Learning the optimal joint operation of the energy systems of Uruguay, Brazil, Paraguay and Argentina

Ruben Chaer
Facultad de Ingeniería
Universidad de la República
Electricity Market Administration
Montevideo, Uruguay
rchaer@simsee.org

Ignacio Ramírez
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
nacho@fing.edu.uy

Vanina Camacho
Administración del Mercado Eléctrico
Montevideo, Uruguay
vcamacho@adme.com.uy

Ximena Caporale
Facultad de Ingeniería
Universidad de la República
Administración del Mercado Eléctrico
Montevideo, Uruguay
xcaporale@adme.com.uy

Gonzalo Casaravilla
Facultad de Ingeniería
Universidad de la República
gcp@fing.edu.uy

***Abstract*—In the continuous fight against Bellman's Curse of Dimensionality, this work presents the first steps towards learning the Optimal Operation Policy of the electricity generation system of Uruguay, Brazil, Paraguay and Argentina with the infrastructures projected for the year 2030. The Operation Policy under consideration involves 76 state variables: one associated to the surface temperature anomaly of the Pacific Ocean in the N34 area, and 75 related to the hydroelectric reservoirs. The proposed methodology includes the design and training of two alternate neural network architectures combined with modern techniques devised for variance reduction and exploration, which were key to the success achieved.**

***Keywords*—Approximate Stochastic Dynamic Programming, Reinforcement Learning, Machine Learning, Optimal operation of hydrothermal systems.**

I. Introduction

THE optimal operation of hydro-thermal systems has been a challenge for decades for the operators of the electro-energy systems. This is a specially important topic for Latin American countries, which are characterized by a high hydroelectric component. Programming the operation of an electro-energy system involves determining which resources will be used to guarantee the supply of the energy demand in the following hours, days, months, years, so that the overall cost is minimized and quality and safety standard are complied with. In the presence of energy reservoirs (e.g., hydroelectric lakes), the problem becomes a Stochastic Dynamic Programming (SDP) problem. In 1957 Richard Bellman published [1] a detailed solution to the SDP problem which is now known as the Bellman recursion. In the same publication, Bellman stated that the solution suffers from what he called the "curse of dimensionality" which expresses that the proposed algorithm quickly becomes unusable with the increase in the dimension of the space of states of the system and the stochastic processes to be considered.

One of the seminal works in the fight against the Bellman Curse is [2], in which the technique known as Stochastic Dual Dynamic Programming (SDDP), widely used in Brazil, is developed. The approach used by SDDP is based approximating the Future Cost function based on a series of successive relaxations. In each iteration, the Lagrange multipliers associated with the dynamic constraint of the system allow, in theory, to adjust the representation of the Cost-to-go function (also called state value function) by adding, to the approximate representation, planes tangent to said function. The method is elegant because it has a convergence criterion associated with obtaining an upper and lower bound for the expected cost of the future operation in the deterministic case. In the case of systems with relevant random components, the technique suffers from problems similar to those of the Bellman curse due to the need to apply successive approximations on a tree of possible future scenarios. Although some techniques have been proposed to reduce the variability of solutions [3] with the massive incorporation of variable renewable energy, the representation of stochastic processes becomes of paramount importance and therefore it is expected that the SDDP technique must be improved.

The present work was possible thanks to the financing received from ANII of Uruguay for the development of the project: FSE_1_2017_1_144926 "Investment planning with variable energies, network restrictions and demand management".

Another strategy to solve the SDP problem approximately is known as Rolling Horizons (RH) [4]. This strategy has advantages when it comes to resolving the use of resources in a relatively short time horizon. This method performs a forward exploration step after which, with the information gathered, the time step is advanced and a new exploration is performed with a new time horizon. The success of the strategy relies on the hypothesis that there is a time horizon after which the decisions of the present have no effect and therefore it is possible to decide a present action by evaluating the consequences over a limited horizon. In practice, in systems with reservoirs capable of storing energy for months or years, the aforementioned hypothesis clearly does not hold, and the Operation Policy achieved is far from optimal.

Traditionally, the problems derived from Bellman's Curse of Dimensionality were exclusively of hydro-thermal system operation. Nowadays, with the trend towards decarbonized systems, variable energies and energy stores, the operators of all systems are joining the cause against this curse [5], [6] and [7].

