

# Modelling of a fully renewable energy grid with hydrogen storage: time aggregation for a scalable Capacity Expansion Problem

Bianca Urso · Gabor Riccardi

Received: date / Accepted: date

**Abstract** In recent years, the integration of renewable energy sources into electrical grids has become a critical area of research due to the increasing need for sustainable and resilient energy systems. To address the variability of wind and solar power output over time, electricity grids expansion plans need to account for multiple scenarios over large time horizons. This significantly increases the size of the resulting Linear Problem (LP), making them computationally challenging for large scale grids. To tackle this, we propose an approach that aggregates time steps to reduce the problem size, followed by an iterative refinement of the aggregation. **We provide sufficient conditions under which the aggregated problem is equivalent to the original, unaggregated one, and refine the time intervals that do not meet these conditions. write better: Additionally, we introduce a validation function to assess the feasibility of the aggregated solutions. Our method is tested on a fully renewable energy grid with hydrogen storage.** We generate scenarios for wind and photovoltaic (PV) power output using scenario generation techniques that capture temporal dependencies. These dependencies are modeled with marginal distributions coupled with a Gaussian copula, ensuring that the generated scenarios reflect realistic temporal correlations observed in historical data.

**Keywords** First keyword · Second keyword · need 4-6

**Mathematics Subject Classification (2020)** 90-10 · 90B15

---

B. Urso  
IUSS School of Advances Studies, Palazzo del Broletto, Piazza della Vittoria, 15 – 27100 Pavia PV, Italy  
Tel: +39 0382 375811  
Fax: +39 0382 375899  
E-mail: bianca.urso@iusspavia.it

G. Riccardi  
Dipartimento di Matematica "F. Casorati" Via Adolfo Ferrata, 5 – 27100 Pavia

## 1 INTRODUCTION

**Context; literature overview. Small introduction on LP/MIP. Definition of stochastic optimization, definition of CEP, ED problems.**

The threat of climate change is pushing policy-makers to pursue greater integration of renewable energy sources into electrical grids, while at the same time ensuring reliability and resilience through digital optimization of electric power distribution and transmission in smart grids [?]. One of the main difficulties arising when designing an electric power system relying on renewables is the great variability of the generation of electricity through wind and solar, since these resources are highly dependent on weather conditions, **curtailment = inefficient, need hydrogen, cite source for hydrogen (eg: EU super investing in it). The stochastic nature of the problem....** making it impossible to plan long-term by optimizing on forecasts, and requiring a statistical approach to ensure a robust model. Up to now, common approaches have adopted Stochastic Programming (SP) or Robust Optimization (RO) models, along with hybrid models involving Information Gap Decision Theory or Chance Constraint [1]. While initially favored, the SP approach comes with high computational burden, so RO models have seen more popularity in recent years, despite the drawback of being conservative methods with higher average cost of operation and planning of energy systems.

In the typical setting, the problem to solve is a Capacity Expansion Problem (CEP) regarding infrastructure investments: solar and wind farms, fuel cells, hydrolizers, grid upgrades to augment Net Transfer Capacity (NTC) and so on. Nested within the CEP is an Economic Dispatch (ED) problem concerning the operational costs of said infrastructure. The problem is well suited to be modelled through mixed integer linear programming (MILP), as is explained in detail in [?] and ((**other examples?**))

The CEP for investment planning requires to look at long time horizons, and on the other hand intra-day variability in generation **is the main complexity driver** for the ED problem, so the time horizon must be modelled by a large quantity of fairly tight time steps. Furthermore, **spacial structure big, even though we ignore it here and just slam everything into country nodes**. Thus the temporal and spacial characteristics of the model bring the MILP size to increase rapidly. This is especially demanding in the case of SP, since all the variables from the inner ED problem must be reproduced over all scenarios.

To reduce these costs, one possible approach is to use a Rolling Horizon (RH) **[explain better]**. The basic technique is described in [?], along with some results regarding quality guarantees for the optimality of the solution. In [?], a rolling horizon approach is used within a RO model to optimize the operation of a micro-grid composed of **check** serving an isolated area in Chile. **other examples of RH used for the ED** A similar idea is applied in [?], where integer variables representing capital investments are initially relaxed and then progressively fixed in successive time steps, reducing the computational costs associated to the search for integer solutions.

A big drawback of the RH approach is that **Problem: it is not optimal. In fact, [?]. But better reflects actual decision making based on info that is only available**

progressively, vs perfect foresight approach.

Our work aims to build a hybrid model, exploring the use of the RH method as a tool to guide progressively tighter relaxations of the perfect foresight model, rather than as a stand-alone technique. **Explain better**. Further, RH is then used for the validation of the results for the CEP obtained in the perfect foresight optimization, to ensure that given a solution to the CEP, the ED problem admits feasible solutions even in a limited-foresight environment. The idea is to solve the CEP as to build a grid that can be operated as a "smart grid", **through a control system that optimizes on day ahead forecasts**.

The model we optimize on in **introduce model: components, aim, use Gurobi...**

Finally, to generate the scenarios needed to train and test the SP model, a Gaussian Copula method was used. This has been previously done **here and here and it's not new right?**.

This paper is organized as follows: **HOW?**

## 2 MODEL

### 2.1 LP FORMULATION

Our model describes a network in Europe that is to be powered and supplied of hydrogen through power generated by photovoltaic panels and wind turbines, converted to hydrogen through electrolysis and potentially reconverted in fuel cells.

The network is represented by an undirected multigraph  $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ , where  $\mathcal{N}$  corresponds to the nodes in the network, and  $\mathcal{E} = \mathcal{E}_H \cup \mathcal{E}_P$  represents transmission lines ( $e \in \mathcal{E}_P$ ) and hydrogen lines ( $e \in \mathcal{E}_H$ ). Each of these nodes can be in different countries in Europe, and the power generated by wind and solar power depends on the node location. In particular, if node  $n \in \mathcal{N}$  is in France, the scenarios for power generation at  $n$  will be generated using parameters fitted to France's data.

Each node has its generators, hydrogen storage, fuel cells and hydrolyzers, for which the capacity is to be decided. Likewise, the CEP aims to solve for transmission line NTC and hydrogen pipe transmission capacity for each edge of the network.

The basic formulation of the LP **described here** allows us to solve the **CEP with perfect foresight**. **Note: we consider continuous variables and then round up even for ns and nw because it decreased computational costs a lot without changing results much. Wouldn't be possible when optimizing over micro-grids.**

The model takes in input the generation and load scenarios of the given area along with various parameters indicating costs and efficiency of the current state of technology and possible upper bounds for the decision variables. The CEP and the ED problems for all scenarios are solved concurrently. When optimizing over multiple scenarios jointly, the solver returns the minimal amount of infrastructure and capacities that is needed to have feasibility (that is, demand met at all time and no blackouts) over all scenarios in input, with minimal cost. Cost is considered to be the sum of cap-

**Table 1** Decision variables

Name	Unit	Description
$ns_n$	-	Number of solar units at node $n \in \mathcal{N}$
$nw_n$	-	Number of wind units at node $n \in \mathcal{N}$
$nh_n$	kg	Storage capacity at node $n \in \mathcal{N}$
$mhte_n$	kg	Maximum hydrogen to electricity capacity, <i>i.e. the maximum amount of hydrogen that needs to be converted within a 1h time frame into electricity</i> at node $n \in \mathcal{N}$
$meth_n$	MWh	Maximum electricity to hydrogen capacity at node $n \in \mathcal{N}$
$addNTC_l$	MWh	Additional net transfer capacity on line $l$ ;
$addMH_l$	kg	Additional hydrogen transfer capacity on pipe $l$
$H_{j,t,n}$	kg	Stored hydrogen at node $n$ , time $t$ , scenario $j$
$HtE_{j,t,n}$	kg	Hydrogen converted to electricity at time $t$ , scenario $j$
$EtH_{j,t,n}$	kg	Electricity converted to hydrogen at time $t$ , scenario $j$
$P\_edge_{j,t,l}$	MWh	Power passing through line $l$ at time $t$ , scenario $j$
$H\_edge_{j,t,l}$	kg	Hydrogen transported through line $l$ at time $t$ , scenario $j$

ital costs and average operational costs of the infrastructure over the scenarios.

