Efficient Decision Making under Uncertainty in a Power System Investment Problem

Jialin Liu
Shenzhen Key Laboratory of Computational Intelligence
University Key Laboratory of Evolving Intelligent Systems of Guangdong Province
Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen 518055, China
Email: liujl@sustech.edu.cn

Olivier Teytaud
Inria TAU, LRI, UMR 8623
CNRS - Univ. Paris-Saclay
Gif-sur-Yvette 91190, France
Email: olivier.teytaud@inria.fr

***Abstract*—The optimization of power systems involves complex uncertainties, such as technological progress, political context, geopolitical constraints. These uncertainties are difficult to modelize as probabilities, due to the lack of data for future technologies and due to partially adversarial geopolitical decision makers. Tools for such difficult decision making problems include Wald and Savage criteria, probabilistic reasoning and Nash equilibria. We investigate the rationale behind the use of a two-player Nash equilibrium approach in such a difficult context, and show that the approach is computationally efficient for large problems. Moreover, it automatically provides a selection of interesting decisions and critical scenarios for decision makers and is computationally cheaper than the Wald or Savage, thanks to the use of the sparsity of Nash equilibrium. It also has a natural interpretation in the sense that Nature does not make decisions taking into account our own decisions. The proposed approach was tested on instances of an artificial power system investment problem and can be applied to other problems, that can be modelled as a two-player matrix game or of which a payoff matrix can be built.**

***Index Terms*—Power system investment, scenario-based decision making, Nash equilibrium, two-player matrix game**

I. Introduction

Planning in power systems relies on many uncertainties. Some of them, originating in nature or in consumption, can be tackled through probabilities [1]–[4], others, such as technology evolution, geopolitics or CO2 penalization laws, are somewhere between stochastic and adversarial. Planning in power systems usually involves several agents (e.g., countries), therefore the decision making also depends on the decisions of other agents. For instance, the United Nations Climate Change Conference, COP21, aims at achieving a new universal agreement on climate agreement, which is an issue of, dynamic and changeable, cooperation and competition. Different type of energy sources include particular uncertainties. The construction of a country's nuclear reactors or its uranium supply could highly depend on another more developed country. The convention and trade wars between countries are not trivial to predict. The curtailment of alternative energy, such as wind and solar, may occur for several reasons including transmission congestion (or local network constraints), global oversupply and operational issues [5]. Each type of curtailment occurs with different frequencies depending on the generation and electrical characteristics of the regional and local systems. Another example is the risk of terrorism in the congested traffic, which cannot be represented by any stochastic model. Scenario-based decision tools are needed to handle such uncertainties without probabilistic model. Handling such uncertainties is a challenge. For example, how should we modelize the risk of gas curtailment, the evolution of oil prices, and the status of uranium supply?

There exists various non-probabilistic decision-making models, such as the Wald's maximin criterion (maximizing the worst-case reward) and Savage's minimax criterion (minimizing the worst-case regret). However, the Wald's and Savage's criteria are conservative. Additionally, both return only one decision (detailed later in Section II). In this paper, we propose to provide a set of policies for decision maker by approximating Nash equilibrium using adversarial bandit algorithm and the computational cost is reduced thanks to the sparsity of Nash equilibrium. The proposed approach is tested on instances of an artificial power investment problem with different levels of randomness. Experimental results show that the sparsity helps to cope with uncertainties and extract automatically the critical scenarios and the most interesting polices. The proposed approach is computationally efficient for decision making in large problems.

The remainder of this paper is structured as follows. We review and compare existing methodologies in Section II. Section III describes our proposed approach. Empirical study is provided in Section IV. Section V concludes.

II. Background

To facilitate the presentation, Section II-A introduces the notations used in this paper. In Section II-B, we review some

This work was supported by National Key R&D Program of China (Grant No. 2017YFC0804003), National Natural Science Foundation of China (Grant No. 61906083), Shenzhen Peacock Plan (Grant No. KQTD2016112514355531), the Science and Technology Innovation Committee Foundation of Shenzhen (Grant No. ZDSYS201703031748284), the Program for University Key Laboratory of Guangdong Province (Grant No. 2017KSYS008) and the ADEME project POST ("Plateforme d'Optimisation des Supergrids Transcontinentaux").

classic decision methods under uncertainty and their weaknesses/advantages in the context of power system investment, then compare them in Section II-C.

A. Notations

The notations are as follows: $\mathcal{S}$ is the set of possible scenarios and $\mathcal{K}$ is the set of possible policies. $\mathbf{R}$ is the matrix of rewards and and the associated reward function is $\mathbf{R}_{k,s} = R(k,s)$, i.e., $R(k,s)$ is the reward when applying policy $k \in \mathcal{K}$ in case the outcome of uncertainties is $s \in \mathcal{S}$. The reward function is also called a utility function or a payoff function. A *strategy* (a.k.a. policy) is a random variable k with values in $\mathcal{K}$. A *mixed strategy* is a probability distribution of possible policies; this is the general case of a strategy. A *pure strategy* is a deterministic policy, i.e., it is a mixed strategy with probability 1 for one and exactly one element, the others having probability 0. The *exploitability* of a (deterministic or randomized) strategy k is

$$\left(\max_{k' stochastic} \min_{s\in\mathcal{S}} \mathbb{E}_{k'} R(k',s)\right) - \min_{s\in\mathcal{S}} \mathbb{E}_k R(k,s). \quad (1)$$

We refer to the choice of s as Nature's choice. This does not mean that only natural effects are involved; geopolitics and technological uncertainties are included as well. k is the decision we are maximizing. In fact, natural phenomena can usually be modelized with probabilities, and are included through random perturbations - they are not the point in this work - contrarily to climate change uncertainties.

To make the notation simpler, we will use "m.s." and "p.s." as acronyms for *mixed strategy* and *pure strategy*, respectively, in the equations.

