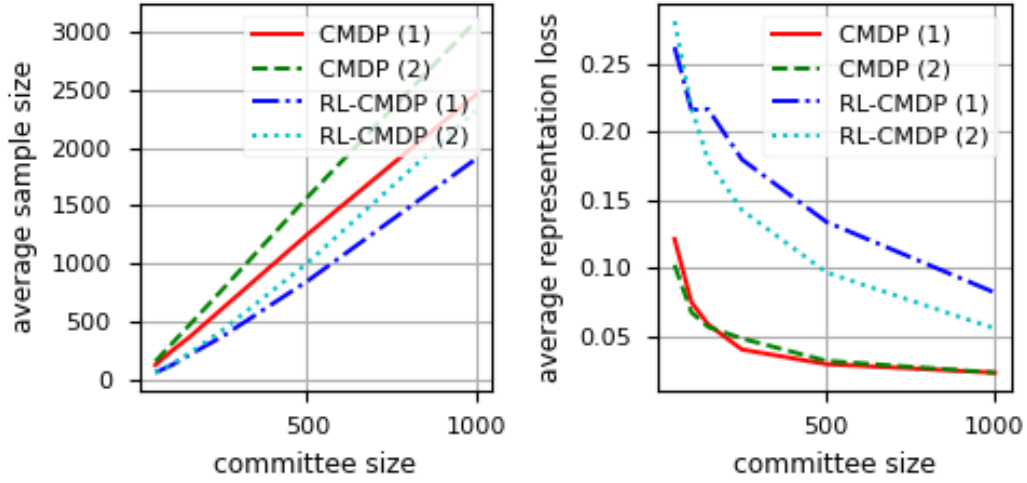Figure 3: Bayesian network structures for the Census Income dataset.

Figure 4: Effect of committee size K on sample complexity and representation loss for CMDP (**known** p) and RL-CMDP (**unknown** p), on the two different Bayesian networks (1) and (2) fitted on the Census Income dataset.

The target vectors are $\rho^{\text{gender}} = (1/2, 1/2)$ and $\rho^{\text{age}} = (1/2, 1/2)$, that is, an ideal committee contains as many women as men and as many senior than junior.

We solve the linear program

$$
\max \quad \frac{\pi(MS,1)}{3} + \frac{\pi(FS,1)}{4} + \frac{\mu(MJ,1)}{4} + \frac{\mu(FJ,1)}{6}
$$

$$
\text{u.c.} \quad \frac{\pi(MS,1)}{3} + \frac{\pi(FS,1)}{4} = \frac{\mu(MJ,1)}{4} + \frac{\mu(FJ,1)}{6}
$$

$$
\frac{\pi(MS,1)}{3} + \frac{\mu(MJ,1)}{4} = \frac{\pi(FS,1)}{4} + \frac{\mu(FJ,1)}{6}
$$

(15)

Its solution is

gender \ age	S	J
M	$^1/_3$	$^1/_4$
F	$^1/_4$	$^1/_6$

$$\pi^*(MS, 1) = {}^1/_2$$
$$\pi^*(FJ, 1) = 1$$
$$\pi^*(MJ, 1) = 1 \tag{16}$$
$$\pi^*(FS, 1) = 1$$

Thus, each time the current volunteer is a senior male, we select him with probability $^1/_2$; all other volunteers are selected with probability 1. The expected final composition of the pool is 30% of junior male, 30% of senior female, 20% of junior female and 20% of senior male. As the policy selects in average $^5/_6$ of the volunteers, the expected time until we select K candidates is $\mathbb{E}^{p,\pi^*}[\tau] = (^6/_5)K$.