In [8], after comparing different Approximate Dynamic Programming alternatives, the authors conclude that "none of these techniques works reliably in a way that would scale to more complex problems". Our work challenges this view by developing a system capable of operating a complex electro-energetic system in a satisfactory way By learning an optimum Operation Policy (OP) through a reinforcement learning loop.This work can be sseen as an extension of [9] , where the authors presented a success case of learning the optimal operation of an energy system in Uruguay. In the present work, the problem is scaled to consider the joint operation of the systems of Uruguay, Brazil, Paraguay and Argentina, with the infrastructure projected for the year 2030. The work [9] implies the learning of an OP in a state space of 6 dimensions, whereas in the present case, the state space has 76 dimensions.

II. NOTATION

The dynamic of the power system is modeled as in (1), where k is an integer that identifies the time-step, X_k is the state-vector of the system at the beginning of the step k , r_k is the vector of non-controlled inputs (like rainfall, wind, etc.) and u_k is the vector of controllable inputs (typically the power to be delivered for each generation unit, or power line).

$$X_{k+1}=f\left(X_k,r_k,u_k,k\right) \tag{1}$$

The funcion (2) represents the cost of operation during the step k as the sum of the fuel consumed by the thermal generators, the imports minus the exports, and any other operational cost including the cost of rationing, in the event that not all the energy demand is fulfilled.

$$c_k=c\left(X_k,r_k,u_k,k\right) \tag{2}$$

The OP (3) is a mapping that assigns a control vector u_k to different values of the system state and the non-controlled variables at step k .

$$u_k=OP\left(X_k,r_k,k\right) \tag{3}$$

Let $J\left(X_s,k+1\right)$ be the state-value function. This function represents the expected value of the optimal future operation beginning at state X_s. The *Optimal Operation Policy* is the one that minimizes the expected value of the sum of (2) and $J\left(X_s,k+1\right)$ and corresponds to the solution of the optimization problem:

$$J\left(X,k\right)=\left\langle \min_{u}\left\{ce\left(X_k,u_k,r_k,k\right)+J\left(X_s,k+1\right)\right\} \;@\left|\begin{matrix}u\in\Omega\left(X_k,r_k,k\right)\\ X_s=f\left(X_k,u_k,r_k,k\right)\end{matrix}\right.\right\rangle_{r_k} \tag{4}$$

Notice that equations (1)-(3) assume that the non-controlled inputs are known at the begging of each time-step, but nothing is assumed about the future of them.

The non-controlled inputs are modeled as random process without memory (white noise) with given distributions. If the random processes need to be modeled with dynamics, a corresponding model with its state variables is incorporated in (1) and the white noise that feeds such models is represented in the r_k vector. Notice that, whether or not to consider dynamics in the random processes involved may depend on the time scale of the steps. For a time-step of one hour, the wind power must be represented as a process with dynamics because the wind can not change at the same time in all wind farms of the country during one hour. There is a strong correlation of the wind power between consecutive hours. But if the time-step is a week or a month, representing the wind power as a process with dynamics does not make sense.

III. The learning loop

Knowing (1) and (2) and having an initial estimation of J, it is possible to simulate possible realizations of the operation of the system. In our case, to perform the simulations we use the SimSEE [10] platform. Carrying out a set of

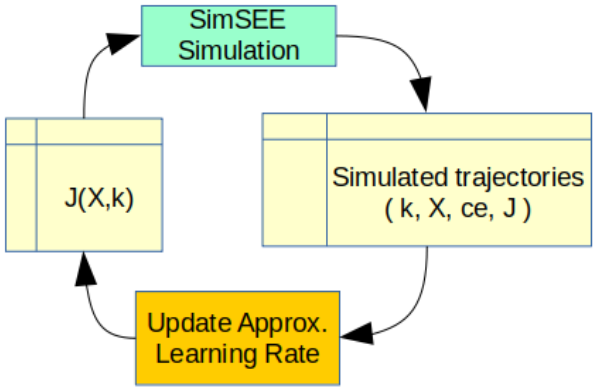

Fig. 1: The learning loop.

simulations, adding for each trajectory the costs actually incurred, a new estimation of the function J is obtained. Fig. 1 shows that learning loop.