B. Decision making under uncertainty

1) Scenario-based planning: Maybe the most usual solution consists in selecting a small set $\{s_1, \dots, s_M\}$ of possible s, assumed to be most realistic. Then, for each s_j, an optimal k_i is obtained. The human then checks the matrix of the $R(k_i, s_j)$ for $j \in \{1, \dots, M\}$ and corresponding i. Variants of this approach have been studied in scenario planning [6]–[8]. Y. Feng [9] provides examples with more than 1000 scenarios. When optimizing the transmission network, we must take into account the future installation of power plants, for which there are many possible scenarios - in particular, the durations involved in power plant building are not necessarily larger than constants involved in big transmission lines. The scenarios involving large wind farms, or large nuclear power plants, lead to very specific constraints depending on their capacities and locations.

2) Wald criterion: The Wald criterion [10] consists in optimizing in the worst case scenario. For a maximization problem, the *Wald-value* is

$$v_{wald} = \max_{k \text{ p.s. on } \mathcal{K}} \min_{s\in\mathcal{S}} R_{k,s}, \quad (2)$$

and the recommended policy is k realizing the $\max$. We choose a policy which provides the best solution (maximal reward) for the worst scenario. Wald's maximin model provides a reward which is guaranteed in all cases. Implicitly, it assumes that Nature will make its decision in order to bother us, and, in a more subtle manner, Nature will make its decision while knowing what we are going to decide. It is hard to believe, for example, that the ultimate technological limit of photovoltaic units will be worse if we decide to do massive investments in solar power. Therefore, Wald's criterion is too conservative in many cases; hence the design of the Savage criterion.

3) Savage criterion: For a maximization problem, the *Savage-value* [11] is:

$$v_{savage} = \min_{k \text{ p.s. on } \mathcal{K}} \max_{s\in\mathcal{S}} regret(k,s), \quad (3)$$

where $regret(k,s) = \max_{k'\in\mathcal{K}} (R_{k',s} - R_{k,s})$. The Savage criterion is an application of the Wald's maximin model to the regret. Contrarily to Wald's criterion, it does not focus on the worst scenario. Its interpretation is that we optimize the guaranteed loss compared to an anticipative choice (anticipative in the sense: aware of all future outcomes) of decision. On the other hand, Nature still makes its decision after us, and has access to our decision before making its decision - Nature, in this model, can still decide to reduce the technological progress of wind turbines just because we have decided to do massive investments in wind power.

4) Nash equilibria: The principle of the Nash equilibrium is that contrarily to what is assumed in Wald's criterion, there is no reason for Nature (the opponent) to make a decision *after* us, and to know what we have decided. The *Nash-value* is

$$v_{nash} = \max_{k \text{ m.s. on } \mathcal{K}} \min_{s\in\mathcal{S}} \mathbb{E}_k R(k,s), \quad (4)$$

where "m.s." stands for "mixed strategy". As a mixed strategy is used, the fact that the $\max$ is written before the $\min$ does not change the result [12]; v_{nash} is also equal to $\min_{s \text{ r.v. on } \mathcal{S}} \max_k \mathbb{E}_k R(k,s)$, where "r.v." stands for random variable.

The *exploitability* (1) of a (possibly mixed) strategy k is equivalent to

$$v_{nash} - \min_{s\in\mathcal{S}} \mathbb{E}_k R(k,s). \quad (5)$$

A Nash strategy is a strategy with exploitability equal to 0. A Nash strategy always exists, and it is not necessarily unique. A Nash equilibrium, for a finite-sum problem, is a pair of Nash strategies for us and for Nature, respectively. In the general case, a Nash strategy is not pure. Criteria for Nash equilibria corresponds to Nature and us making decision privately, i.e. without knowing what each other will do. In this sense, it is more intuitive than other criteria. Our proposed criterion is a combination of Nash and Savage as discussed in Section III.

5) Other decision tools: Other possible tools for partially adversarial decision making are multi-objective optimization (i.e. for each s, there is one objective function $k \mapsto R(k,s)$) [13] and possibilistic reasoning [14]. These tools rely intensively on human experts, a priori (selection of scenarios) or a posteriori (selection in the Pareto set).

C. Comparison between various decision tools

Let us compare the various discussed policies in Table I. We see that the Nash approach has a lower computational cost and some advantages in terms of modeling compared to Wald or Savage; Nature makes its decision privately (which means we do not know the uncertainties), but not with access to our decisions. On the other hand, its output is stochastic, which might be a drawback for users. Pure scenario-based method requires human expertise and lots of human resources.

It would be beneficial by selecting a number of displayed policies and scenarios using Nash approach, based on which the final decision will be made by human experts. In this work, we aim at extract automatically a such set of interesting policies and crucial scenarios for decision makers. The extracted policies achieve better performance than the decisions made using Wald, which is more conservative.

III. Our proposal: Nash Uncertainty Decision

Our proposed tool is as follows. We use Nash equilibria, for their principled nature and (as discussed later) low computational cost in large scale settings. We compute the equilibria thanks to adversarial bandit algorithms, as detailed in Section III-A. We use sparsity (Section III-B), for improving the precision and reducing the number of pure policies in our recommendation. The resulting algorithm (detailed in Section III-C) has the following advantages: (i) It is fast; this is not intuitive, but Nash equilibria, in spite of the complex theories behind this concept, can be approximated quickly, without computing the entire matrix of $\mathbf{R}$. A pioneering work in this direction was [15]; within logarithmic terms and dependency in the precision, the cost is roughly the square root of the size of the matrix. (ii) It naturally provides a submatrix of $\mathbf{R}$, for the best k and the most critical s.

We believe that such outcomes are natural tools for including in platforms for simulating large-scale power systems involving huge uncertainties.

