Decomposition-Coordination Method for Finite Horizon Bandit Problems

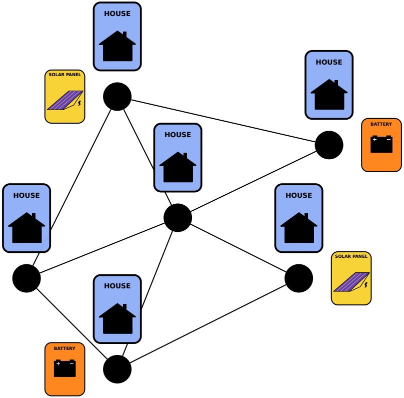
M. De Lara[1], B. Heymann[2], J.-P. Chancelier[1]

Workshop on restless bandits
IMAG, Grenoble
20-21 November 2023

[1] CERMICS, École des Ponts, Marne-la-Vallée, France
[2] Criteo, Paris, France

Motivation

We consider a *peer-to-peer* microgrid where houses exchange energy, and we formulate it as a large-scale stochastic optimization problem



How to manage such network in an (almost) optimal way?

Mix of spatial and temporal decompositions

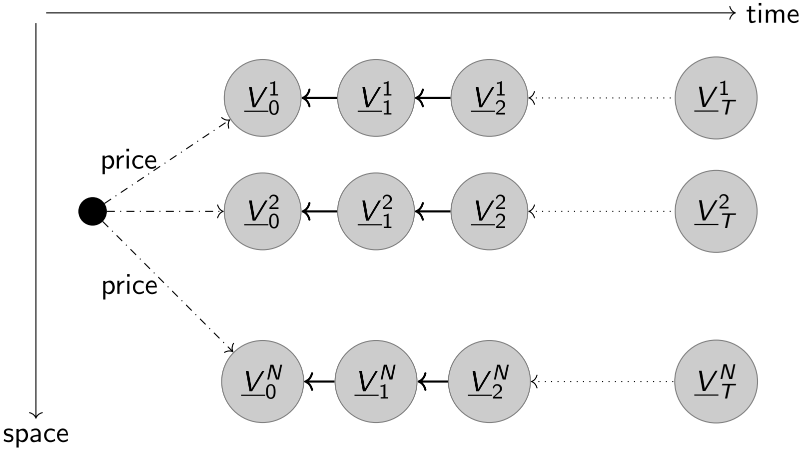
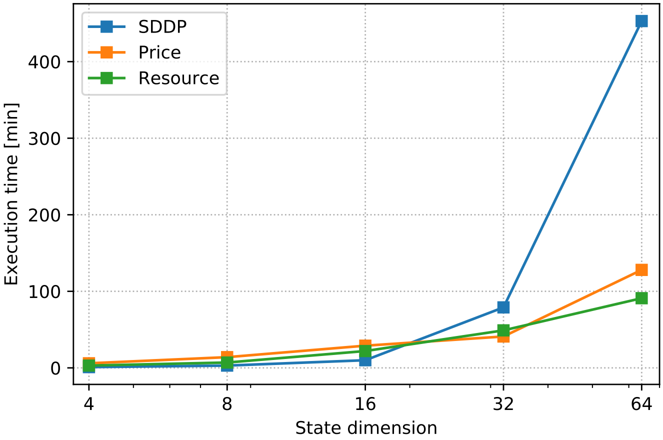


Figure: The case of price decomposition

Increase in execution time with state dimension



Outline of the presentation

Dynamic programming and arm decomposition

Numerical results

Conclusion

Appendix: The DeCo algorithm

Appendix: Bayesian bandits as multistage stochastic optimization

Outline of the presentation

Outline of the presentation

Multistage stochastic optimal control formulation

- Let $T \geq 1$ be an integer (finite) representing the horizon
- At each discrete time stage $t \in [\![0, T-1]\!]$, a decision-maker (DM) makes a decision and gets a reward as follows
  - At the beginning of the time interval $[t, t+1[$, the DM selects an arm $a \in A$ (finite set)
  - At the end of the time interval $[t, t+1[$, the arm a delivers a random variable $\mathbf{W}^a_{t+1} \in \{\mathtt{B}, \mathtt{G}\}$, ("bad" $\mathtt{B}$, "good" $\mathtt{G}$)
- The corresponding probabilities are unknown to the DM

$$p^a = (p^{\mathtt{B}a}, p^{\mathtt{G}a}) = \left(\mathbb{P}\{\mathbf{W}^a_{t+1} = \mathtt{B}\}, \mathbb{P}\{\mathbf{W}^a_{t+1} = \mathtt{G}\}\right) \in \Sigma$$

where $\Sigma = \{p = (p^{\mathtt{B}}, p^{\mathtt{G}}) \in \mathbb{R}^2_+ \mid p^{\mathtt{B}} + p^{\mathtt{G}} = 1\}$
is the one-dimensional simplex