The optimization problem is solved using the Gurobi solver [4].

### 2.1.1 Decision Variables

See Table 2 for the summary of all decision variables. The main variables that are of interest to the policy maker and are explicitly returned in output are the following: Stored hydrogen is considered to be the total of liquid and gas hydrogen to be stored. Our model does not assume a distinction between the two forms, and considers hydrogen to be immediately ready for long-term storage as soon as it is converted from electricity, as well as instantaneously convertible to electricity in fuel cells at need. To give an indication of how much hydrogen should be kept in gas form at fuel cells, ready to be converted, and how much storage for gas hydrogen should be available at electrolyzers to accomodate abundant generation moments, we consider the following decision variables as well (N.B.: they are values *per hour*, to be compared to the time necessary for gas to liquid hydrogen transformation and vice versa): To accomodate for possible improvements on capacity of existing power lines or hydrogen transport infrastructure, the following variables are added: And parameters for the cost of such new infrastructure have to be given, indexed by edge of the respective graph:  $cNTC_l$  and  $cMH_l$ .

The second stage variables, indexed by scenario  $j$ , time step  $t$ , and node  $n \in \mathcal{N}$  are: Note: all variables are set to be non-negative.

### 2.1.2 Parameters

There are a series of parameters that characterize the model and can be modified by the policy maker through the GUI. The main ones are related to capital costs of the infrastructure to be built:

If not specified through the GUI, the following values are assumed for panels and turbines:  $cs = 400\text{€}$ ,  $cw = 3\,000\,000\text{€}$ . In this model we assume no marginal costs for PV and wind power production: the operating costs of the farms throughout their life-cycle can be factored into the capital costs, and there is no additional cost linked to the production itself. We assume the flow of electricity has no marginal cost nor power loss (the modelling of that problem is beyond the scope of this project). Conversely, we do set a cost for the use of hydrogen pipes (or other means of transfer):

Conversely, the marginal costs of conversion within electrolyzers and power cells are relevant. Thus one can set the following parameters:

According to the European Hydrogen Market Landcape November 2023 Report [2], “Hydrogen production costs via electrolysis with a direct connection to a renewable energy source in Europe vary from 4.18 to 9.60 €/kg H<sub>2</sub> of hydrogen, with the average for all countries being 6.86 €/kg H<sub>2</sub>”. For electrolyzers, we consider the Levelised cost of hydrogen (LCOH) to account for both marginal costs and capital costs. Such cost is dependent on the country’s specific market condition and can be calculated through the European Hydrogen Observatory tool. Unless specified through the GUI,  $ceth = 20\text{kg/MWh} \cdot 10\text{€/kg} = 200\text{€/MWh}$  is assumed (see the discussion on conversion efficiency below), and  $chte$  is assumed to be 2€/kg. Furthermore, the storage of the hydrogen itself has a cost that depends on various factors: capital cost of the technology used for storage, operating costs, length of time that the hydrogen is kept in storage. Thus the following parameters can be set:

Thus  $ch$  will be simply multiplied by the maximum storage needed ( $nh$ ), representing capital cost of storage infrastructure, whereas  $ch_t$  represents the marginal cost of keeping the hydrogen stored. Unless specified, it is assumed to be  $ch = 10\text{€/kg}$  and  $ch_t = 0\text{€/kg}\cdot\text{h}$

Within the electrolyzers and fuel cells, the conversion itself can be more or less efficient:

It is assumed that 1kg of hydrogen has an energy value of 33kWh. Thus if we consider an electrolyzer operating at maximum efficiency ( $feth = 1$ ), one MWh of electricity yields  $1000/33 \simeq 30\text{kg}$  of hydrogen. If not specified through the GUI, a value of  $feth = 0.66$  is considered, thus 1MWh yields 20kg of hydrogen. Conversely, in a fuel cell operating at maximum efficiency ( $fhfe = 1$ ) 1kg of hydrogen yields 33kWh. If not specified through the GUI, a value of  $fhfe = 0.75$  is considered, yielding 24.75kWh per kg of hydrogen. Actual efficiencies vary a lot depending on the technology used. Furthermore, chemical and physical constraints make it so that efficiencies higher than 0.80-0.85 are currently considered unachievable [?].

Finally, the GUI gives the option to place upper bounds to the variables, based on either technological and physical constraints (dimension of the facilities) or because of political choices (local population unfavourable to wind turbines):

If no bound is given through the GUI, these values will be set to  $Mns = 10^6$ ,  $Mnw = 500$ ,  $Mnh = 10^9$ ,  $Mhte = 10^6$ ,  $Meth = 10^5$ ). Note: computation time increases significantly when increasing the upper bound for  $Mnw$ .

**Table 2** Model parameters

Name	Unit	Description
$cs_n$	€	Cost of one Solar Panel at node $n$
$cw_n$	€	Cost of one Wind Turbine at node $n$
$cH\_edge_l$	€	Cost of transferring 1kg of hydrogen through edge $l$
$chte$	€/kg	Conversion cost of $H_2$ to electricity
$ceth$	€/MWh	Conversion cost of electricity to hydrogen
$ch_n$	€/kg	Cost of hydrogen storage capacity
$ch\_t$	€/(kg·h)	Cost of storing hydrogen for 1h, per unit of hydrogen
$fhte$	(scalar between 0 and 1)	efficiency of hydrogen to electricity conversion
$feth$	(scalar between 0 and 1)	efficiency of electricity to hydrogen conversion
$Mns$	(integer)	Maximum number of solar panels that can be installed
$Mnw$	(integer)	Maximum number of wind turbines that can be installed
$Mnh$	kg	Maximum hydrogen storage capacity
$Mhte$	kg	Upper bound for $mhte$
$Meth$	MWh	Upper bound for $meth$
$ES_{j,t,n}$	MWh	Power output of a single solar panel
$EW_{j,t,n}$	MWh	Power output of a single wind turbine
$EL_{j,t,n}$	MWh	Electricity load
$HL_{j,t,n}$	kg	Hydrogen load
$P\_edge_{j,t,l}$	MWh	Power passing through line $l$ during time step $t$ in scenario $j$
$H\_edge_{j,t,l}$	kg	Hydrogen flowing through pipe $l$ during time step $t$ in scenario $j$
$NTC_l$	MWh	Net Transfer Capacity, that is, maximum amount of electricity that can pass on line $l$ of the electric grid in the span of 1h;
$MH_l$	kg	Maximum amount of hydrogen that can flow on edge $l$ in 1h.