A. Computing Nash equilibria with adversarial bandit algorithms

For the computational cost issue for computing Nash equilibria, there exist algorithms reaching approximate solutions much faster than the exact linear programming approach [16]. Some of these fast algorithms are based on the bandit formalism. The multi-armed bandit problem [17]–[19] is a model of exploration/exploitation trade-offs, aimed at optimizing the expected payoff. Let us define an adversarial multi-armed bandit with $K \in \mathbb{N}^+$ ($K > 1$) arms and let $\mathcal{K}$ denote the set of arms. Let $\mathcal{T} = \{1, \dots, T\}$ denote the set of time steps, with $T \in \mathbb{N}^+$ a finite time horizon. At each time step $t \in \mathcal{T}$, the algorithm chooses $i_t \in \mathcal{K}$ and obtains a reward $R_{i_t,t}$. The reward $R_{i_t,t}$ is a mapping $(\mathcal{K}, \mathcal{T}) \mapsto \mathbb{R}$.

The generic adversarial bandit is detailed in Algorithm 1. In the case of adversarial problems, when we search for a Nash equilibrium for a reward function $(k, s) \mapsto R(k, s)$, two bandit algorithms typically play against each other. The decision making process is modelled as playing a two-player game. One of them is Nature, and the other plays our role. At the end, our bandit algorithm recommends a (possibly mixed) strategy over the K arms. This recommended distribution is often the empirical distribution of play during the games against the Nature bandit. A related work is the use of approximated Nash equilibria for selecting random seeds to boost AI agents in playing fully and partially observable board games [20], [21].

Such a fast approximate solution can be provided by $Exp3$ (Exponential weights for Exploration and Exploitation) [22] and its variant $Exp3.P$ [23], presented in Algorithm 2. $Exp3$ has the same efficiency as the Grigoriadis and Khachiyan method [15] for finding approximate Nash equilibria, and can be implemented with two bandits playing one against each other, e.g. one for us and one for Nature. Note that $Exp3.P$ is not anytime: it requires the time horizon in order to initialize some input meta-parameters.

Algorithm 1 Generic adversarial multi-armed bandit. The problem is described through the arm set $\mathcal{K}$, the budget T, and most importantly the *get reward* method, i.e. the mapping $R : (\mathcal{K}, \mathcal{T}) \mapsto \mathbb{R}$, where $\mathcal{T} = \{1, \dots, T\}$.

Require: a time horizon (computational budget) $T \in \mathbb{N}^+$
Require: a set of arms $\mathcal{K}$
Require: a probability distribution π on $\mathcal{K}$
1: **for** $t \leftarrow 1$ to T **do**
2: Select arm $i_t \in \mathcal{K}$ based upon π
3: *Get reward* $R_{i_t,t}$
4: Update the probability distribution π using $R_{i_t,t}$
5: **end for**

B. Sparsity of Nash equilibrium

Teytaud and Flory [25] proposed a truncation technique on sparse problem. Considering the Nash equilibria for two-player finite-sum matrix games, if the Nash equilibrium of the problem is sparse, the small components of the solution can be removed and the remaining submatrix is solved exactly. This technique can be applied to some adversarial bandit algorithm such as $Grigoriadis$' algorithm [15], $Exp3$ [22] or Inf [26]. The properties of this sparsity technique are as follows. Asymptotically in the computational budget, the convergence to the Nash equilibria is preserved [25]. The computation time is lower if there exists a sparse solution [27]. The support of the obtained approximation has at most the same number of pure policies and often far less [25]. Essentially, we get rid of the random exploration part of the empirical distribution of play.

C. Overview of our method

A high level view of our method is given in Algorithm 3. All the algorithmic challenge is hidden in the computation engine, $tExp3.P$ (detailed in Algorithm 4), obtained by applying the truncation technique [25] (lines 13-20 of Algorithm 4) to $Exp3.P$ (previously presented in Algorithm 2).

TABLE I: Comparison between several tools for decision making under uncertainty. $K = |\mathcal{K}|$ is the number of possible investment policies, $S = |\mathcal{S}|$ is the number of scenarios, K' ($\ll K$) is the number of displayed policies, and S' ($\ll S$) is the number of displayed scenarios.

Method	Extraction of policy	Extraction of scenario	Computational cost	Interpretation
Scenario-based	Handcrafted	Handcrafted	$K' \times S'$	Human expertise.
Wald	One	One per policy	$K \times S$	Nature decides later, minimizing our reward.
Savage	One	One per policy	$K \times S$	Nature decides later, maximizing our regret.
Nash	Nash-optimal	Nash-optimal	$(K+S) \times \log(K+S)$	Nature decides privately, before us.

Algorithm 2 $Exp3.P$: variant of $Exp3$, proved to have a high probability bound on the weak reward [24]. η and γ are two input parameters.

Require: $\eta \in \mathbb{R}$
Require: $\gamma \in (0, 1]$
Require: a time horizon (computational budget) $T \in \mathbb{N}^+$
Require: $K \in \mathbb{N}^+$ is the number of arms
$y \leftarrow 0$
for $i \leftarrow 1$ to K **do** ▷ initialization
  $\omega_i \leftarrow \exp(\frac{\eta\gamma}{3}\sqrt{\frac{T}{K}})$
end for
for $t \leftarrow 1$ to T **do**
  for $i \leftarrow 1$ to K **do**
    $p_i \leftarrow (1-\gamma)\frac{\omega_i}{\sum_{j=1}^{K}\omega_j} + \frac{\gamma}{K}$
  end for
  Generate i_t according to $(p_1, p_2, \dots, p_K)$
  Compute reward $R_{i_t,t}$
  for $i \leftarrow 1$ to K **do**
    if $i == i_t$ **then**
      $\hat{R}_i \leftarrow \frac{R_{i_t,t}}{p_i}$
    else
      $\hat{R}_i \leftarrow 0$
    end if
    $\omega_i \leftarrow \omega_i \exp\left(\frac{\gamma}{3K}(\hat{R}_i + \frac{\eta}{p_i\sqrt{TK}})\right)$
  end for
end for
return probability distribution $(p_1, p_2, \dots, p_K)$

Algorithm 3 The Sparse-Nash algorithm for solving decision making problems under uncertainty.