- We suppose that the DM holds a prior beta distribution $\pi^a_0 = \beta(n^{\mathtt{B}a}, n^{\mathtt{G}a})$ over the unknown $p^a = (p^{\mathtt{B}a}, p^{\mathtt{G}a}) \in \Sigma$

Decision model for arm selection

- We consider a sequence $\mathbf{U} = \{\mathbf{U}_t\}_{t \in [\![0, T-1]\!]}$ of r.v on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where
  - $\mathbf{U}_t = \{\mathbf{U}_t^a\}_{a \in A}$, $\forall t \in [\![0, T-1]\!]$
  - $\mathbf{U}_t^a \in \{0, 1\}$, $\forall a \in A$, $\forall t \in [\![0, T-1]\!]$
- Values $\mathbf{U}_t^a \in \{0, 1\}$ represent that,
  at the beginning of the time interval $[t, t+1[$,
  - either arm a has been selected ($\mathbf{U}_t^a = 1$)
  - or arm a has not been selected ($\mathbf{U}_t^a = 0$)
- Since, at each given time, one and only one arm has to be selected, we add the constraint

$$\sum_{a \in A} \mathbf{U}_t^a = 1 \;, \quad \forall t \in [\![0, T-1]\!]$$

This way of modeling the selection of a unique arm
is not the most common in the bandit literature

Multistage stochastic optimal problem

We formulate a maximization problem

$V_0(\pi_0) = V_0\Big(\big(n_0^{\mathtt{B}a}\big)_{a\in A}, \big(n_0^{\mathtt{G}a}\big)_{a\in A}\Big) =$

$$\sup \underbrace{\int_{\Delta(\Sigma)^A} \prod_{a\in A} \overbrace{\pi_0^a(\mathrm{d}p^a)}^{\beta(n_0^{\mathtt{B}a}, n_0^{\mathtt{G}a})(\mathrm{d}p^a)} \mathbb{E}_{\{p^a\}_{a\in A}}}_{\text{probabilistic model}} \Big[\sum_{t=0}^{T-1} \sum_{a\in A} \overbrace{\mathbf{U}_t^a}^{\text{decision}} \underbrace{L_t^a(\mathbf{W}_{t+1}^a)}_{\text{reward}}\Big]$$

The supremum is taken over $\mathbf{U} = \{\mathbf{U}_t^a\}_{a\in A, t\in [\![0, T-1]\!]}$
subject to constraints $(\forall t \in [\![0, T-1]\!])$

$$\sum_{a\in A} \mathbf{U}_t^a = 1 \quad \text{(only one arm is selected)}$$

$$\underbrace{\sigma(\mathbf{U}_t) \subset \sigma\Big(\mathbf{U}_0, \{\mathbf{U}_0^a \mathbf{W}_1^a\}_{a\in A}, \dots, \mathbf{U}_{t-1}, \big\{\mathbf{U}_{t-1}^a \mathbf{W}_t^a\big\}_{a\in A}\Big)}_{\text{inclusion of } \sigma\text{-algebras}} \quad \text{(information)}$$

Outline of the presentation

Dynamic programming and arm decomposition (1/2)

By weak duality — for the the coupling constraint $\sum_{a \in A} \mathbf{U}_t^a = 1$ with a deterministic multiplier μ_t — we obtain the upper bound

$$V_0\big((n_0^{\mathrm{Ba}})_{a \in A}, (n_0^{\mathrm{Ga}})_{a \in A}\big) \leq \inf_{\mu \in \mathbb{R}^T} \left(\sum_{a \in A} \underbrace{V_0^a[\mu](n_0^{\mathrm{Ba}}, n_0^{\mathrm{Ga}})}_{\text{arm } a \text{ value function}} + \sum_{t=0}^{T-1} \mu_t \right)$$

where, for any vector $\mu = \{\mu_t\}_{t \in [\![0, T-1]\!]} \in \mathbb{R}^T$ of multipliers,

$$V_T^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}}) = 0 \ , \quad \forall (n^{\mathrm{Ba}}, n^{\mathrm{Ga}}) \in \mathbb{N} \times \mathbb{N}$$

$$V_t^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}}) = \max \Big\{ V_{t+1}^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}}), -\mu_t + \frac{n^{\mathrm{Ba}}}{n^{\mathrm{Ba}} + n^{\mathrm{Ga}}} \big(L_t^a(\mathrm{B}) + V_{t+1}^a[\mu](n^{\mathrm{Ba}} + 1, n^{\mathrm{Ga}}) \big) + \frac{n^{\mathrm{Ga}}}{n^{\mathrm{Ba}} + n^{\mathrm{Ga}}} \big(L_t^a(\mathrm{G}) + V_{t+1}^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}} + 1) \big) \Big\}$$