At the end of each set of simulations there are trajectories of the type: (X_{ki}, J^h_{ki}) where the subscript $k=1..N_{Steps}$ denotes the time step and the index $i=1...N_{Trajectories}$ the number of the simulated trajectory (or realizations). Each trajectory is determined by an initial state X_{1i} and by a random seed that uniquely determines the performance of all stochastic processes during the simulation of that trajectory. The J^h_{ki} values are calculated from the simulation result and the $J^{h-1}(X,k)$ estimate using the equation:

$$J^h_{ki} = \sum_{p=k}^{p=k+n_{TD}} q^{(p-k)} ce_{ki} + q^{n_{TD}+1} J^{h-1}(X^s_{(k+n_{TD}+1,i)}, k+n_{TD}+1) \quad (5)$$

Where q is the money discount factor and ce_{ki} is the cost incurred in stage k of the simulation of trajectory i computed with (2) and $X^s_{(k+n_{TD}+1,i)} = f(X_{(k+n_{TD},i)}, r_{ki}, u_{ki}, k)$, that is the state projected by (1).

The n_{TD} (number of time-difference steps) determines the numbers of ce_{ki} added in the sum of the (5).

So far the proposed formulation is standard, the details described below on the initialization of the trajectories, the evolution of the state of the system in sections of n_{TD} steps within the trajectories, the use of Common Random Numbers and the modeling of the variations of the state value function are, in our opinion, what makes the difference between whether the Robot learns or not.

A. Initial states, random seeds and trajectories

At the beginning of each simulation stage, a set of initial states and random seeds is fixed and the trajectories are initialized with the cartesian product of both sets. For reasons of variance reduction, the information of each random seed is treated separately. This brings with it the need to consider a set of different initial states, so that the trajectories associated with the same random seed are different and therefore manage to collect information.

B. Where is the OP information?

Note that in determining the control vector u_k in the problem of (4), the absolute value of function J is irrelevant. The control solution is the same if we add a constant value to J . The OP's information is found in directional derivatives $\frac{\partial}{\partial X} J$. Taking this into consideration, the representation of J is adjusted to represent the differences $J_{ki} - J_{kj}$ for each set of new information instead of $J(X_{ki}, k)$.

C. Common Random Numbers (CRN)

As shown, the important thing is to represent the differences of the value function of the state by the movement of the state (possible in a time step as a consequence of the control action). In the case of generation systems, the J-value function is a distribution with a huge dispersion compared to the possible variation of its expected value by movement of the

state in a time step. For this reason, the comparison of the $J(X_k^i,k)-J(X_k^j,k)$ differences through simulations requires the use of variance reduction techniques such as the use of CRN. This is a key aspect, it is decisive!, as is mentioned in [11].

D. State evolution mode

In order to apply (5), the system must evolve following the dynamics of (1) during simulation. Fixing the value of n_{TD} in our algorithm, it is determined during how many steps, within each simulated trajectory, the system will evolve according to (1). For ex. if $n_{TD}=30$, the trajectories will start with an initial state and an average of 30 steps will be simulated, evolving the system according to (1) and after those 30 steps, the position of the system will be fixed (for each trajectory) in a new randomly determined state to continue with the system dynamics for another 30 steps and so on until the entire simulation horizon is covered.

In each iteration of the learning loop, (5) carries information n_{TD} steps from the future to the present. In that sense, the longer the time horizon considered, the more iterations of the learning loop will be necessary for $J^h(X,k)$ to reflect the future consequences of moving the state. In this sense, increasing the n_{TD} parameter would seem convenient. In the example case presented, with multi-annual reservoir stocks, the consequences of the decisions are observed for at least the following 3 years. With a weekly step simulation, if $n_{TD}=1$ were set this would imply at least 52*3 = 156 iterations of the learning loop for the relevant future information to reach the present at least once. On the other hand, if $n_{TD}=156$ were set, in one iteration there would already be information on the possible future consequences within 3 years. But possibly, depending on the time constants associated with the state variables, the trajectories associated with the same random seed, although associated with different initial states, converge to a single trajectory, thus losing the ability to collect information on the spatial differences of the state value function. This is why it is important that the n_{TD} value be less than the emptying time of the lakes for which an operation policy is to be formed.

In the implementation carried out, the sections of steps with evolution according to the dynamics of the system are randomly offset for each group of trajectories associated with the same random seed at the beginning of each iteration of the learning loop. With this, it is possible to cushion more quickly the effects caused by the consideration of said sections.