Require: A family $\mathcal{K}$ of possible decisions (e.g., investment policies)
Require: A family $\mathcal{S}$ of scenarios
Require: A mapping $(k, s) \mapsto R_{k,s}$, providing the rewards, where $k \in \mathcal{K}$ and $s \in \mathcal{S}$
Run $tExp3.P$ on the mapping R, get a probability distribution on $\mathcal{K}$ and a probability distribution on $\mathcal{S}$
Output $k_1, \dots, k_m$ the policies with positive probability and $s_1, \dots, s_n$ the scenarios with positive probability. Emphasize the policy with highest probability
Output the matrix of $R(k_i, s_j)$ for $i \leq m$ and $j \leq n$

Algorithm 4 $tExp3.P$, combining $Exp3.P$ and the truncation method. α is the truncation parameter.

Require: $\mathbf{R}_{m\times n}$, matrix defined by mapping $(i, j) \mapsto \mathbf{R}_{i,j}$
Require: a time horizon (computational budget) $T \in \mathbb{N}^+$
Require: α, truncation parameter
Run $Exp3.P$ during T iterations; get an approximation (p, q) of the Nash equilibrium
$\zeta = \max_{i\in\{1,\dots,m\}} \frac{(Tp_i)^\alpha}{T}$ ▷ compute the threshold for p
for $i \leftarrow 1$ to m **do** ▷ truncation
  if $p_i \geq \zeta$ **then**
    $p'_i = p_i$
  else
    $p'_i = 0$
  end if
end for
for $i \leftarrow 1$ to m **do**
  $p''_i = \frac{p'_i}{\sum_{j=1}^{m} p'_j}$
end for
$\zeta' = \max_{i\in\{1,\dots,n\}} \frac{(Tq_i)^\alpha}{T}$ ▷ compute the threshold for q
for $i \leftarrow 1$ to n **do** ▷ truncation
  if $q_i \geq \zeta'$ **then**
    $q'_i = q_i$
  else
    $q'_i = 0$
  end if
end for
for $i \leftarrow 1$ to n **do**
  $q''_i = \frac{q'_i}{\sum_{j=1}^{n} q'_j}$
end for
return p'' and q'' as an approximate Nash equilibrium of the problem

IV. Experiments and Discussions

We propose a simple model of investments in power systems. Our model is not supposed to be super realistic, it is aimed at being easy to reproduce.

A. Power investment problem

We consider each investment *policy*, sometimes called action or decision, a vector

$$\mathbf{k} = (C, F, X, S, W, P, T, U, N, A) \in \{0, \frac{1}{2}, 1\}^{10}.$$

A *scenario* is a vector

$$\mathbf{s} = (Z, WB, PB, TB, XB, UB, SB, CC, NT) \in \{0, \frac{1}{2}, 1\}^9.$$

The parameters and detailed corresponding descriptions of policy variables and scenario variables are provided in Tables IIa and IIb, respectively.

Let $\mathcal{S}$ be the set of possible scenarios and $\mathcal{K}$ be the set of possible policies. The utility function R is a mapping $(\mathcal{K}, \mathcal{S}) \mapsto \mathbb{R}$. Given decision $\mathbf{k} \in \mathcal{K}$ and scenario $\mathbf{s} \in \mathcal{S}$, a reward can be computed by

$$\begin{aligned} R(\mathbf{k}, \mathbf{s}) = \frac{2}{3}(1 + rand) \cdot (N(1 - Z)/5 \\ - cost \cdot (N + U + T + P + W + S + X + F + C) \\ + c \cdot ((X == XB) + (C! = CC) + (F! = NT) + (P == PB)) \\ + 7XB \cdot X + W(1 + WB)(SB + \sqrt{S})/2 \\ + 3P(PB + SB) - 4C \cdot CC \\ - F \cdot NT + S(1 - Z) + P \cdot Z + U \cdot UB \\ + T \cdot S \cdot (1 + TB - SB/2) - F \cdot NT \\ + A \cdot (1 + W + P - 2SB)), \end{aligned} \tag{6}$$

where $rand$ is a uniform random generator in the range $\in (0, 1)$, $cost$ and c are meta-parameters. The meta-parameter c determines how sensitive the reward is to the disasters and breakthrough of technologies.

This provides a reward function $R(\mathbf{k}, \mathbf{s})$, with which we can build a matrix $\mathbf{R}$ of rewards. However, with a ternary discretization for each variable we get a huge matrix, that we will not construct explicitly - more precisely, it would be impossible to construct it explicitly with a real problem involving hours of computation for each $R(\mathbf{k}, \mathbf{s})$. Fortunately, approximate algorithms can solve Nash equilibria with precision ϵ with $O(K \log(K)/\epsilon^2)$ requests to the reward function [19], i.e. far less than the quadratic computation time K^2 needed for reading all entries in a matrix of size K^2.

B. Experimental setting

We perform experiments on the designed noisy investment problem (6) and apply the algorithm $tExp3.P$ using the input parameter values $\eta = 2\sqrt{\log \frac{KT}{\epsilon}}$ and $\gamma = \min(0.6, 2\sqrt{\frac{3K \log(K)}{5T}})$ proposed by Busa-Fekete and Kégl [28].