Dynamic programming and arm decomposition (2/2)

- The global stochastic optimal control problem $V_0(\pi_0)$ is, theoretically, solvable by dynamic programming using value functions $\{V_t\}_{t\in[\![0,T]\!]} : \prod_{a\in A}\Delta(\Sigma) \to \mathbb{R}\cup\{+\infty\}$
- However, computing $V_0(\pi_0)$ using Dynamic Programming faces the curse of dimensionality, as the priors are of the form $\pi_0 = \{\pi_0^a\}_{a\in A} \in \prod_{a\in A}\Delta(\Sigma)$
- The DECO algorithm consists in replacing

$$V_{t+1}\big(\{\pi^a_{t+1}\}_{a\in A}\big) \rightsquigarrow \sum_{a\in A} V^a_{t+1}[\mu](\pi^a_{t+1})$$

for a suitable vector $\mu \in \mathbb{R}^T$ in order compute a policy by

$$\mathcal{U}_t(\pi_t) \in \operatorname*{arg\,max}_{\substack{u_t=\{u^a_t\}_{a\in A}\in\{0,1\}^A \\ \sum_{a\in A} u^a_t=1}} \Big(\tilde{L}_t(\pi_t, u_t) + \int_{\Delta(\Sigma)} \overbrace{V_{t+1}(\pi_{t+1})}^{\sum_{a\in A} V^a_{t+1}[\mu](\pi^a_{t+1})} k_t(\mathrm{d}\pi_{t+1} \mid \pi_t, u_t)\Big)$$

The DeCo algorithm as a nonstationary index policy

For a suitable value of μ, when the state of the multi-armed bandit is given by $(n^{\mathtt{B}a}, n^{\mathtt{G}a})_{a \in A}$ at stage t,
the DeCo algorithm selects an arm

$$
\begin{aligned}
& \mathcal{A}_t^*[\mu]\Big(\big\{(n^{\mathtt{B}a}, n^{\mathtt{G}a})\big\}_{a\in A}\Big) \in \arg\max_{a\in A} \Big[\\
\text{(expected reward)} \quad & \frac{n^{\mathtt{B}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}} L_t^a(\mathtt{B}) + \frac{n^{\mathtt{G}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}} L_t^a(\mathtt{G}) \\
& + \\
\text{(value} \quad & \frac{n^{\mathtt{B}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}} V_{t+1}^a[\mu](n^{\mathtt{B}a}+1, n^{\mathtt{G}a}) + \\
\text{of} \quad & \frac{n^{\mathtt{G}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}} V_{t+1}^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}+1) \\
\text{information)} \quad & - V_{t+1}^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}) \quad \Big]
\end{aligned}
$$

Outline of the presentation

Outline of the presentation

Numerical experiments (small number $|A|$ of arms)

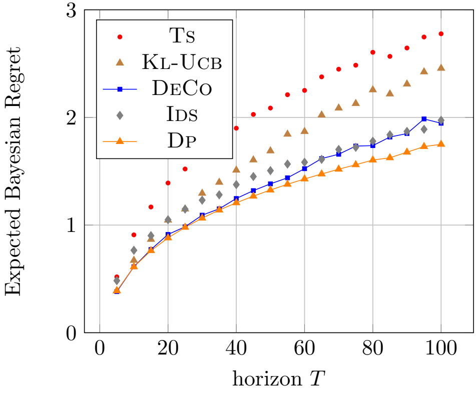
The DeCo algorithm compared to
brute force Bf algorithm (global DP)
for small number $|A|$ of arms and short horizon T

arms $\|A\|$	horizon T	DeCo	Bf
3	10	6.411	6.409
3	20	13.458	13.465
5	10	6.645	6.659

Comparison in term of
estimated total expected reward (higher is better)

Numerical experiments (small number $|A|$ of arms)

The DeCo algorithm compared to Bf algorithm and others for $|A| = 2$ arms and horizon T up to 100



Comparison in term of (to be defined later)
Expected Bayesian Regret (lower is better)

Outline of the presentation

Numerical experiments for larger number $|A|$ of arms

BF cannot be used anymore because of the curse of dimensionality

FH-GITTINS is used as a proxy supposed to be close to the optimal solution

arms $\lvert A \rvert$	horizon T	DECO	FH-GITTINS
5	20	14.21	14.28
5	40	29.85	30.06
15	20	14.59	14.67
15	40	31.54	31.63

Comparison in term of
estimated total expected reward (higher is better)

The performance of DECO is close to the optimal solution while keeping the computational cost reasonable