A scenario consists in a different realizations of the following variables, given as input in the model and indexed by scenario  $j$ , time step  $t$ , and node  $n \in \mathcal{N}$ :

Lastly the following parameters are indexed by edge of the respective graph:

### 2.1.3 Objective Function

The cost function is given by the sum of all capital costs of installing infrastructure, all marginal costs of the hour-by-hour hydrogen to electricity and electricity to hydrogen conversions, and minimal costs associated to the variables  $mhte$  and  $meth$  so that they are minimized through the model. Let  $Nnodes$ ,  $NEedges$  and  $NHedges$  be the number of nodes, edges on the electric grid graph and edges of the hydrogen transfer graph

respectively. The objective function is modified as follows:

$$\begin{aligned}
\min \quad & \sum_{k=1}^{Nnodes} cs_k \cdot ns \cdot cw_k \cdot nw_k + ch_k \cdot nh_k + \\
& + \sum_{l=1}^{NEdges} cNTC_l \cdot addNTC_l + \sum_{l=1}^{NHedges} cMH_l \cdot addMH_l + \\
& + \frac{1}{d} \sum_{j=1}^d \sum_{i=1}^{inst} \left( \sum_{k=1}^{Nnodes} (ch_{t,k} \cdot H_{j,t,k} + chte_k \cdot HtE_{j,t,k} + ceth_k \cdot EtH_{j,t,k}) + \right. \\
& \quad \left. + \sum_{l=1}^{NHedges} (cH\_edge_l \cdot H\_edge_{j,t,l}) \right) + \\
& + \sum_{k=1}^{Nnodes} 0.01 \cdot (mhte_k + meth_k)
\end{aligned}$$

The  $1/d$  factor in front of the marginal costs allows to average over the scenarios, whereas the capital costs are the same for all scenarios. Thus, ignoring the costs of  $mhte$  and  $meth$ , the objective function value gives an estimate of the actual costs (in €) for the set up and maintenance of the system throughout the length of the scenario (one year).

#### 2.1.4 Constraints

The following constraints are to ensure that for all time steps  $t$  and all scenarios  $j$ , the electricity load and the hydrogen load are met. The measure units are MWh and kg respectively, conversion factors are considered for  $HtE$  and  $EtH$  respectively. Let  $Out(n)$  and  $In(n)$  indicate the outgoing and incoming edges from node  $n$  on the respective graph. Then for each node  $n$ , we have the following flow balance constraints:

$$\begin{aligned}
\text{Electricity Balance: } \quad & ns_n \cdot ES_{j,t,n} + nw_k \cdot EW_{j,t,n} + 0.033 \cdot fhthe_k \cdot HtE_{j,t,n} - EL_{j,t,n} - EtH_{j,t,n} + \\
& \sum_{l \in Out(n)} P\_edge_{j,t,l} + \sum_{l \in In(n)} P\_edge_{j,t,l} \geq 0;
\end{aligned}$$

$$\begin{aligned}
\text{Hydrogen Storage: } \quad & H_{j,t+1,n} = H_{j,t,n} + 30 \cdot feth_k \cdot EtH_{j,t,n} - HtE_{j,t,n} - HL_{j,t,n} - \\
& - \sum_{l \in Out(n)} H\_edge_{j,t,l} + \sum_{l \in In(n)} H\_edge_{j,t,l}
\end{aligned}$$

We ask that the consumed electricity be less or equal than the produced electricity at all times. On the grid itself, the two sides should be equal, but we observe that  $ns \cdot ES_{j,t} + nw \cdot EW_{j,t}$  indicate the maximum power that can be generated with set weather conditions, whereas actual production will be regulated to meet demand through curtailment.

The stored hydrogen at time  $t + 1$  is the result of what was stored at time  $t$  adjusted by what was converted and what was sent to the industrial load. For  $t = 24 * 365$  we set the same constraint on hydrogen by considering  $t + 1$  to be index 1: this way we

avoid placing a “start time” at an arbitrary place within the year (time is rendered modulo the year) and we avoid the model asking for conveniently high initial storage values of hydrogen appearing out of thin air.

The total storage and conversion capacities are calculated by minimizing the maximum over time and scenarios of the variables  $H_{j,t}$ ,  $EtH_{j,t}$  and  $HtE_{j,t}$ , for all scenarios  $j$ , time steps  $t$  and nodes  $n$ :

$$\begin{aligned} \text{Storage Capacity Limit: } H_{j,t,n} &\leq nh_n; \\ \text{EtH Conversion Limit: } EtH_{j,t,n} &\leq meth_n; \\ \text{HtE Conversion Limit: } HtE_{j,t,n} &\leq mhte_n. \end{aligned}$$

Finally, edge capacities on the respective graphs are considered for all scenarios  $j$ , time steps  $t$  and nodes  $n$ :

$$\begin{aligned} \text{Net Transfer Capacity: } |P_{\text{edge}_{j,t,l}}| &\leq NTC_l + \text{addNTC}_l; \\ \text{Hydrogen Transfer Capacity: } |H_{\text{edge}_{j,t,l}}| &\leq MH_l + \text{addMH}_l. \end{aligned}$$

### 3 Optimization and Time Resolution

The time horizon generated by the scenarios has a time resolution where each time step has a length of one hour. Each value represents the total power (hydrogen) production or demand in the corresponding hour at the node. The smaller the length of each time step, the more accurate the results. However, the number of variables and constraints grows linearly with the number of time steps, making the model intractable (especially in the context of an application) with just a few scenarios.

Moreover, considering every hour in each day of the year is partly redundant, as each day will be similar to neighboring days. Yet, simply considering a sample of days for each season might undermine long-term storage capacity representation.

Given an initial time horizon  $\mathcal{T} = \{1, \dots, T\}$ , we can consider partitions of  $\mathcal{T}$  as a family of disjoint subsets whose union is  $\mathcal{T}$ . We only consider those partitions where every subset is an interval of  $\mathcal{T}$ . We refer to these as time partitions. Given a time partition  $P$ , we can consider the corresponding model obtained by considering each interval in  $P$  as a single time step. For every  $I$  in  $P$ , we define:

$$ES_{j,I,n} := \sum_{i \in I} ES_{j,i,n}, \quad EW_{j,I,n} := \sum_{i \in I} EW_{j,i,n}$$

and similarly for  $HL_{j,I,n}$  and  $HR_{j,I,n}$ . We denote the model obtained by the time partition  $P$  as  $CEP_P$ .

It is evident that the optimal value of  $CEP_P$  is a lower bound for the original problem  $CEP_{\mathcal{T}}$ , as given a feasible solution  $(ns, nw, nh, mhte, meth, H, HtE, EtH, Pedge, Hedge)$  of the latter, we can obtain a solution of the former by taking  $(ns, nw, nh, mhte, meth)$  the same as in  $CEP_{\mathcal{T}}$  and:

$$Pedge_{j,I,e} = \sum_{i \in I} Pedge_{j,i,e}, \quad Hedge_{j,I,e} = \sum_{i \in I} Hedge_{j,i,e}$$



and similarly for  $EtH$  and  $HtE$ , and  $H_{j,I,n} = H_{j,i_0,n}$  where  $I = [i_0, \dots, i_{|I|}] \in P$ . In particular, there is a cost-preserving linear map from the feasible space of  $CEP_{\mathcal{J}}$  to the feasible space of  $CEP_P$ , making the latter a relaxation of the former.

This is generally true when considering any time partition  $P'$  finer than  $P$ , where for every  $t' \in P'$ , there exists  $t \in T$  such that  $t' \subset t$ . In particular, we have the following observation:

**Observation 3.1** *Let  $V_{\mathcal{J}} \subset \mathbb{R}^{N_{\mathcal{J}}}$  and  $V_{P'} \subset \mathbb{R}^{N_{P'}}$  be the space of feasible solutions of  $CEP_{\mathcal{J}}$  and  $CEP_{P'}$ , respectively. There exists a linear map  $L_{P'P} : \mathbb{R}^{N_{P'}} \rightarrow \mathbb{R}^{N_P}$  such that  $L(V_{P'}) \subset V_P$  and  $c_P(L(x)) = c_{P'}(x)$ , where  $c_P$  is the cost function of  $CEP_P$  and  $c_{P'}$  is the cost function of  $CEP_{P'}$ .*

Thus, by iteratively solving finer time partitions, we converge to the optimal solution of  $\mathcal{J}$ .