We consider policies and scenarios in discrete domains: $\mathcal{K} = \{0, \frac{1}{2}, 1\}^{10}$, $\mathcal{S} = \{0, \frac{1}{2}, 1\}^9$. The reward matrix $\mathbf{R}_{3^{10} \times 3^9}$ can be defined by $\mathbf{R}_{i,j} = R(\mathbf{k}_i, \mathbf{s}_j)$, where $\mathbf{k}_i$ denotes the i^{th} policy in $\mathcal{K}$ and $\mathbf{s}_i$ denotes the j^{th} scenario in $\mathcal{S}$ ($\forall i \in \{1, \ldots, 3^{10}\}$, $\forall j \in \{1, \ldots, 3^9\}$). Note that the reward is noisy as previously mentioned. Thus, each line of the matrix is a possible policy and each column is a scenario, $\mathbf{R}_{i,j}$ is the stochastic reward obtained by apply the policy $\mathbf{k}_i$ to the scenario $\mathbf{s}_j$.

Experiments are performed for different numbers of time steps in the bandit algorithms, i.e. we consider T simulations for each $T \in \{1, 10, 50, 100, 500, 1000\} \cdot K$. Therefore, for each T, the input meta-parameters η and γ are different, as they depend on the budget T. In the entire paper, when we show an expected reward $R(\mathbf{k}, \mathbf{s})$ for some $\mathbf{s}$ and for $\mathbf{k}$ learned by one of our methods, we refer to $10,000$ independent trials; $R(\mathbf{k}, \mathbf{s})$ are played for $10,000$ randomly drawn pairs $(\mathbf{k}_{i_n}, \mathbf{s}_{j_n})$ i.i.d. according to the random variables i_n and j_n proposed by the considered policies. The performance is the average reward of these $10,000$ trials $R(\mathbf{k}_{i_1}, \mathbf{s}_{j_1}), \ldots, R(\mathbf{k}_{i_{10000}}, \mathbf{s}_{j_{10000}})$. There is an additional averaging, over learning. Namely, each learning (i.e. the sequence of $Exp3$ iteration for approximating a Nash equilibrium) is repeated 100 times. The meta-parameters $cost$ is set to 1 and c is set to $\{1, 2, \ldots, 10\}$ in our experiments. The reward matrix is normalized in the experiments due to the assumptions when recommending the theoretically optimal values for the parameters η and γ in $Exp3.P$.

C. Experimental results and discussions

We present the the sparsity level (i.e. the number of pure policies in the support of the obtained approximation), the robust score (defined as the worst of the rewards against pure policies) and the proxy exploitability (defined as the difference between the best robust score in the table, minus the robust score) with $c = 1$ and $c = 10$ in Tables III and IV, respectively.

1) Sparsity helps in various horizons: Teytaud and Flory [25] proposed $\alpha = 0.7$ as truncation parameter and Auger et al. [27] used the same value. In both of our testcases (c.f. Tables III and IV), $\alpha = 0.9$ does not provide good results when $T = K$, however $\alpha = 0.7$ is always better than the baseline, to which the truncation technique is not applied (rows with heading "NT" in Tables III and IV). We validate the good performance of $\alpha = 0.7$.

We observe that when the number of simulations is bigger than the cardinality of the search domain, i.e. the number of possible pure policies, then $\alpha \simeq 0.9$ leads to better empirical mean reward against the pure policies. For example, for the testcase with $c = 1$, $\alpha = 0.9$ outperforms the other values of α at most of time; when the budget is huge ($T = 1000K$), $\alpha = 0.99$ provides better results.

2) The parameters of $Exp3.P$: When learning with few simulations ($T = K$), the non-truncated solutions and non-sparse solutions are as weak as a random strategy. Along with the increment of simulation times, the non-truncated solutions and non-sparse solutions become stronger, but still weaker than the truncated solutions. Sparsity level "0.01" means that one and only one solution of the 100 learnings has one element above the threshold ζ, the other 99 solutions of the 99 learnings have no element above the threshold ζ. This situation is not far from the non-truncated or non-sparse case. If the solution is sparse, we get a better empirical mean reward even with a small horizon, i.e. the $tExp3.P$ succeeds in finding better pure policies.

We see that truncated algorithms outperform their non-truncated counterparts, in particular, truncation clearly shows its strength when the number of simulations is small in front of the size of search domain.

TABLE II: Parameters and descriptions of policy variables ($\mathbf{k}$) and scenario ($\mathbf{s}$) in the designed model of power investment.

(a) Parameters and descriptions of policy variables ($\mathbf{k}$).

$\mathbf{k} \in \{0, \frac{1}{2}, 1\}$	**Corresponding investment**
C	Coal
F	Nuclear fission
X	Nuclear fusion
S	Supergrids
W	Wind power
P	PV units
T	Solar thermal
U	Unconventional renewable
N	Nanogrids
A	massive storage in Scandinavia

(b) Parameters and descriptions of scenario ($\mathbf{s}$).

$\mathbf{s} \in \{0, \frac{1}{2}, 1\}$	**Nature's action**
Z	Massive geopolitical issues
WB	Wind power technological breakthrough
PB	PV Units breakthrough
TB	Solar thermal breakthrough
XB	Fusion breakthrough
UB	Unconventional renewable breakthrough
SB	Local storage breakthrough
CC	Climate change disaster
NT	Nuclear terrorism

TABLE III: Results for reward matrix R computed with $c = 1$. In these tables, the result is the average value of 100 learnings. The reward matrix is normalized in the experiments. The standard deviation is shown after $\pm$. "NT" means that the truncation technique is not applied; "non-sparse" means that all elements of the solution provided are above the threshold ζ. **Top:** Average sparsity level (over $3^{10} = 59049$ arms), i.e. number of pure policies in the support of the obtained approximation, of solutions provided by $Exp3.P$ in power investment problem. **Middle:** Robust score (to be maximized) using different truncation parameter α for solutions provided by $Exp3.P$ in power investment problem. The robust score is the worst of the rewards against pure policies. **Bottom:** Proxy exploitability (to be minimized) using different truncation parameter α for solutions provided by $Exp3.P$ in power investment problem. The proxy exploitability is the difference between the best robust score in the table, minus the robust score.