IV. Parametric network series

Luckily for us, the signals and processes involved in the planning of energy dispatch usually exhibit smooth regular patterns. This can be exploited to impose parsimonious approximations which extrapolate reasonably to unseen states. Our proposed method combines the flexibility of Neural Networks (NNs) with prior information about the problem. In a nutshell, the value function, which is a function of state and time, is approximated by a time-step neural network. The architecture of the network is the same for all time slots, reflecting the fact that the structure of the system itself does not change abruptly. The parameters vary across the networks, although in a controlled way: the variation of each parameter is penalized during the training process.

The general idea is depicted in Fig. 2.

In generic form, we can represent the estimate of the value function of iteration h as:

$$J^h(X,k)=M(X,k,\theta_k) \tag{6}$$

Where θ_k are vectors of fitting parameters for our model, that are trained by minimizing the following loss function:

$$L=\sum_{k,g} L_{kg}+\lambda\sum_k \|\theta_k\|^2+\beta\sum_{k=2}\|\theta_k-\theta_{k-1}\|^2 \tag{7}$$

Where, the elements of the first sum have the expression:

$$L_{kg}=\frac{1}{4N^2}\sum_{i\neq j\in g}((M(X_{kj},\theta_k)-M(X_{ki},\theta_k))-(J_{kj}-J_{ki}))^2 \tag{8}$$

Where g is the set of indexes that identify the trajectories associated with each random seed. As already mentioned, the information collected during the simulation is used to adjust the model based on the spatial differences of the value function associated with the same random seed.

The second summation in (7) corresponds to a Ridge regularization on the set of parameters with weight λ .

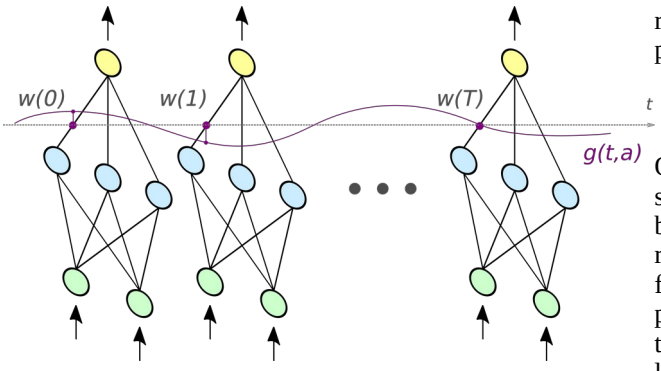

Fig. 2: Parametric network series.

And finally the third summation corresponds to the regularization that penalizes the abrupt variation of the model parameters with the passage of time with weight β .

The learning starts with a Null Policy (NP) defined as: $J^0(X,k)=0$ for all state and time. This would be the Operation Policy with zero derivatives in all directions of the state space. The NP is not as silly as it may seem at first glance, because the operating restrictions of the hydroelectric lakes are represented in the dispatch problem as restrictions with penalties for going below certain levels, which put at risk the availability of power from hydroelectric plants, and by restrictions that penalize the operation at high levels due to the effects of flooding of the lake on the surrounding lands. The penalties are established in MUS$/(m.day) which leads to the fact that even with the NP, the operation is generally within the reasonable operation zones.

The expected value of the future cost of operation using the NP, estimated by simulating the operation on a set of 100 realizations of the stochastic processes, resulted in: kMUS$ 28 (twenty-eight thousand million dollars). Value that will be used as a point of comparison of the operation policies learned by the Robot, based on simulations on the same set of realizations.

A. *Why not a single net?*

Naturally, one could use a single parametric function to model the whole value function across all time steps. We carry out tests defining a single NN for the entire time horizon, adding inputs representing the time and the first five annual harmonics to allow the Robot to have the notion of the time location and adapt to the annual cycles. Two time inputs marking the distance (with exponential decay) to the start and end of the simulation horizon were added to allow the Robot to adapt to the availability of forecasts (at the start) and to the end of the horizon (end times, emptying of the lakes). After 200 iteration of the learning loop with this NN structure the achieved cost-to-go was of kMUS$ 26, this is a 7% reduction from the null policy.

The structure of a NN per time Fig.2 , with a hidden layer of 12 neurons followed by a output layer of one neuron. This structure achieved a cost-to-go of kMUS$ 21 after 20 iteration of the learning loop, this is a 25% reduction from the null policy.

From these two tests we said that the structure of Fig.2 is better that a single NN with time positioning inputs, but we will still continue testing both structures, since for some systems in particular, in shorter time horizons such as weekly programming, the structure of a single neural network may be competitive.