### 3.1 Variables and Constraints aggregation

TODOS:

- When is the cost of the aggregated solutions equal to the cost of the original problem respect to the corresponding unaggregated solution? **done**
- When is the corresponding unaggregated solution optimal for the original problem? **done**
- How to extend obs so that it holds for ED.
- rimuovere che combinazioni devono essere convesse, fa casino e non cambia niente **done**
- How to get a scenario for which the aggregated ED solution is feasible
- How to easily compute when an aggregated interval is "far" from such feasible scenario
- Unaggregate on those intervals first. **done**
- Add examples
- If the non aggregated variables are mixed integers, everything works the same.
- For  $\rho_r$  constraints which behave well under generalised convex combinations of variables can be excluded. **done**

GR: Secondo me ci sta scriverlo in maniera un filo generale perchè: Così questo metodo non è ristretto a solo (ed esattamente) a questo modello, magari diventa chiaro che aggiungendo generatori tradizionali tutto funziona comunque e top. Oppure per problemi totalmente diversi ma con una struttura simile. Poi così alcune dimostrazioni si "semplificano", ad esempio qui il problema della conservazione dei costi nel caso le variabili sono negative non sorgeva perchè ci si è ristretti a una formulazione con variabili positive (alla quale ci si può poi ricondurre come abbiamo visto).

Varying time aggregation can be viewed as performing row and column aggregation on the original linear programming (LP) model. Consider the following general

linear problem:

$$\min_{x \in \mathbb{R}^n} c^T x \quad (1)$$

$$\text{s.t. } Ax = b \quad (2)$$

$$x \geq 0 \quad (3)$$

Here,  $A$  is an  $m \times n$  matrix. Now, let  $\sigma = \{S_1, S_2, \dots, S_{\tilde{n}}\}$  be a partition of  $[n]$  (the columns) and  $\delta = \{R_1, R_2, \dots, R_{\tilde{m}}\}$  a partition of  $[m]$  (the rows), corresponding to a partition of the rows and columns of  $A$ .

We obtain the corresponding aggregated problem by replacing each set  $S$  in  $\sigma$  with a single row, and each set  $R$  in  $\delta$  with a single column. One way to aggregate a set of rows (or columns) is by taking a linear combination of the rows (or columns), known as *weighted aggregation*. We denote the weights of the aggregation by  $\omega_r$  for  $r \in \sigma$ , and  $\tau_c$  for  $c \in \delta$ . The corresponding aggregated LP problem becomes:

$$\min_{\tilde{x} \in \mathbb{R}^{\tilde{n}}} \tilde{c}^T \tilde{x} \quad (4)$$

$$\text{s.t. } \tilde{A}\tilde{x} = \tilde{b} \quad (5)$$

$$\tilde{x} \geq 0 \quad (6)$$

where  $\tilde{A}$  is a  $\tilde{m} \times \tilde{n}$  matrix.

In the problem under consideration, we have various types of constraints: Electricity Balance, Hydrogen Balance, Hydrogen Storage, and bounds on the variables. Given a time partition  $P$ , we define  $\sigma$  and  $\delta$  such that each set  $S \in \sigma$  corresponds to all constraints of the same type, scenario, and time index  $t$  that falls within the same time interval in  $T$  as  $P$ . Similarly, the variables (such as Power generation, Hydrogen generation, etc.) are partitioned in  $\delta$  based on the same criteria. Rows and columns are combined via weight aggregation. This aggregation maintains the structure of the original problem, meaning that had we formulated the model directly with the aggregated time steps, we would have arrived at the same model. Before defining a *structure-preserving aggregation* for a general LP, we introduce some notation: Given a matrix  $B$  with row and column index sets  $I$  and  $J$ , respectively, for any subsets  $I' \subset I$  and  $J' \subset J$ , we denote the submatrix of  $B$  with rows in  $I'$  and columns in  $J'$  as  $B_{I', J'}$ .

Let  $\tilde{A}$  be formed by aggregating the rows and columns of  $A$  according to the partitions  $\sigma$  and  $\delta$ , respectively. For each  $R \in \sigma$ , we denote by  $\tilde{A}_R$  the row in  $\tilde{A}$  resulting from aggregating the rows of  $A$  corresponding to  $R$ , while  $A_R$  refers to the submatrix of  $A$  consisting of all rows in  $R$ . Similarly, for each  $C \in \delta$ , we define  $\tilde{A}_C$  as the column in  $\tilde{A}$  obtained by aggregating the columns of  $A$  in  $C$ , and  $A_C$  as the submatrix of  $A$  containing all columns in  $C$ . Thus,  $\sigma$  and  $\delta$  serve as the index sets for  $\tilde{A}$ .

For a family of sets  $F$ , we denote the subsets of  $F$  with size exactly  $k$  and greater than  $k$  by  $F_{=k}$  and  $F_{>k}$ , respectively. Specifically,  $\text{supp}(\tilde{A}_R)_{>1} \subset \delta$  represents the set of indices corresponding to partitions  $C \in \delta$  with size greater than 1, and  $\text{supp}(A_r)_{>1}$  refers to the set of indices where  $c \in C \in \delta$ , with  $C$  having a size greater than 1.

**Definition 3.2** Given an LP problem (1), we say that a weighted aggregation with respect to partitions  $\sigma, \delta$  is *structure-preserving* if for each  $R \in \sigma$  and each  $r \in R$ , there exists  $f^r : [\tilde{n}] \rightarrow [n]$  such that:

1.  $f^r|_{\text{supp}(\tilde{A}_R)} : \text{supp}(\tilde{A}_R)_{>1} \rightarrow \text{supp}(A_r)_{>1}$  is a bijection such that

$$\tilde{A}_{R,C} = A_{r,f^r(C)} \text{ for all } C \in \text{supp}(\tilde{A}_R)_{>1}$$

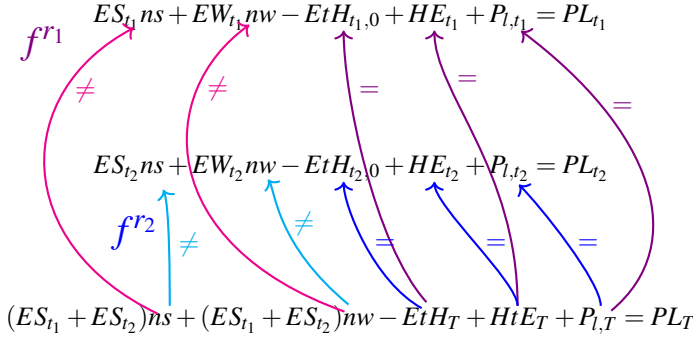
2. If  $f^{r'}(C') = f^r(C)$  then  $C = C'$ .
3.  $C = \{f^r(C)\}_{r \in R}$  for all  $R \in \sigma_{>1}$  and  $C \in \delta$ .

From condition 3 follows the following:

**Observation 3.3** For all  $\{c\} \in \delta_{=1}$  and constraints  $r \in R \in \sigma_{>1}$ ,  $f^r(\{c\}) = c$ .

*Example 3.4* As an example, we consider the function  $f^r$  corresponding to the time step aggregation in CEP as defined in ???. Let us examine the Power Balance constraints  $r_1$  and  $r_2$  for a fixed node  $n \in \mathcal{N}$  and time steps  $t_1$  and  $t_2$ , where  $T := \{t_1, t_2\} \in P$  represents an interval in the time partition of the aggregated problem. We note that both  $f^{r_1}$  and  $f^{r_2}$  satisfy the conditions outlined in Definition 3.2.