α	**Average sparsity level over $3^{10} = 59049$ arms**					
	$T = K$	$T = 10K$	$T = 50K$	$T = 100K$	$T = 500K$	$T = 1000K$
0.1	13804.380 ± 52.015	non-sparse	non-sparse	non-sparse	non-sparse	non-sparse
0.3	2810.120 ± 59.083	non-sparse	non-sparse	non-sparse	non-sparse	non-sparse
0.5	395.920 ± 15.835	non-sparse	non-sparse	59048.960 ± 196.946	49819.430 ± 195.016	non-sparse
0.7	43.230 ± 2.624	58925.340 ± 26.821	55383.140 ± 150.057	46000.020 ± 277.653	9065.180 ± 159.610	non-sparse
0.9	3.600 ± 0.260	992.940 ± 64.474	796.500 ± 41.724	503.600 ± 24.927	97.670 ± 5.445	52632.820 ± 522.505
0.99	1.110 ± 0.031	2.250 ± 0.171	2.500 ± 0.180	2.310 ± 0.156	1.790 ± 0.121	6.700 ± 0.612

α	**Robust score**					
	$T = K$	$T = 10K$	$T = 50K$	$T = 100K$	$T = 500K$	$T = 1000K$
NT	4.922e-01 ± 5.649e-07	4.928e-01 ± 1.787e-06	4.956e-01 ± 4.016e-06	4.991e-01 ± 5.892e-06	5.221e-01 ± 1.404e-05	4.938e-01 ± 1.687e-06
0.1	4.948e-01 ± 5.739e-05	4.928e-01 ± 1.787e-06	4.956e-01 ± 4.016e-06	4.991e-01 ± 5.892e-06	5.221e-01 ± 1.404e-05	4.938e-01 ± 1.687e-06
0.3	5.004e-01 ± 1.397e-04	4.928e-01 ± 1.787e-06	4.956e-01 ± 4.016e-06	4.991e-01 ± 5.892e-06	5.221e-01 ± 1.404e-05	4.938e-01 ± 1.687e-06
0.5	**5.059e-01 ± 2.272e-04**	4.928e-01 ± 1.787e-06	4.956e-01 ± 4.016e-06	4.991e-01 ± 5.891e-06	5.242e-01 ± 5.491e-05	4.938e-01 ± 1.687e-06
0.7	5.054e-01 ± 1.327e-03	4.928e-01 ± 3.835e-06	4.965e-01 ± 3.896e-05	5.031e-01 ± 1.016e-04	5.317e-01 ± 9.573e-05	4.938e-01 ± 1.687e-06
0.9	4.281e-01 ± 6.926e-03	**5.137e-01 ± 4.199e-04**	**5.151e-01 ± 5.007e-04**	**5.140e-01 ± 4.965e-04**	**5.487e-01 ± 9.413e-04**	4.960e-01 ± 1.828e-04
0.99	3.634e-01 ± 8.191e-03	4.357e-01 ± 6.873e-03	4.612e-01 ± 5.380e-03	4.683e-01 ± 4.834e-03	5.242e-01 ± 3.302e-03	**5.390e-01 ± 3.167e-03**
Pure	3.505e-01 ± 7.842e-03	3.946e-01 ± 7.181e-03	4.287e-01 ± 6.203e-03	4.489e-01 ± 5.410e-03	5.143e-01 ± 3.597e-03	4.837e-01 ± 5.558e-03

α	**Proxy exploitability**					
	$T = K$	$T = 10K$	$T = 50K$	$T = 100K$	$T = 500K$	$T = 1000K$
NT	1.369e-02	2.092e-02	1.946e-02	1.492e-02	2.669e-02	4.525e-02
0.1	1.109e-02	2.092e-02	1.946e-02	1.492e-02	2.669e-02	4.525e-02
0.3	5.485e-03	2.092e-02	1.946e-02	1.492e-02	2.669e-02	4.525e-02
0.5	**0.000e+00**	2.092e-02	1.946e-02	1.492e-02	2.454e-02	4.525e-02
0.7	4.328e-04	2.091e-02	1.854e-02	1.083e-02	1.705e-02	4.525e-02
0.9	7.778e-02	**0.000e+00**	**0.000e+00**	**0.000e+00**	**0.000e+00**	4.304e-02
0.99	1.425e-01	7.806e-02	5.385e-02	4.564e-02	2.456e-02	**0.000e+00**
Pure	1.554e-01	1.191e-01	8.638e-02	6.503e-02	3.443e-02	5.537e-02

TABLE IV: Results for reward matrix R computed with $c = 10$. In these tables, the result is the average value of 100 learnings. The reward matrix is normalized in the experiments. The standard deviation is shown after $\pm$. "NT" means that the truncation technique is not applied; "non-sparse" means that all elements of the solution provided are above the threshold ζ. **Top:** Average sparsity level (over $3^{10} = 59049$ arms), i.e. number of pure policies in the support of the obtained approximation, of solutions provided by $Exp3.P$ in power investment problem. **Middle:** Robust score (to be maximized) using different truncation parameter α for solutions provided by $Exp3.P$ in power investment problem. The robust score is the worst of the rewards against pure policies. **Bottom:** Proxy exploitability (to be minimized) using different truncation parameter α for solutions provided by $Exp3.P$ in power investment problem. The proxy exploitability is the difference between the best robust score in the table, minus the robust score.