Numerical experiments: comparison with other methods

- We then tested DeCo against
  - Thomson Sampling (Ts) [10, 2]
  - Kullback-Leibler upper-confidence bound (Kl-Ucb) [3]
  - Information-Directed Sampling (IDS) [9][3]
  - Finite Horizon Gittins index (FH-Gittins) [6, 8, 7]
  - In the case of two arms, exact DP
- The solutions $\mathbf{U} = \{\mathbf{U}_t^a\}_{a \in A, t \in [\![0, T-1]\!]}$ are compared using the Expected Bayesian Regret given by

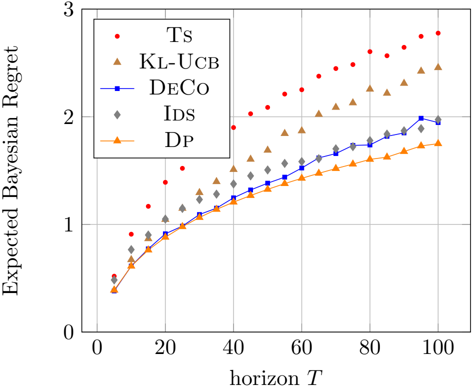
$$\mathcal{R}(\mathbf{U}) = \int_{\Delta(\Sigma)^A} \prod_{a \in A} \pi_0^a(\mathrm{d}p^a) \left\{ \mathbb{E}_{\{p^a\}_{a \in A}} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \left(\mathbf{U}_t^{\mathrm{BA},a} - \mathbf{U}_t^a \right) \mathbf{W}_{t+1}^a \right] \right\}$$

  - L_t^a equal to 1 on $\mathtt{G}$ and 0 on $\mathtt{B}$
  - Ba: best arm policy is, for all $a \in A$, given by $\mathbf{U}_t^{\mathrm{BA},a} = 1 \iff a \in \arg\max_{a' \in A} p_{\mathtt{G}}^{a'}$
  - the prior is supposed to be the uniform distribution for all arms

[3] For IDS we used the library [1]

Numerical experiments: comparison with other methods

The two arms case where brute force algorithm can be used

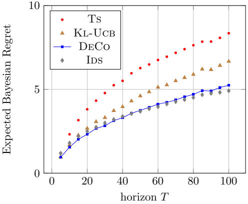


Comparison in term of
Expected Bayesian Regret (lower is better)

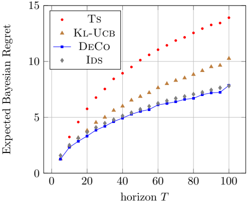
Numerical experiments: comparison with other methods

Increasing the number of arms

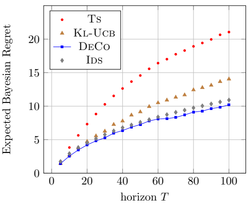
(a) 5 arms



(b) 10 arms



(c) 20 arms



On all cases, DeCo

- beats both Ts and Kl-Ucb with a comfortable margin
- and is comparable to IDS

Numerical experiments: comparison with other methods

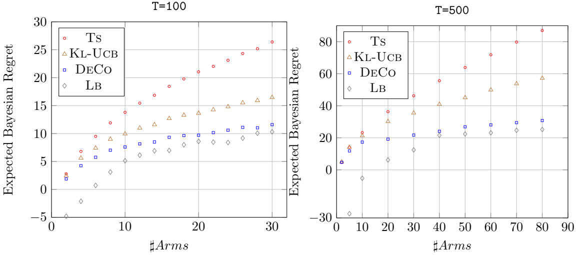
- On all cases, DeCo beats both Ts and Kl-Ucb with a comfortable margin, and is comparable to IDS
- For the two arms case DeCo is very close to the optimal solution, computed by DP
  (we used the Julia BinaryBandit library)
- Expected Bayesian Regret is numerically obtained by Monte Carlo simulations
  - Expectation with respect to the prior:
    a sample of size 1000
  - Expectation with respect to the arms parameters:
    a sample of size 1000 or of size 100 (for large T)
  - Same samples for all the evaluated policies

Numerical experiments: DeCo provides a lower bound

- $\mathcal{R}^{\text{LB}}$: lower bound provided by DeCo (using the dual bound)

$$\mathcal{R}(\mathbf{U}) \geq \mathcal{R}^{\text{LB}} = \frac{|A|}{|A|+1} T - \left(\sum_{a \in A} V_0^a[\mu^*](\pi_0^a) + \sum_{t=0}^{T-1} \mu_t^* \right)$$

- $\mathcal{R}^{\text{LB}}$, DeCo, Ts and Kl-Ucb regret as a function of the number of arms for $T = 100$ and $T = 500$