Condition ?? is met as all aggregated variables,  $EtH_T$ ,  $HtE_T$ , and  $P_{l,T}$ , are mapped (through violet arrows for  $f^{r_1}$  and blue arrows for  $f^{r_2}$ ) to unaggregated variables with the same coefficients. Condition 3 implies that all variables appearing in the two constraints are included in the image of either  $f^{r_1}$  or  $f^{r_2}$ . Finally, Condition ?? holds trivially.



This implies that the coefficients of the aggregated variables in the aggregated problem match those in the original problem for the corresponding unaggregated variables,  $f^r$  can be seen as a function mapping the aggregated variables to variables of the same "type" in the unaggregated constraint. While obtaining a feasible solution to (1) from (4) is not always guaranteed, it is possible under certain assumptions. .

**Observation 3.5** If  $(\sigma, \delta)$  is a structure-preserving aggregation, let  $R \in \sigma$  and  $r \in R$ . Let  $\tilde{x}$  be a solution to the aggregated problem (4). If  $\tilde{b}_r - \tilde{A}_{R,\delta_{=1}} \tilde{x}_{\delta_{=1}} \neq 0$ , define

$$\rho_r := \frac{b_r - A_{r,\delta_{=1}} \tilde{x}_{\delta_{=1}}}{\tilde{b}_r - \tilde{A}_{R,\delta_{=1}} \tilde{x}_{\delta_{=1}}}.$$

If  $A_{r,\delta_{=1}} = 0$  and  $b_r = 0$  for all  $r \in R$ , then  $\rho_r$  can be chosen arbitrarily.

If  $\rho_r \geq 0$  and  $x \in \mathbb{R}^n$  satisfies  $x_{\delta_{=1}} = \tilde{x}_{\delta_{=1}}$  and  $x_{f^r(C)} = \rho_r \tilde{x}_C$  for all  $C \in \text{supp}(\tilde{A}_R)_{>1}$ , then  $x$  satisfies the constraint  $A_r x = b_r$  of the original problem.

*Proof* Consider,  $A_r x = \sum_{i \in \text{supp}(A_r)} A_{r,i} x_i$ . This sum can be divided over the aggregated and unaggregated variables:

$$A_r x = A_{r,\delta_{=1}} x_{\delta_{=1}} + \sum_{c \in \text{supp}(A_r)_{>1}} A_{r,c} x_c. \quad (7)$$

If  $A_{r,\delta_{=1}} = 0$  and  $b_r = 0$  then  $A_{r,\delta_{=1}} x_{\delta_{=1}} = 0$ . Fix  $\rho_r \geq 0$ , then from the definition of structure-preserving aggregation, we know that  $f^r(\text{supp}(\tilde{A}_R)_{>1}) = \text{supp}(A_r)_{>1}$ , so equation (7) becomes:

$$(7) = \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} A_{R,f^r(S)} x_{f^r(S)} = \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} \tilde{A}_{R,S} \rho_r \tilde{x}_S = \rho_r \tilde{A}_R \tilde{x} = 0, \quad (8)$$

where the second equality holds because  $\tilde{A}_{R,S} = A_{r,f^r(S)}$  and  $x_{f^r(S)} = \rho_r \tilde{x}_S$ . Thus,  $x$  satisfies the constraint  $A_r x = b_r$ .

When  $A_{r,\delta_{=1}} \neq 0$  or  $b_r \neq 0$ , we proceed similarly:

$$(7) = A_{r,\delta_{=1}} x_{\delta_{=1}} + \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} A_{r,f^r(S)} x_{f^r(S)} \quad (9)$$

$$= A_{r,\delta_{=1}} \tilde{x}_{\delta_{=1}} + \rho_r \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} \tilde{A}_{R,S} \tilde{x}_S. \quad (10)$$

By the definition of  $\rho_r$  the first sum in the last line of equation (9) is equal to:

$$\begin{aligned} \rho_r \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} \tilde{A}_{R,S} \tilde{x}_S &= \rho_r (\tilde{A}_R \tilde{x} - \tilde{A}_{R,\delta_{=1}} \tilde{x}_{\delta_{=1}}) \\ &= \rho_r (\tilde{b}_R - \tilde{A}_{R,\delta_{=1}} \tilde{x}_{\delta_{=1}}) \\ &= b_r - A_{r,\delta_{=1}} \tilde{x}_{\delta_{=1}}. \end{aligned}$$

Thus, we obtain:

$$A_r x = b_r.$$

□

A structure-preserving aggregation does not inherently ensure the feasibility of all constraints in the original problem. However, Observation 3.5 demonstrates how to partially reconstruct a solution  $x$  for a specific constraint  $r$  by scaling the aggregated variables appropriately within the support of  $A_r$ .

**Definition 3.6** Let  $\rho_r$  be defined as in Observation 3.5 for all  $r \in R \in \sigma_{>1}$ . Let  $x \in \mathbb{R}^n$  be defined as  $x_{\delta_{=1}} := \tilde{x}_{\delta_{=1}}$  and  $x_{f^r(C)} := \rho_r \tilde{x}_C$  for all  $C \in \delta_{>1}$  and  $r \in R \in \sigma_{>1}$ . Then  $x$  is well defined if for all  $r, r' \in R \in \sigma_{>1}$  such that  $f^r(C) = f^{r'}(C)$ , we have  $\rho_r = \rho_{r'}$ . In such case we refer to  $x$  as a *rho*-solution.

If  $x$  is a  $\rho$ -solution.  $x$  is a feasible solution for the constraints in  $\sigma_{>1}$ , we also need to ensure that  $x$  is also feasible for the remaining constraints in  $\sigma_{=1}$ .

*Example 3.7* Consider the constraint that the initial hydrogen stored must be equal to the final hydrogen stor for a one node network:

$$\sum_{t=1}^n \Delta H_t = 0 \quad (11)$$

The corresponding aggregated constraint is for a time partition  $P$  is:

$$\sum_{T \in P} \Delta H_T = 0 \quad (12)$$

For all  $\rho$  such that  $\sum_{t \in T} \rho_t = 1$  for all  $T$  in  $P$ , given a feasible solution for the aggregated problem  $\Delta \tilde{H}_T$ , let  $\Delta H_t := \rho_t \Delta \tilde{H}_T$ , then constraint (11) holds:

$$\sum_{t=1}^n \Delta H_t = \sum_{t=1}^n \rho_t \Delta \tilde{H}_T = \sum_{T \in P} \sum_{t \in T} \rho_t \Delta \tilde{H}_T = \sum_{T \in P} \Delta \tilde{H}_T = 0 \quad (13)$$

Thus constraint (11) holds for  $\rho$ -solutions

This is a special instance of a general class of constraints that always hold for  $\rho$ -solutions, this follows from the following propriety of  $\rho$ -solutions:

**Observation 3.8** Let  $\omega_r \in \mathbb{R}$  for all  $r \in R \in \sigma$  be the weights of the row aggregation. If  $x$  is a  $\rho$ -solution, then we have, for all  $R \in \sigma_{>1}$ :

$$\omega_R^T \rho_R = 1 \quad (14)$$

*Proof*

$$\omega_R^T \rho_R = \sum_{r \in R} \omega_r \rho_r = \frac{\sum_{r \in R} \omega_r (b_r - A_{r, \delta_{=1}} \tilde{x}_{\delta_{=1}})}{\tilde{b}_r - \tilde{A}_{R, \delta_{=1}} \tilde{x}_{\delta_{=1}}} = 1$$

Note that if  $x$  is a  $\rho$ -solution, then for all  $C \in \text{supp}(A_r)_{>1}$ , we can pick  $R^{(C)} \in \sigma_{>1}$  so that  $C \in \text{supp}(\tilde{A}_{R^{(C)}})$  and  $x_{f^r(C)} = \rho_r \tilde{x}_C$  for all  $r \in R^{(C)}$  and the definition of  $x$  does not depend on the choice of  $R^{(C)}$ .