α	**Average sparsity level over $3^{10} = 59049$ arms**					
	$T = K$	$T = 10K$	$T = 50K$	$T = 100K$	$T = 500K$	$T = 1000K$
0.1	6394.625 $\pm$ 84.308	non-sparse	non-sparse	non-sparse	non-sparse	non-sparse
0.3	1337.896 $\pm$ 40.491	non-sparse	non-sparse	non-sparse	non-sparse	non-sparse
0.5	206.146 $\pm$ 12.647	non-sparse	non-sparse	non-sparse	non-sparse	non-sparse
0.7	25.563 $\pm$ 2.045	non-sparse	non-sparse	non-sparse	59048.750 $\pm$ 0.250	non-sparse
0.9	3.729 $\pm$ 0.353	42616.313 $\pm$ 1476.644	47581 $\pm$ 1015.506	38361.182 $\pm$ 1091.373	4510.125 $\pm$ 726.595	58323.125 $\pm$ 157.971
0.99	1.208 $\pm$ 0.072	4.479 $\pm$ 0.575	5.333 $\pm$ 0.565	6.000 $\pm$ 0.969	2.875 $\pm$ 1.076	8.500 $\pm$ 2.204

α	**Robust score**					
	$T = K$	$T = 10K$	$T = 50K$	$T = 100K$	$T = 500K$	$T = 1000K$
NT	1.151e-01 $\pm$ 6.772e-08	1.151e-01 $\pm$ 2.175e-07	1.153e-01 $\pm$ 3.707e-07	1.154e-01 $\pm$ 5.797e-07	1.167e-01 $\pm$ 2.046e-06	1.152e-01 $\pm$ 1.297e-06
0.1	1.158e-01 $\pm$ 1.843e-05	1.151e-01 $\pm$ 2.175e-07	1.153e-01 $\pm$ 3.707e-07	1.154e-01 $\pm$ 6.019e-07	1.167e-01 $\pm$ 2.046e-06	1.152e-01 $\pm$ 1.297e-06
0.3	1.160e-01 $\pm$ 3.441e-05	1.151e-01 $\pm$ 2.175e-07	1.153e-01 $\pm$ 3.707e-07	1.154e-01 $\pm$ 6.019e-07	1.167e-01 $\pm$ 2.046e-06	1.152e-01 $\pm$ 1.297e-06
0.5	**1.166e-01 $\pm$ 9.751e-05**	1.151e-01 $\pm$ 2.175e-07	1.153e-01 $\pm$ 3.707e-07	1.154e-01 $\pm$ 5.797e-07	1.167e-01 $\pm$ 2.046e-06	1.152e-01 $\pm$ 1.297e-06
0.7	1.165e-01 $\pm$ 6.176e-04	1.151e-01 $\pm$ 2.175e-07	1.153e-01 $\pm$ 3.707e-07	1.154e-01 $\pm$ 5.797e-07	1.167e-01 $\pm$ 2.051e-06	1.152e-01 $\pm$ 1.297e-06
0.9	1.068e-01 $\pm$ 3.176e-03	**1.156e-01 $\pm$ 4.586e-05**	1.160e-01 $\pm$ 6.348e-05	1.172e-01 $\pm$ 9.722e-05	**1.266e-01 $\pm$ 3.829e-04**	1.152e-01 $\pm$ 4.288e-06
0.99	8.423e-02 $\pm$ 3.118e-03	1.119e-01 $\pm$ 2.316e-03	**1.189e-01 $\pm$ 1.888e-03**	**1.202e-01 $\pm$ 2.101e-03**	1.145e-01 $\pm$ 4.519e-03	**1.186e-01 $\pm$ 7.684e-04**
Pure	7.810e-02 $\pm$ 2.570e-03	8.354e-02 $\pm$ 2.710e-03	9.327e-02 $\pm$ 2.202e-03	9.658e-02 $\pm$ 2.097e-03	1.120e-01 $\pm$ 3.625e-03	8.755e-02 $\pm$ 6.497e-03

α	**Proxy exploitability**					
	$T = K$	$T = 10K$	$T = 50K$	$T = 100K$	$T = 500K$	$T = 1000K$
NT	1.494e-03	4.594e-04	3.592e-03	4.772e-03	9.903e-03	3.388e-03
0.1	7.727e-04	4.594e-04	3.592e-03	4.772e-03	9.903e-03	3.388e-03
0.3	5.838e-04	4.594e-04	3.592e-03	4.772e-03	9.903e-03	3.388e-03
0.5	**0.000e+00**	4.594e-04	3.592e-03	4.772e-03	9.903e-03	3.388e-03
0.7	9.391e-05	4.594e-04	3.592e-03	4.772e-03	9.903e-03	3.388e-03
0.9	9.758e-03	**0.000e+00**	2.860e-03	2.992e-03	**0.000e+00**	3.371e-03
0.99	3.236e-02	3.647e-03	**0.000e+00**	**0.000e+00**	1.211e-02	**0.000e+00**
Pure	3.848e-02	3.204e-02	2.559e-02	2.362e-02	1.463e-02	3.103e-02

V. Conclusion

We proposed in Section III a new criterion (based on Nash equilibria and sparsity) and a new methodology (based on adversarial bandits and sparsity) for decision making with uncertainty. Technically speaking, we tuned a parameter-free adversarial bandit algorithm $tExp3.P$, for large scale problems, efficient in terms of performance itself, and also in terms of sparsity. $tExp3.P$ performed better than $Exp3.P$ (without truncation). Moreover, $tExp3.P$ with truncation parameter $\alpha = 0.7$, which is theoretically guaranteed [25], got stable performance in the experiments.