The lower bound is of no use (lower than 0) for $|A| \leq 5$

When $|A| \uparrow$ the regrets of DeCo and $\mathcal{R}^{\text{LB}}$ become quite close, which indicates that DeCo is close to being optimal

Outline of the presentation

Conclusion (pros)

- The numerical results illustrate the value of the decomposition-coordination approach (observed in other applications): DeCo is a simple algorithm and its performances are close to the optimal Bayesian solution for several configurations of arms and horizons, while keeping the computing time reasonable
- Empirically, DeCo offers performances comparable to FH-Gittins but with a much smaller computation burden
- DeCo can deal with time varying reward functions, and can even include a final reward
- In particular, DeCo can be applied to nonstationary settings, whereas FH-Gittins cannot

Conclusion (cons)

- As of now, the approach main limitation is that the horizon T is supposed to be known in advance and to be reasonably small, whereas many multi-armed bandit algorithms do not require T as an input
- In addition, the usage of dynamic programming might make DeCo too burdensome for some applications with long horizon T
- Also, since the DeCo algorithm requires a Bayesian prior, the question of the impact of a wrong prior on the performance is left open

Conclusion (perspectives)

- We could explore the possibility to adapt the multiplier μ as time goes on and we receive bandit feedback
- Further works include
  - a theoretical analysis of the DeCo policy
  - and an extension to the discounted infinite horizon case
  - as well as adapting the heuristic to other use cases

References

[1] D. Baudry, Y. Russac, and A. Filiot. Information_directed_sampling. https://github.com/DBaudry/, 2019.

[2] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[3] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[4] J. C. Gilbert and X. Jonsson. LIBOPT – An environment for testing solvers on heterogeneous collections of problems, 2007.

[5] J. Gittins and Y.-G. Wang. The Learning Component of Dynamic Allocation Indices. *The Annals of Statistics*, 20(3):1625 – 1636, 1992.

[6] E. Kaufmann. On Bayesian index policies for sequential resource allocation. *The Annals of Statistics*, 46(2):842 – 865, 2018.

[7] T. Lattimore. Regret analysis of the finite-horizon gittins index strategy for multi-armed bandits. In *Conference on Learning Theory*, pages 1214–1245. PMLR, 2016.

[8] J. Nino-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.

[9] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[10] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[11] R. Weber. On the Gittins Index for Multiarmed Bandits. *The Annals of Applied Probability*, 2(4):1024 – 1033, 1992.

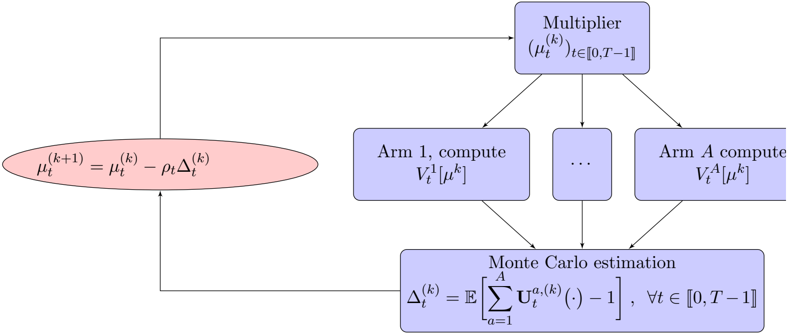
Outline of the presentation

The DeCo algorithm

- DeCo (decomposition-coordination algorithm)
- Stands for the decentralized control policy obtained by arm decomposition
- By contrast with the (brute force) dynamic programming solution (Bf), we have to solve Bellman equations for each arm,
  $\Longrightarrow$ dynamic programming with state of dimension 2 no matter the number of arms
- The DeCo algorithm is made of
  - an offline computation phase
  - an online computation phase

Offline phase of the DeCo algorithm

Minimization of the dual function
$$\varphi(\mu) = \left(\sum_{a \in A} V_0^a[\mu](\pi_0^a) + \sum_{t=0}^{T-1} \mu_t \right)$$
for a given family $\{\pi_0^a\}_{a \in A} = \left\{\beta\left(n_0^{\mathtt{B}a}, n_0^{\mathtt{G}a}\right)\right\}_{a \in A}$ of beta priors



Offline phase of the DECO algorithm (continued)

1. Choose an initial vector $\mu^{(0)} \in \mathbb{R}^T$ of multipliers.
2. Iteration k, given multipliers $\mu^{(k)} \in \mathbb{R}^T$, compute the Bellman functions $\{V_t^a[\mu^{(k)}]\}_{t \in [\![0,T]\!], a \in A}$ and optimal controls.
   - The computation is performed in parallel, arm per arm.
   - $V_t^a[\mu^{(k)}]$ is to be evaluated only on the finite grid $\{(n_0^{\mathtt{Ba}} + n^{\mathtt{Ba}}, n_0^{\mathtt{Ga}} + n^{\mathtt{Ga}}) \mid n^{\mathtt{Ba}} + n^{\mathtt{Ga}} \leq t\}$.
   - If all the arms share the same prior and instantaneous reward, then all the arms share the same sequence of Bellman value functions.