Additionally, the possibility of using the β parameter to soften the control actions indicated by the operation policy is an attractive instrument for system operators. As an example, it is not to be expected that the value of water in a reservoir can change radically from one hour to the next. This type of regularization on the parameters can be introduced by the model structure organized in a time-step model.

V. Modeling of Electro-Energy Systems

The complete generation of the four countries (Uruguay, Argentina, Paraguay and Brazil) was modeled, identifying the following regions, which can be identified in Fig.3 , in the modeling: **UY**: Uruguay. **AR_ComPat**: South of Argentina including Comahue and Patagonia demand. **AR_Mer**: The rest of Argentina (Center and North). Corresponds to the area of greatest demand. **PY**: Paraguay. **BR_SE**: Brazil South-East. Corresponds to the area of greatest demand. **BR_S**: South Brazil. It corresponds to the area of Rio Grande do Sul including Porto Alegre. **BR_NE**: Brazil Northeast. **BR_N**: North Brazil. The **BR_Fic** region is a fictitious region used to reflect the restriction of transmissions within Brazil.

The detailed parameters of the different generators were obtained from the different sources detailed below. Likewise, the information corresponding to the historical series of Demand and water flows to the different hydroelectric plants that allowed the construction of the corresponding stochastic models.

For Uruguay, the model (SimSEE) was obtained from the ADME website [12] corresponding to the 2022 Supply Guarantee Report that models the system until the year 2030.

For Argentina, information was obtained from the CAMMESA seasonal programming database [13], supplemented with information from the Energy Transition Plan [14] of the Ministry of Energy to have the system configuration for the year 2030 (scenario called REN 20).

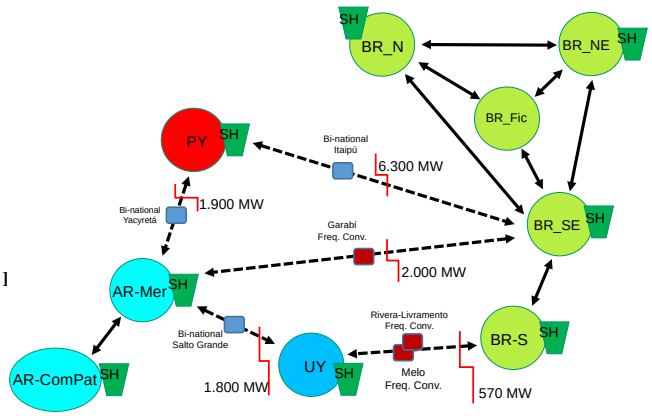

Fig. 3: Set of interconnecting nodes and arcs used in SimSEE to model the four countries.

For Paraguay, the two most important hydroelectric plants, Itaipú and Yacyretá, were included when modeling the system of Brazil and Argentina respectively. Two additional hydroelectric plants were considered. The Acaray hydroelectric plant, modeled as a run-of-river plant, with 4 units of 70 MW, adding two units in 2030 and one hydroelectric plant, modeled with a reservoir, on the Iguazú River of 100 MW from the year 2028. The rest of the

information on the current system was obtained from the ANDE website, in particular, from the Generation and Transmission Master Plan 2021-2040 [15], [16].

The information for Brazil was obtained from the database of the monthly programming for December 2021 [17] published by the CCEE, covering until the year 2025. This configuration was complemented with the information from the Plano Decenal de Expansão de Energia 2031 [18].

VI. Modelling regional exchanges

Currently there are two types of exchanges between countries. Some based on agreements associated with the construction of large infrastructures, such as the Salto Grande, Itaipú and Yacyretá hydroelectric plants, and others based on occasional offers. The logic with which the latter occur has been changing and will surely change in the future.

Each country values the water in its reservoirs, prioritizing its use for national demand, and only those resources that do not compromise national supply are traded in occasional exchanges.

In order to give a signal about the resources available for both modalities, it was decided to evaluate as occasional exchanges those that occur between the countries when the difference in marginal costs exceeds 60 US$/MWh. Additionally, the energy spilled by each country but that could have been generated (turbinable spills hydraulic, wind, solar, etc.) is quantified.

VII. Some results

The Fig.4 shows the expected value of the annual marginal cost, by time slot, in each area represented (Nodes in Fig.3). The resulting values are compatible with those projected for 2030 by the respective countries. In the case of Paraguay, the marginal cost is almost equal to that of the south-eastern region of Brazil, to which it is strongly interconnected. Since Paraguay has a surplus in energy, the marginal cost reflects the loss of income from exports for the supply of 1 MWh of incremental demand.