**Observation 3.9** Let  $(\sigma, \delta)$  be a structure-preserving, row and column aggregation. If  $x$  is a  $\rho$ -solution and  $r$  is a constraint in  $\sigma_{=1}$ , such that

$$A_{r, f^{r'}(C)} = \omega_{r'} \tilde{A}_{r, C} \text{ for all } r' \in R^{(C)}, C \in \delta_{>1},$$

then  $x$  is a feasible solution for constraint  $r$ .

*Proof* Since  $\omega_R^T \rho_R = 1$ , we have:

$$A_r x = \sum_{C \in \delta_{=1}} A_{r,C} x_C + \sum_{C \in \text{supp}(A_r)_{>1}} \sum_{r' \in R^{(C)}} A_{r,f^{r'}(C)} x_{f^{r'}(C)} = \sum_{C \in \delta_{=1}} \tilde{A}_C \tilde{x}_{r,C} + \sum_{C \in \text{supp}(A_r)_{>1}} \sum_{r' \in R^{(C)}} \omega_{r'} \rho_{r'} \tilde{A}_{r,C} \tilde{x}_C = \tilde{b}_r = b_r$$

We now define the hypergraph associated to the aggregation  $(\sigma, \delta)$ .

**Definition 3.10** The *hypergraph associated to the aggregation*  $(\sigma, \delta)$  is the hypergraph  $\mathcal{N}, \mathcal{E}$  having as nodes the aggregated variables  $\mathcal{N} := \delta_{>1}$  and as edges the subset of  $\mathcal{N}$  that appear together in aggregated constraints.

When two edges (constraints) in the hypergraph,  $r$  and  $r'$ , share aggregated variables, the scaling factors  $\rho_r$  and  $\rho_{r'}$  must be equal for Observation 3.5 to hold for both  $r$  and  $r'$ . From this follows the following:

**Proposition 3.11** *If  $(\sigma, \delta)$  is a structure-preserving aggregation and the constraints in  $\sigma_{=1}$  hold for rho-solutions. Let  $\tilde{x}$  be a solution to the aggregated problem (4). For all  $r \in R \in \sigma_{>1}$  define  $\rho_r$  as in Observation 3.5. If  $\rho_r \geq 0$  and is constant over the connected components of the hypergraph associated to  $(\sigma, \delta)$ . Then  $x_{\delta_{=1}} := \tilde{x}_{\delta_{=1}}$  and  $x_{f^r(C)} := \rho_r \tilde{x}_C$  for all  $C \in \text{supp}(\tilde{A}_R)$  and  $C \in \delta_{>1}$  is well defined and thus  $x$  is a  $\rho$ -solution. Further more if  $x$  is feasible for all constraints in  $\delta_{=1}$ , then  $x$  is feasible solution to the unaggregated problem (1).*

Until now we have only considered feasibility, ignoring the relationship between the cost of  $\tilde{x}$  and the cost of  $x$ . The following observation gives a condition for the cost of  $\tilde{x}$  to be equal to the cost of  $x$ . For all  $C \in \delta_{>1}$  let  $R^{(C)} \in \sigma_{>1}$  be so that  $C \in \text{supp}(\tilde{A}_{R^{(C)}})$  and  $x_{f^r(C)} = \rho_r \tilde{x}_C$  for all  $r \in R^{(C)}$ .

**Observation 3.12** *Let  $x$  be a  $\rho$ -solution. If  $\omega_r \tilde{c}_C = c_{f(r,C)}$  for all  $r \in R^{(C)} \in \sigma_{>1}$ , then the cost of  $\tilde{x}$  for the aggregated problem is equal to the cost of  $x$  in the unaggregated problem.*

*Proof* Let  $\tilde{x}$  be a solution to the aggregated problem (4). Using observation 3.8, for all  $C \in \delta_{>1}$  the cost corresponding to the variable  $\tilde{x}_C$  is

$$\tilde{c}_C \tilde{x}_C = \tilde{c}_C \left( \sum_{r \in R} \omega_r \rho_r \right) \tilde{x}_C = \sum_{r \in R} \tilde{c}_C \omega_r \rho_r \tilde{x}_C = \sum_{r \in R} c_{f(r,C)} x_{f(r,C)}$$

Which correspond to the cost of the variables  $\{x_{f(r,C)}\}_{r \in R}$ . Thus

$$\tilde{c}\tilde{x} = \sum_{C \in \delta_{=1}} \tilde{c}_C \tilde{x}_C + \sum_{C \in \delta_{>1}} \tilde{c}_C \tilde{x}_C = \sum_{C \in \delta_{=1}} c_C x_C + \sum_{C \in \delta_{>1}} \sum_{r \in R} c_{f(r,C)} x_{f(r,C)} = \sum_{C \in \delta_{=1}} c_C x_C + \sum_{j \in \cup_{C \in \delta_{>1}} C} c_j x_j = cx$$

While row aggregation of a linear problem is a relaxation of the original problem, the same does not apply to column aggregation. However, the column aggregation used for the Capacity Expansion Problem in this work is still a relaxation. In general a column aggregation of a linear problem is a relaxation of the original problem whenever it is a *constant-coefficients column aggregation*, that is:

**Definition 3.13** A column aggregation of a linear problem, respect to the column partition  $\delta$ , is a *constant-coefficients column aggregation* if, for all  $C \in \delta_{>1}$ , the non-zero rows in  $A_C$  are identical.

**Proposition 3.14** *If  $(\sigma, \delta)$  is a structure-preserving, constant-coefficients column aggregation, if the hypothesis of proposition 3.11 and observation 3.12 are true, then the aggregated problem (4) is exact and  $x$  is an optimal solution.*

Since for observation 3.12, the cost of the aggregated problem is equal to the cost of  $x$  in the unaggregated problem, we only need that the aggregated problem is a relaxation of the unaggregated problem. But this is true since both row aggregations and constant-coefficients column aggregations are relaxations.

### 3.2 Application to Capacity Expansion Problem

We now apply the results of the previous section to the Capacity Expansion Problem. TODO:

- What are the connected components of the hypergraph as defined in the previous sections? each connected component looks like the hypergraph of the ED at a fixed timestep. Why can we not consider the edges corresponding to hydrogen storage? because however  $\rho_s$  are chosen, they hold, so they don't force  $\rho$  to be equal over the nodes it connects.
- What does it mean? Given an aggregated problem, we are interested in which time intervals are problematic, since the ones which have well define  $\rho$  could be disaggregated and obtain the same solution, we disaggregate the ones in which the  $\rho_s$  are far from being well defines, that is we have two constraints in the same connected components with really different  $\rho_r$ .
- What is  $\rho_r$ , since the aggregated variables are only second stage variables. Let's consider  $\rho_r$  with  $r \in R$ , and let  $t_r$  and  $n \in \mathcal{N}$  be respectively the time step and the node corresponding to the constraint  $r$ . Let the net energy production at time  $t$  in  $n$ : that is demand in  $n$  minus any renewable power produced in  $n$ . Then  $\rho_r$  corresponds to fraction between the net energy production at time  $t$  and the net energy production during the interval  $T$  with  $t \in T$ .
- How to disaggregate: Thus we calculate  $\rho(n, t)$  for all nodes in the network and all timestep, and disaggregate those with highest variance for fixed  $t$ . (since, fixed  $t$  the should all be equal for the solution be feasible for the problem obtained by disaggregating the time interval  $T$ ).
- How much to disaggregate: One could either subdivide each time interval into finer time intervals, or to the single timesteps. For the former, one could keep the intervals in which the variance of  $\rho_r$  is small, together, and divide in single timesteps the rest.
- Initial aggregation: to calculate  $\rho_r$  we need to have solved already an aggregated problem, so it's not too insightful. But we can start by grouping together intervals by putting "peaks" on the extremes, that is peak of demand and production, and then subdividing equally the so obtained intervals.
- Write well how to calculate  $\rho_r$