Offline phase of the DECO algorithm (continued)

3. Once gotten $\left\{V_t^a[\mu^{(k)}]\right\}_{a\in A}$ at time $t=0$ and iteration k
   - update the multipliers by a gradient step to obtain $\mu^{(k+1)}$
   - The gradient of the dual function φ with respect to the multipliers is obtained by computing the expectation of the dualized constraint.
   - Numerically, the expectation is obtained by Monte Carlo simulations.
   - The gradient phase can be replaced by a more sophisticated algorithm such as the conjugate gradient or the quasi-Newton method.
   - In some of our numerical experiments, we use a solver (limited memory BFGS) of the MODULOPT library from INRIA [4]. To obtain a global $O(T^3)$ running time, the computing budget allocated to this iterated gradient phase does not depend on T.
4. Stop the iterations (stopping criterion) or go back to 2 with multiplier $\mu^{(k+1)}$.

Online phase of the DECO algorithm

- The global stochastic optimal control problem is, theoretically, solvable by dynamic programming
- Using the Bellman value functions $\{V_t\}_{t \in [\![0,T]\!]}$, an optimal policy would be given by the feedback (where $\pi_t = \{\pi_t^a\}_{a \in A} = \big\{\beta(n_t^{\mathtt{Ba}}, n_t^{\mathtt{Ga}})\big\}_{a \in A}$)

$$U_t(\pi_t) \in \mathop{\arg\max}_{\substack{u_t = \{u_t^a\}_{a \in A} \in \{0,1\}^A \\ \sum_{a \in A} u_t^a = 1}} \Big(\tilde{L}_t(\pi_t, u_t) + \int_{\Delta(\Sigma)} V_{t+1}(\pi_{t+1}) k_t(\mathrm{d}\pi_{t+1} \mid \pi_t, u_t)\Big)$$

- The DECO algorithm consists in replacing the Bellman value function V_{t+1} by $\sum_{a \in A} V_{t+1}^a[\mu]$, using the collection $\big\{V_{t+1}^a[\mu]\big\}_{a \in A}$, of Bellman value functions given by the offline phase and a suitable vector $\mu \in \mathbb{R}^T$

Online phase of the DeCo algorithm (continued)

- We obtain the following policy: when the state of the multi-armed bandit is given by $(n^{\mathtt{B}a}, n_t^{\mathtt{G}a})_{a\in A}$ at time t, the DeCo algorithm selects an arm $\mathcal{A}_t^*[\mu]\big(\{(n^{\mathtt{B}a}, n_t^{\mathtt{G}a})\}_{a\in A}\big)$ in

$$\begin{aligned}\arg\max_{a\in A}\Big[& -V_{t+1}^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}) \\ & + \frac{n^{\mathtt{B}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}}\big(L_t^a(\mathtt{B}) + V_{t+1}^a[\mu](n^{\mathtt{B}a}+1, n^{\mathtt{G}a})\big) \\ & + \frac{n^{\mathtt{G}a}}{n^{\mathtt{B}a}+n^{\mathtt{G}a}}\big(L_t^a(\mathtt{G}) + V_{t+1}^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a}+1)\big)\Big]\end{aligned}$$

- This is a *nonstationary index policy*
- The DeCo policy used in numerical experiments is the policy $\mathcal{A}^*[\mu^*]$,
where μ^* is given by the offline phase of the DeCo algorithm

Interpretation

- The index in DeCo is the sum of an exploration term and of an exploitation term
- We define the value of the information to be gained from pulling arm a at time t as

$$\delta_t^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}}) = \frac{n^{\mathrm{Ba}}}{n^{\mathrm{Ba}} + n^{\mathrm{Ga}}} V_{t+1}^a[\mu](n^{\mathrm{Ba}} + 1, n^{\mathrm{Ga}}) + \frac{n^{\mathrm{Ga}}}{n^{\mathrm{Ba}} + n^{\mathrm{Ga}}} V_{t+1}^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}} + 1) - V_{t+1}^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}})$$

- Using $\delta_t^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}})$, we can write

$$V_t^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}}) = V_{t+1}^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}}) + \left(\delta_t^a[\mu](n^{\mathrm{Ba}}, n^{\mathrm{Ga}}) + \frac{n^{\mathrm{Ba}}}{n^{\mathrm{Ba}} + n^{\mathrm{Ga}}} L_t^a(\mathrm{B}) + \frac{n^{\mathrm{Ga}}}{n^{\mathrm{Ba}} + n^{\mathrm{Ga}}} L_t^a(\mathrm{G}) - \mu_t\right)^+$$