The Fig.5 shows the volumes of energy that make up the annual balance for the four countries.

The Sink category corresponds to the energy spilled that could have been turbined if, for example, the occasional exchange were allowed with a difference of marginal costs lower than the simulated value of 60 US$/WMh.

	h00h06	h06h12	h12h18	h18h26	Daily
AR_ComPat	40.1	40.1	40.2	40.6	40.3
AR_Mer	41.0	40.9	41.0	41.4	41.1
BR_N	17.1	17.2	17.8	21.5	18.4
BR_NE	15.9	16.0	16.6	20.3	17.2
BR_NO	16.9	16.9	17.6	21.2	18.2
BR_S	17.2	17.2	17.9	21.7	18.5
BR_SE	17.0	17.1	17.7	21.5	18.3
PY	17.4	17.4	18.1	21.9	18.7
UY	46.8	48.0	48.4	50.5	48.4

Fig. 4: Expected value, by hourly band, of the annual marginal cost [US$/MWh].

The category TVC_0 corresponds to the energy from thermal power plants associated with co-generation or inflexible processes that are considered for dispatch with zero variable cost.

The item TVC_60 corresponds to the energy from thermal power plants, subject to centralized dispatch with a variable cost less than or equal to 60 US$/MWh. These plants are generally combined cycles fueled by natural gas.

The item TVC_150 corresponds to the energy from flexible thermal power plants, subject to centralized dispatch, with a variable cost greater than 60 US$/MWh and less than or equal to 150 US$/MWh. These plants are generally combined cycle fueled with diesel or turbines or motor-generators fueled with diesel or natural gas.

The Fig.6 shows the expected value of international exchanges between the four countries for the year 2030.

	Demand	Exports	Sink	Transmission losses	Wind	Solar	Hydro	TVC_0	TVC_60	TVC_150	Imports	Rationing
Argentina	17997	111	0	285	2564	743	4650	365	8760	0	1311	0
Brasil	93381	80	3878	203	12566	6500	63521	12657	531	325	1434	8
Paraguay	4227	2619	0	3	0	0	6844	0	0	0	0	5
Uruguay	1537	17	149	18	580	84	675	284	0	0	98	0

Fig. 5: Annual energy balance by country [mean-MW].

Exports					
	Brasil	1.9	x	0.0	78.0
	Paraguay	1308.9	1310.5	x	0.0
	Uruguay	0.0	16.7	0.0	x

Fig. 6: Annual excanges [mean-MW].

VIII. Conclusions

This paper presented an application case of reinforcement learning applied to the optimal dispatch of a hydrothermal generation system in a group of four countries with a high integration of hydraulic, wind and solar renewable energies; showing that it is possible to achieve reasonable operating policies even in systems of the size and level of randomness proposed. It is highlighted that for the success of the learning of the Robot, the use of the Common Random Numbers technique was key to reduce the variance given the level of randomness that the state value function presents, mainly due to the randomness of hydroelectricity in the countries. considered.

The use of an exploration technique was also key, during the simulations, in which the dynamics of the system are followed during stretches of pre-established length, at the end of which the system is positioned in random states, within the possible space, with a distribution that attempts to arrive at a uniform sampling of the values of the state value function.

IX. FUTURE WORKS

One of the best ways to identify/guide future work is to start by identifying the difficulties and limitations encountered.

The main difficulty in the implementation carried out is in the resolution of the dispatch problem in a single MIP-Simplex with dimensions of more than 5000 variables and of the order of 4000 restrictions. The SimSEE platform has its own MIP-Simplex solver and for this work it was implemented, that in case the own routine failed, it would call the commonly used routine of the GLPK project. It was not possible to use only the GLPK routine, since it fails in more cases than SimSEE's own and sometimes remains in a loop, hanging the Robot. This seems like a minor technical detail, but for those of us who need a working implementation it is not. The conclusion is that the problem must be divided into dispatch zones, interconnected by transmission channels, and the dispatch problem must be solved iteratively. We identify this implementation as one of the necessary future tasks due to the technical feasibility of covering major problems and, additionally, this implementation would resolve a limitation imposed today regarding the sub-division of the time-stop into equal time bands for all countries. If you go to a solution by zones with hourly known power-flows by the interconnections, it is possible to use the subdivision into time bands with different definitions for each zone.