By reformulating the problem in [ref](#), we introduce a new set of variables  $h_{t,j,n} := H_{t+1,j,n} - H_{t,j,n}$ , replacing the original variables  $H_{t,j,n}$ , and add the following constraints:

$$\sum_{t=1}^{\bar{t}} h_{t,j,n} = H_{\bar{t},j,n}, \quad \text{for all } \bar{t} \leq T, \quad (15)$$

It can be easily checked that:

**Observation 3.15** *The aggregation of the Capacity Expansion Problem (CEP) is a structure-preserving aggregation with constant coefficients. Furthermore, the constraints (15) are  $\rho$ -agnostic.*

Since the only constraints linking variables across different time steps are those in (15), the connected components of the hypergraph representing the aggregation correspond to the hypergraphs of the Energy Dispatch (ED) at each time step. Specifically, there is one connected component per time step, with nodes representing the variables at that step. To construct a feasible solution from the aggregated problem, it is sufficient for  $\rho_r$  to be constant across constraints within the same time step.

Using the definition of  $\rho_r$ , it can be shown that for each time step  $t \in T \in \mathcal{T}$ , this corresponds to the ratio between the net energy production at time  $t$  and the net energy production over the interval  $T$ . Hence, the conditions of Proposition 3.11 are met if, for a fixed time step  $t$ , this ratio is the same for all nodes in the network, and if, during the interval  $T$ , the net production at each node is consistently positive or consistently negative. Thus, we state:

**Observation 3.16** *If for each time step  $t \in T \in \mathcal{T}$ , the ratio of the net energy production at time  $t$  to the net energy production over the interval  $T$  is the same for all nodes in the network, and if, throughout the interval  $T$ , the net production at each node is either always positive or always negative, then the aggregated problem is exact.*

Although this is not generally the case, it suggests an iterative procedure for refining the solution of the aggregated problem. At each iteration, we refine the time partition of the aggregated problem, selecting the time interval to be refined based on the extent to which it violates the conditions of Observation 3.16. Specifically, we split intervals where the net production changes sign, and those intervals  $T$  in which, for a fixed time step  $t \in T$ , the ratio between the net energy production at time  $t$  and the net energy production over the interval  $T$  exhibits the largest variance.

### 3.3 SCENARIO GENERATION

To estimate the optimal capacities for the CEP through a stochastic approach, realistic and diverse weather scenarios are needed, so to capture the variability and uncertainty of power generation through renewable sources over extended periods. In order to generate such scenarios, samples are extracted from a joint probability density function (PDF) fit on historical data. [In the following subsection](#), we use  $Y_t$  to



denote the stochastic process of generated power observations for either solar or wind in a single country. In our project, we used an hourly time step ( $T=\{1...8760\}$ ) and fit the wind and solar distributions separately for each country considered. To model the marginal probability distributions corresponding to the power output of wind turbines for each hour of the year, a Weibull distribution was used, justified by its proven effectiveness in capturing the variability and skewness of wind power distributions [?]. For solar power, Beta distributions were employed, as in [?].

To fit our model, we used a dataset containing 30 years of data for various European countries, which was collected by [3]. On the other hand, electricity load is taken from the [ENTSO-E Statistical Reports](#). In this simple model, while fitting on historical data we did not account for possible changes in future climate, since the focus lies mostly in the computational aspect.

To account for interdependence between temporally near time steps, we coupled these distributions using a Gaussian Copula approach, which captures the dependencies between hourly power outputs effectively. This approach accurately mimics common weather phenomena: The Gaussian Copula represents well the coupled behavior in renewable stochastic systems [?].

A possible improvement of the generation process could be to fit wind and PV data jointly in the copula step, potentially also including load scenarios with the generation scenarios through the same approach. This would consider dependence between Energy Demand and weather conditions, but it would necessitate of the historical dataset provided for the corresponding grid, and would also further increase computational costs.

### 3.3.1 Parametric Estimation of Wind Power distribution

The parameters defining the Weibull Distribution are estimated using the Maximum Likelihood Estimation (MLE). The Weibull density function is given by:

$$f(x; \theta, \gamma) = \left(\frac{\gamma}{\theta}\right) x^{\gamma-1} \exp\left(-\left(\frac{x}{\theta}\right)^\gamma\right)$$

where  $\theta, \gamma > 0$  are the scale and shape parameters, respectively. Given observations  $X_1, \dots, X_n$ , the log-likelihood function is:

$$\log L(\theta, \gamma) = \sum_{i=1}^n \log f(X_i | \theta, \gamma)$$

The optimum solution is found by searching for the parameters for which the gradient is zero :

$$\frac{\partial \log L}{\partial \theta} = -\frac{n\gamma}{\theta} + \frac{\gamma}{\theta^2} \sum_{i=1}^n x_i^\gamma = 0 \quad (16)$$

Eliminating  $\theta$ , we get:

$$\left[ \frac{\sum_{i=1}^n x_i^\gamma \log x_i}{\sum_{i=1}^n x_i^\gamma} - \frac{1}{\gamma} \right] = \frac{1}{n} \sum_{i=1}^n \log x_i \quad (17)$$

This can be solved to get the MLE estimate  $\hat{\gamma}$ . This can be accomplished with the aid of standard iterative procedures such as the Newton-Raphson method or other numerical procedures. This is done with the aid of the package *scipy*. Once  $\hat{\gamma}$  is found,  $\hat{\theta}$  can be determined in terms of  $\hat{\gamma}$  as:

$$\hat{\theta} = \left( \frac{1}{n} \sum_{i=1}^n x_i^{\hat{\gamma}} \right)^{\frac{1}{\hat{\gamma}}} \quad (18)$$

### 3.3.2 Parametric Estimation of Solar Power distribution

To estimate the  $\alpha$  and  $\beta$  parameters defining the Beta distribution  $Y$ , we use the Method of Moments. The mean of the random variable  $Y$  can be expressed as  $\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta}$  and the variance as  $\text{Var}[Y] = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}$ . In particular by explicating  $\beta$  in the first equation and substituting it in the second equation we obtain that:

$$\begin{cases} \alpha = \mathbb{E}[X] \left( \frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\text{Var}[X]} - 1 \right) \\ \beta = (1 - \mathbb{E}[X]) \left( \frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\text{Var}[X]} - 1 \right) \end{cases} \quad (19)$$

By substituting the mean and the variance with their empirical approximation we obtain the Method of Moments estimator for  $\alpha$  and  $\beta$ .