- The arm is pulled in the decomposed problem only if the sum of the information gain (δ_t) and the expected reward is greater than μ_t

Interpretation (continued)

- μ_t interpreted as an equilibrium price of a "bandit market"
- Each bandit is handled by an independent profit maximizing agent, which is required to pay the market price μ_t to pull the arm of its bandit at time t
- This is different but connected to the fair charge metaphore proposed in [11] for the Gittins index. Here the price depends on a market made of several arms, whereas for the Gittins index, the fair charge is arm specific.
- Last, the selected arm (in online phase) is the one maximizing

$$\underbrace{\delta_t^a[\mu](n^{\mathtt{B}a}, n^{\mathtt{G}a})}_{\text{exploration}} + \underbrace{\frac{n^{\mathtt{B}a}}{n^{\mathtt{B}a} + n^{\mathtt{G}a}} L_t^a(\mathtt{B}) + \frac{n^{\mathtt{G}a}}{n^{\mathtt{B}a} + n^{\mathtt{G}a}} L_t^a(\mathtt{G})}_{\text{exploitation}}$$

- Such exploration term is reminiscent of the exploration term encountered in UCB. Also, [5] refers to a learning component in the Gittins index as the difference between the index value and the immediate expected reward. More recently, the notion of information gain is also important in [9].

Computational complexity

- Solving the global maximization problem by DP
  is only possible for $|A|$ small and T small
  computational cost $O((2|A|)^T)$
- FH-GITTINS: time complexity $O(T^6)$
- DECO: DP phase cost running time $O(T^3)$
  indeed for each time $t \in [\![1, T]\!]$, we need a grid of $T \times T$
  for the 2 dimensional prior parameter
  (number of successes and failures)
- In the experiment, we fixed the number of gradient calls,
  so that the overall computing cost was $O(T^3)$ in time

Outline of the presentation

Probabilistic model

- Let $\Sigma = \left\{ p = (p^{\mathtt{B}}, p^{\mathtt{G}}) \in \mathbb{R}^2_+ \,\middle|\, p^{\mathtt{B}} + p^{\mathtt{G}} = 1 \right\}$ be the one-dimensional simplex
- For any $p = (p^{\mathtt{B}}, p^{\mathtt{G}}) \in \Sigma$, we consider on the space $\{\mathtt{B}, \mathtt{G}\}^T$ the probability $\mathcal{B}(p^{\mathtt{B}}, p^{\mathtt{G}}) = \bigotimes_{t=1}^{T} \left(p^{\mathtt{B}} \delta_{\mathtt{B}} + p^{\mathtt{G}} \delta_{\mathtt{G}} \right)$ — probability law of a sequence of independent (Bernoulli) random variables with values in $\{\mathtt{B}, \mathtt{G}\}$
- For $\{p^a\}_{a \in A} = \left\{ (p^{\mathtt{B}a}, p^{\mathtt{G}a}) \right\}_{a \in A} \in \prod_{a \in A} \Sigma$,
  - we consider the probability $\bigotimes_{a \in A} \mathcal{B}(p^{\mathtt{B}a}, p^{\mathtt{G}a})$ on the product space $\prod_{a \in A} \{\mathtt{B}, \mathtt{G}\}^T$ — which corresponds to independence between arms in A
  - we denote by $\mathbb{E}_{\{p^a\}_{a \in A}}$ the corresponding mathematical expectation
- We suppose that the DM holds a prior π_0^a over the unknown $p^a = (p^{\mathtt{B}a}, p^{\mathtt{G}a}) \in \Sigma$, for every arm $a \in A$
  In practice, we consider a beta distribution $\beta(n^{\mathtt{B}}, n^{\mathtt{G}})$ on Σ, with positive integers $n^{\mathtt{B}} > 0$ and $n^{\mathtt{G}} > 0$ as parameters

Probabilistic model (continued)