The second observation, on learning loop convergence and on the selection of the structure of neural networks. Since the learning process is easily parallelizable, it would be easy to launch a cloud of Robots learning the same problem, but selecting a brain (representation of the state value function) in each iteration of the learning loop based on the shared results among all the Robots, thus mixing reinforcement learning with the well-known genetic algorithms in order to have a mechanism for improving the representation in addition to learning its parameters.

X. DISCLAIMER

The content of this article is entirely the responsibility of its authors, and does not necessarily reflect the position of the institutions of which they are part of.

REFERENCES

[1] R. Bellman, *Dynamic Programing*. Princeton University Press, 1957.

[2] M. V. F. Pereira and L. M. V. G. Pinto, “Multi-stage stochastic optimization applied to energy planning,” *Math. Program.*, vol. 52, no. 1–3, pp. 359–375, May 1991, doi: 10.1007/BF01582895.

[3] M. P. Soares, A. Street, and D. M. Valladão, “On the solution variability reduction of Stochastic Dual Dynamic Programming applied to energy planning,” *Eur. J. Oper. Res.*, vol. 258, no. 2, pp. 743–760, Apr. 2017, doi: 10.1016/j.ejor.2016.08.068.

[4] É. Cuisinier, P. Lemaire, B. Penz, A. Ruby, and C. Bourasseau, “New rolling horizon optimization approaches to balance short-term and long-term decisions: An application to energy planning,” *Energy*, vol. 245, p. 122773, 2022, doi: https://doi.org/10.1016/j.energy.2021.122773.

[5] W. B. Powell, *Approximate Dynamic Programming*. Wiley, 2011.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*, Second edition. Cambridge, Massachusetts: The MIT Press, 2018.

[8] D. R. Jiang, T. V. Pham, W. B. Powell, D. F. Salas, and W. R. Scott, “A comparison of approximate dynamic programming techniques on benchmark energy storage problems: Does anything work?,” in *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, Orlando, FL, USA, Dec. 2014, pp. 1–8. doi: 10.1109/ADPRL.2014.7010626.

[9] R. Chaer *et al.*, “Teaching a Robot the optimal operation of an Electrical Energy System with high integration of renewable energies,” in *2021 IEEE URUCON*, Montevideo, Uruguay, Nov. 2021, pp. 364–367. doi: 10.1109/URUCON53396.2021.9647311.

[10] “SimSEE.” IIE-FING-UdeLaR, 2022. [Online]. Available: https://simsee.org

[11] J.-J. Christophe, J. Decock, J. Liu, and O. Teytaud, “Variance Reduction in Population-Based Optimization: Application to Unit Commitment,” in *Artificial Evolution*, vol. 9554, S. Bonnevay, P. Legrand, N. Monmarché, E. Lutton, and M. Schoenauer, Eds. Cham: Springer International Publishing, 2016, pp. 219–233. doi: 10.1007/978-3-319-31471-6_17.

[12] “Informe de Garantía de Suministro 2022.” ADME. [Online]. Available: https://www.adme.com.uy/informes/garantia_suministro.html

[13] “Programación Estacional Nov2021-Abr2022.” CAMMESA. [Online]. Available: https://cammesaweb.cammesa.com/visual-margo/

[14] “Lineamientos para un Plan de Transición Energética al 2030.” Ministerio de Economía, Secretaría de Energía, Argentina, Oct. 2021. [Online]. Available: https://www.argentina.gob.ar/economia/energia/planeamiento-energetico

[15] “PLAN MAESTRO DE GENERACIÓN 2021-2040.” ANDE, Feb. 2021. [Online]. Available: https://www.ande.gov.py/documentos/plan_maestro/PLAN%20MAESTRO%20DE%20GENERACION%20%202021-2040.pdf

[16] “Plan Maestro de Transmisión. Período 2021–2040.” Administración Nacional de Electricidad (ANDE)., Feb. 2021.
[17] “Base de datos del programa NEWAVE, correspondiente a la programación Diciembre 2021.” CCEE, Dec. 2021. [Online]. Available: https://www.ccee.org.br
[18] “Plano Decenal de Expansão de Energia 2031.” Ministerio de Minas y Energía, Secretaria de Planejamento e Desenvolvimento Energético, Empresa de Pesquisa Energética, Brasil, 2022. [Online]. Available: https://www.epe.gov.br/pt/publicacoes-dados-abertos/publicacoes/plano-decenal-de-expansao-de-energia-2031