### 3.3.3 Parametric Copula Estimation

The cumulative density function of both the Weibull and Beta distributions are continuous and invertible. Therefore, the random variables  $U_t := F_{Y_t}(Y_t)$  have a uniform distribution over  $[0, 1]$ . The copula of the random variables  $\{Y_t\}_{t \in T}$  is defined as the function  $C : [0, 1]^T \rightarrow [0, 1]$  such that

$$C(F_{Y_1}(y_1), \dots, F_{Y_T}(y_{|T|})) = P(Y_1 \leq y_1, \dots, Y_{|T|} \leq y_{|T|}). \quad (20)$$

This function always exists because of Sklar's Theorem [cite Sklar?](#). For a given correlation matrix  $\Sigma$ , the Gaussian Copula with parameter matrix  $\Sigma$  is defined as

$$C_{\Sigma}^{\text{Gauss}}(u_1, \dots, u_T) := \Phi_{\Sigma}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_T)),$$

where  $\Phi$ ,  $\Phi_{\Sigma}$  are the cumulative distribution functions of Gaussian variables having distribution  $\mathcal{N}(0, 1)$  and  $\mathcal{N}(\mathbf{0}, \Sigma)$  respectively. In particular if  $C_{\Sigma}^{\text{Gauss}}$  is the copula associated with the random variables  $\{Y_t\}_{t \in T}$  then we have that the random variables  $Z_t = \Phi^{-1}(F_{Y_t}(Y_t)) = \Phi^{-1}(U_t)$  have joint distribution equal to  $\mathcal{N}(0, \Sigma)$ . This follows from:

$$\begin{aligned} P(Z_1 \leq z_1, \dots, Z_T \leq z_T) &= P(\Phi^{-1}(U_1) \leq z_1, \dots, \Phi^{-1}(U_T) \leq z_T) = \\ &= P(U_1 \leq \Phi(z_1), \dots, U_T \leq \Phi(z_T)) = \\ &= C_{\Sigma}^{\text{Gauss}}(\Phi(z_1), \dots, \Phi(z_T)) = \\ &= \Phi_{\Sigma}(z_1, \dots, z_T) \end{aligned}$$

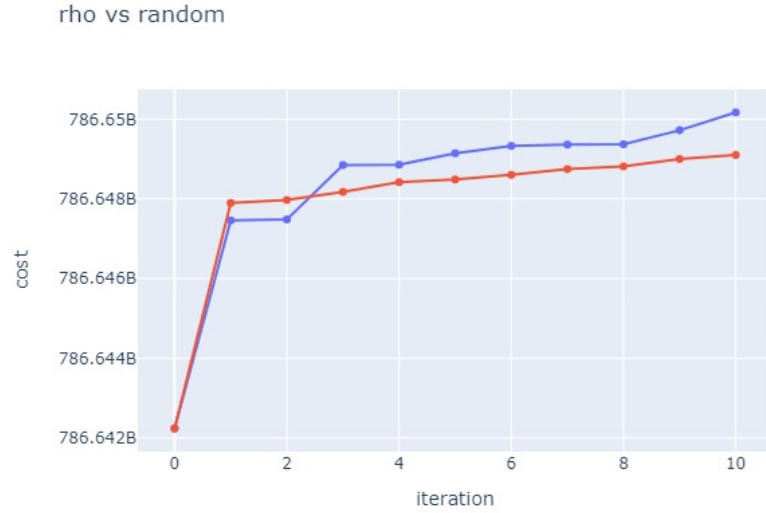
In particular, given the realization  $\{y_{t,j}\}_{t \in T, j \in J}$  of the variables  $\{Y_t\}_{t \in T}$ , an unbiased estimation of the parameter matrix  $\Sigma$  is the empirical covariance matrix  $\hat{\Sigma}$  of the samples  $\{\Phi^{-1}(\hat{F}_{Y_t}(y_{t,j}))\}_{t \in T, j \in J}$ , where  $\hat{F}_{Y_t}$  is the estimated marginal distribution of the variable  $Y_t$  as seen in subsection 3.3.1 and subsection 3.3.2.

Finally, we can generate samples from a Multivariate Gaussian random variable  $(Z_t, t \in T)$  having distribution  $\mathcal{N}(0, \hat{\Sigma})$ . Then the power output scenarios are obtained from these samples by following the previous steps backwards, that is, for each sample, computing  $\hat{F}_t^{-1}(\Phi(Z_t))$  for all  $t \in T$ .

#### 4 COMPUTATIONAL RESULTS

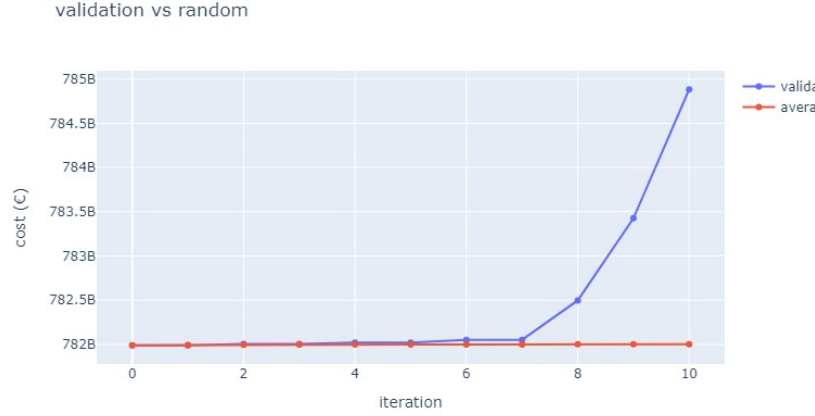
To evaluate the methodologies presented in this paper, we examine a 5-node network over a one-year span, utilizing timesteps of 1 hour across two distinct scenarios. In this section, we compare three approaches for iterating on the aggregated problem: (1) randomly selecting the interval for refinement, (2) selecting the interval with the highest  $\rho$ -variance as defined in Section ??, and (3) selecting the interval where the validation function ?? fails. The scenarios are generated as outlined in Section ?? in the Appendix. The computational tests were conducted on an Intel(R) Core(TM) i7-13700H CPU @ 2.40GHz with 16 GB of RAM.

The following plot illustrates the cost variation at each iteration using the  $\rho$  selection method compared to random interval selection. Notably, the  $\rho$  selection method demonstrates a faster increase in cost reduction than the random selection method, with comparable optimization times: 174 seconds for the  $\rho$  selection method and 155



seconds for the random selection method over 10 iterations.

In the subsequent plot, we compare the cost variation at each iteration using the validation function for interval refinement against the random interval selection method. The results indicate that the former method yields a significantly faster cost reduction, implying quicker convergence to the optimal solution. However, this approach incurs greater computational time, indicated as **many seconds s**. It is worth noting that while both the  $\rho$  and random iteration methods continue even upon encountering a feasible solution for the original problem, the validation function iteration may halt before reaching the maximum iteration limit if the current solution is



feasible, thereby being optimal for the original problem.

#### 4.1 SINGLE NODE NETWORK

First, an electrical grid with a single node is considered (corresponding ideally to an area with uniform weather conditions, highly connected at low cost). A first section will consider realistic parameter combinations and describe the results given by the solver, conducting a parameter sensitivity analysis. A second section will describe a validation function that checks the results of the capacity expansion problem for feasibility on new scenarios. Concurrently, a cost function is designed to give more realistic cost estimates compared to the optimal value given by the solver.

#### 4.2 MULTIPLE NODE NETWORK

Results are computed for a multiple node network, with additional edge variables and parameters. When considering a network with more than a single node, computational costs increase rapidly. Thus in the first section, a small analysis is carried out to determine acceptable time steps on which time dependent data can be aggregated (the gathered data is usually on hourly steps) while maintaining the quality of the solution. Some examples are then considered, and the network dynamics that arise with the introduction of edge variables are described. A mixed approach is then used

---

to design a validation function that can deal with the complexity arising from the introduction of the network structure in the model.

## 5 CONCLUSIONS

### References

1. Michal Jasinski, Arsalan Najafi, et al, Operation and Planning of Energy Hubs Under Uncertainty - A Review of Mathematical Optimization Approaches, 2022
2. Author, European Hydrogen Market Landscape - November 2023 Report, page numbers. Publisher, place (year)
3. Stefan Pfenninger and Iain Staffell. "Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data". In: Energy 114 (2016), pp. 1251–1265. issn: 0360-5442. doi: <https://doi.org/10.1016/j.energy.2016.08.060>.
4. Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2024). <https://www.gurobi.com>