- We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where
  - $\Omega = \prod_{a \in A} \Sigma \times \{\mathtt{B}, \mathtt{G}\}^T$,
  - $\mathfrak{F} = 2^{\Omega}$,
  - $\mathbb{P} = \bigotimes_{a \in A} \pi_0^a\big(\,\mathrm{d}(p^{\mathtt{B}a}, p^{\mathtt{G}a})\big) \otimes \mathcal{B}(p^{\mathtt{B}a}, p^{\mathtt{G}a})$.
- Then, $\mathbf{W}^a = \{\mathbf{W}_t^a\}_{t \in [\![1,T]\!]}$ denotes the coordinate mappings for every arm $a \in A$, with $\mathbf{W}_t^a$ a random variable having values in the set $\{\mathtt{B}, \mathtt{G}\}$.
- For a given family $\{(\bar{p}_{\mathtt{B}}^a, \bar{p}_{\mathtt{G}}^a)\}_{a \in A} \in \prod_{a \in A} \Sigma$ and for $\pi_0^a = \delta_{(\bar{p}_{\mathtt{B}}^a, \bar{p}_{\mathtt{G}}^a)}$, for every arm $a \in A$, the family $\{\mathbf{W}_t^a\}_{a \in A, t \in [\![1,T]\!]}$ consists of independent random variables, where $\mathbf{W}_t^a$ has (Bernouilli) probability distribution with parameter $\bar{p}_{\mathtt{G}}^a \in [0,1]$, that is, $\mathbb{P}\big(\mathbf{W}_t^a = \mathtt{B}\big) = 1 - \bar{p}_{\mathtt{G}}^a$ and $\mathbb{P}\big(\mathbf{W}_t^a = \mathtt{G}\big) = \bar{p}_{\mathtt{G}}^a$. With this probabilistic model, we represent the sequential independent outcomes of $|A|$ independent arms.

Information and admissible controls

- The DM observes the random variable $\mathbf{Y}_{t+1} = \left\{\mathbf{U}_t^a \mathbf{W}_{t+1}^a\right\}_{a \in A}, \forall t \in [\![0, T-1]\!]$
  - When the arm a has been selected at stage t ($\mathbf{U}_t^a = 1$), the DM observes the outcome of the r.v. $\mathbf{W}_{t+1}^a \in \{\mathtt{B}, \mathtt{G}\}$.
  - When the arm a has not been selected at stage t ($\mathbf{U}_t^a = 0$), the DM observes nothing.
- The admissible controls $\mathbf{U} = \left\{\mathbf{U}_t\right\}_{t \in [\![0, T-1]\!]}$ are those that satisfy

$$\sigma(\mathbf{U}_t) \subset \sigma(\mathbf{Y}_0, \mathbf{U}_0, \mathbf{Y}_1, \ldots, \mathbf{U}_{t-1}, \mathbf{Y}_t) \ , \ \ \forall t \in [\![0, T-1]\!] \ ,$$

where $\sigma(\mathbf{Z}) \subset \mathfrak{F}$ is the σ-field generated by the random variable $\mathbf{Z}$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Random rewards

- We consider given a family $\{L_t^a\}_{a\in A, t\in[\![0,T-1]\!]}$ of instantaneous reward functions $L_t^a : \{\mathtt{B},\mathtt{G}\} \to \mathbb{R}$,
- The total random reward associated with the control $\mathbf{U} = \{\mathbf{U}_t\}_{t\in[\![0,T-1]\!]}$ is given by

$$\sum_{t=0}^{T-1} \sum_{a\in A} \mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a)$$

  - When the arm a has been selected at stage t ($\mathbf{U}_t^a = 1$), the r.v. $\mathbf{W}_{t+1}^a$ materializes and the DM receives the payoff $1 \times L_t^a(\mathbf{W}_{t+1}^a) = \mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a)$.
  - When the arm a has not been selected at stage t ($\mathbf{U}_t^a = 0$), the DM receives the payoff $0 = \mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a)$.

Optimality criteria in the Bayesian framework

- Let $\pi_0 = \{\pi_0^a\}_{a \in A} \in \prod_{a \in A} \Delta(\Sigma)$ be the family of initial priors.
  - $\Delta(\Sigma)$: set of probability distributions on Σ.
- We formulate a maximization problem

$$V_0(\pi_0) = \sup \int_{\Delta(\Sigma)^A} \prod_{a \in A} \pi_0^a(\mathrm{d}p^a) \mathbb{E}_{\{p^a\}_{a \in A}} \Big[\sum_{t=0}^{T-1} \sum_{a \in A} \mathbf{U}_t^a L_t^a(\mathbf{W}_{t+1}^a) \Big]$$

  - The supremum is taken over $\mathbf{U} = \{\mathbf{U}_t^a\}_{a \in A, t \in [\![0, T-1]\!]}$ subject to constraints

$$\sum_{a \in A} \mathbf{U}_t^a = 1 \ , \ \ \forall t \in [\![0, T-1]\!]$$

$$\sigma(\mathbf{U}_t) \subset \sigma(\mathbf{Y}_0, \mathbf{U}_0, \mathbf{Y}_1, \dots, \mathbf{U}_{t-1}, \mathbf{Y}_t) \ , \ \ \forall t \in [\![0, T-1]\!]$$