From a user point of view, we get the following advantages: (i) Natural extraction of interesting policies and critical scenarios. However, we point out that $\alpha = .7$ provides stable (and proved) results, but the extracted submatrix becomes easily readable (small enough) with larger values of α. (ii) Faster computational cost than the Wald or Savage classical methodologies. Our methodology only requires a mapping $R : (k, s) \mapsto R(k, s)$, which computes the outcome if we use the policy k and the outcome is the scenario s. Multiple objective functions can be handled: if we have two objectives (e.g. economy and greenhouse gas pollution), we can just duplicate the scenarios, one for which the criterion is economy, and one for which the criterion is greenhouse gas. Given a problem, the algorithm will display a matrix of rewards for different policies and for several scenarios (including, by the trick above, several criteria such as particular matter, greenhouse, and cost).

As a summary, we get a fast criterion, faster than Wald's or Savage's criteria, with a natural interpretation. The algorithm naturally provides a matrix of results, namely the matrix of outcomes in the most interesting decisions and for the most critical scenarios. These decisions and scenarios are also equipped with a ranking.

References

[1] RTE-ft, "Rte forecast team: Electricity consumption in France : Characteristics and forecast method," 2008.

[2] P. Pinson, "Renewable energy forecasts ought to be probabilistic," in *WIPFOR seminar*, 2013.

[3] T. Siqueira, M. Zambelli, M. Cicogna, M. Andrade, and S. Soares, "Stochastic dynamic programming for long term hydrothermal scheduling considering different streamflow models," in *Probabilistic Methods Applied to Power Systems, 2006. PMAPS 2006. International Conference on*. IEEE, 2006, pp. 1–6.

[4] S. Vassena, P. Mack, P. Rousseaux, C. Druet, and L. Wehenkel, "A probabilistic approach to power system network planning under uncertainties," in *IEEE Bologna Power Tech Conference Proceedings*, 2003.

[5] D. Lew, L. Bird, M. Milligan, B. Speer, X. Wang, E. M. Carlini, A. Estanqueiro, D. Flynn, E. Gomez-Lazaro, N. Menemenlis *et al.*, "Wind and solar curtailment," in *Int. Workshop on Large-Scale Integration of Wind Power Into Power Systems*, 2013, pp. 1–9.

[6] P. Saisirirat, N. Chollacoop, M. Tongroon, Y. Laoonual, and J. Pongthanaisawan, "Scenario analysis of electric vehicle technology penetration in thailand: Comparisons of required electricity with power development plan and projections of fossil fuel and greenhouse gas reduction," *Energy Procedia*, vol. 34, no. 0, pp. 459 – 470, 2013, 10th Eco-Energy and Materials Science and Engineering Symposium.

[7] M. Chaudry, N. Jenkins, M. Qadrdan, and J. Wu, "Combined gas and electricity network expansion planning," vol. 113, no. C, pp. 1171–1187, 2014.

[8] P. Schwartz, *The Art of the Long View: Paths to Strategic Insight for Yourself and Your Company*. Random House, 1996.

[9] Y. Feng, "Scenario generation and reduction for long-term and short-term power system generation planning under uncertainties," Theses, Iowa State University, 2014.

[10] A. Wald, "Contributions to the theory of statistical estimation and testing hypotheses," *The Annals of Mathematics*, vol. 10, no. 4, pp. 299–326, 1939.

[11] L. Savage, "The theory of statistical decision," *Journal of the American Statistical Association*, vol. 46, p. 55–67, 1951.

[12] J. v. Neumann, "Zur theorie der gesellschaftsspiele," *Mathematische Annalen*, vol. 100, no. 1, pp. 295–320, 1928. [Online]. Available: http://dx.doi.org/10.1007/BF01448847

[13] S. Pohekar and M. Ramachandran, "Application of multi-criteria decision making to sustainable energy planning—a review," *Renewable and sustainable energy reviews*, vol. 8, no. 4, pp. 365–381, 2004.

[14] D. Dubois and H. Prade, "Possibility theory," in *Computational Complexity*, R. A. Meyers, Ed. Springer New York, 2012, pp. 2240–2252. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-1800-9_139

[15] M. D. Grigoriadis and L. G. Khachiyan, "A sublinear-time randomized approximation algorithm for matrix games," *Operations Research Letters*, vol. 18, no. 2, pp. 53–58, Sep 1995.

[16] B. von Stengel, "Computing equilibria for two-person games," in *Handbook of Game Theory*, R. Aumann and S. Hart, Eds. Amsterdam: Elsevier, 2002, vol. 3, pp. 1723 – 1759.

[17] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[18] M. N. Katehakis and A. F. Veinott Jr, "The multi-armed bandit problem: decomposition and computation," *Mathematics of Operations Research*, vol. 12, no. 2, pp. 262–268, 1987.

[19] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: the adversarial multi-armed bandit problem," in *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*. IEEE Computer Society Press, Los Alamitos, CA, 1995, pp. 322–331.

[20] J. Liu, O. Teytaud, and T. Cazenave, "Fast seed-learning algorithms for games," in *International Conference on Computers and Games*. Springer, 2016, pp. 58–70.

[21] T. Cazenave, J. Liu, F. Teytaud, and O. Teytaud, "Learning opening books in partially observable games: using random seeds in phantom go," in *2016 IEEE Conference on Computational Intelligence and Games*. IEEE, 2016, pp. 1–7.

[22] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[23] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[24] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[25] O. Teytaud and S. Flory, "Upper confidence trees with short term partial information," in *Applications of Evolutionary Computation*. Springer, 2011, pp. 153–162.

[26] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Algorithmic Learning Theory*. Springer, 2009, pp. 23–37.

[27] D. Auger, J. Liu, S. Ruette, D. Saint-Pierre, and O. Teytaud, "Sparse binary zero-sum games," in *Asian Conference on Machine Learning*, 2015, pp. 173–188.

[28] R. Busa-Fekete and B. Kégl, "Fast boosting using adversarial bandits," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 143–150.