

Towards time series aggregation with exact error quantification for optimization of energy systems

Beltrán Castro Gómez, Yannick Werner, and Sonja Wogrin

Institute of Electricity Economics and Energy Innovation, Graz University of Technology, Graz, Austria

{castrogomez,yannick.werner,wogrin}@tugraz.at

Abstract—Energy system optimization models are becoming increasingly popular for analyzing energy markets, such as the impact of new policies or interactions between energy carriers. One key challenge of these models is the trade-off between modeling accuracy and computational tractability. A recently proposed mathematical framework addresses this challenge by achieving exact time series aggregations merging time periods sharing the same active constraint sets. This aggregation, however, is insufficient when the number of unique active constraints is large. We overcome this issue by aggregating data points from different active constraint sets. While this further reduces model size, it inevitably introduces an error compared to the full model. Yet, we show how this error can be exactly quantified without re-solving the optimization problem, enabling users to trade off computational efficiency and model accuracy proactively. This may be especially useful in energy markets to accommodate varying granularity across short- and long-term time horizons.

Index Terms—optimization models, linear programming, data aggregation, greedy algorithms

I. INTRODUCTION

Energy system optimization models (ESOMs) represent highly complex systems in several dimensions, e.g., technological, temporal, spatial, or sectoral [1]. The complexity has risen in recent years, e.g., due to the integration of electricity from intermittent renewable energy sources (RES) [2], or the increasing importance of energy markets to accommodate, e.g., an increasing number of small participants, such as residential prosumers with photovoltaic and storage systems [3], [4], or new and potentially nonlinear sector coupling technologies, like electrolyzers [5], [6] or heat pumps [7], [8]. The main difficulty of large-scale ESOMs relies upon the trade-off between modeling detail and computational tractability [9]. Complexity reduction techniques can improve computational tractability, e.g., by reducing the length of the considered time horizon. To support efficient decision-making in energy markets and systems, however, preserving the outputs of the original ESOM throughout the aggregation process is essential. While we are convinced that the fundamentals of our method extend to other complexity dimensions, we will only focus on time series aggregation (TSA) [10] in this paper.

Among TSA methods for ESOMs, we can differentiate between two main categories: a-priori and a-posteriori methods [11]. Traditional a-priori TSA methods try to best approximate input time series data, e.g., hourly electricity demand, by mini-

mizing the difference between the original and the aggregated time series according to a given metric, such as Euclidean distance in k-means [12]. Such a-priori TSA methods can be improved for ESOMs, e.g., by adding *extreme* periods to the aggregated time series [13]–[15], or including extreme values during the aggregation process itself [14], [16], [17]. However, a-priori TSA methods may not necessarily lead to a good approximation of the ESOM outputs compared to the model with the complete time horizon – as illustrated in [18] – because they are unaware of how the optimization model relates input data to model outputs. A good TSA method should allow one to control the input data aggregation depending on the *output error*, defined between the *full* and *aggregated* optimization model outputs¹. Neglecting this relation may result in suboptimal solutions and poor decision-making in energy markets and systems.

On the other hand, a-posteriori TSA methods additionally consider optimization model outputs in the aggregation process. In [19], the authors propose an iterative method that estimates investment decisions by bounding the output error of the corresponding hourly operational problem. Reference [20] obtains representative days by clustering the cost of optimal investment decisions of individual days and [21] samples *extreme* and *regular* periods, measured based on aggregated model outputs. However, we identify a gap in a-posteriori TSA: the methods that we have found in the literature conceive heuristic components in the aggregation process, e.g., [20] relies on the assumption that similarity in objective function value relates to similarity in overall optimization outcome, and both [19], [21] involve preliminary a-priori clustering outcomes based solely on the input data domain. Thus, they fail to provide analytical expressions to compute exact output error bounds or provide metrics to relate the aggregation to the output error.

In previous work [18], an a-posteriori TSA framework that aggregates data points from time periods that share the same set of active constraints, referred to as *basis*, was proposed², and we use both terminologies interchangeably. Under certain assumptions for linear programs (LPs), this TSA method has been proven to be exact and yield zero error in aggregated model outputs [18]. However, if the number of unique active

¹We refer to the full model as the one that is based on the original time series, while the aggregated model uses the aggregated time series.

²Following the basic feasible solution terminology of the Simplex method.

constraint sets is large, the aggregation of data points solely within each set may only lead to a minor reduction in model size. We address this issue in this paper.

The original contributions of this paper are: we extend the aforementioned theoretical framework with a methodology to merge data points from different active constraint sets; this allows us to derive an exact analytical expression to quantify the associated output error without having to solve the new model. Using this expression, we elaborate on how to optimally choose the active constraint sets from which to merge data points considering the combinatorial complexity of this decision. The proposed methodology assists the modeler in efficiently navigating the trade-off between computational tractability and model accuracy.

The remainder of this paper is organized as follows: Section II recaps the TSA framework based on active constraint sets and introduces the optimal transport problem, which serves as an example. Section III extends the aforementioned TSA framework by proposing a methodology to merge data points from different active constraint sets. Section IV presents and discusses our experimental results to showcase our proposed method applied to the optimal transport problem. Finally, Section V concludes.

II. PRELIMINARIES

First, Section II-A shows how TSA can be performed based on active constraint sets. Afterward, Section II-B introduces the optimal transport problem.

A. TSA based on active constraint sets

Let $t \in \mathcal{T} = \{1, 2, \dots, T\}$ denote the set of time steps. For the sake of simplicity, we assume that there are no time-linking constraints³. Let us define a standard-form primal (left) and dual (right) linear program (LP):

$$\begin{aligned} \min_{x_t} \quad & \sum_{t \in \mathcal{T}} c^T x_t & (1a) \quad & \max_{y_t} \quad \sum_{t \in \mathcal{T}} b_t^T y_t & (2a) \\ \text{s. t.} \quad & Ax_t \leq b_t, \quad \forall t, & (1b) \quad & \text{s. t.} \quad A^T y_t \geq c, \quad \forall t, & (2b) \\ & x_t \geq 0, \quad \forall t, & (1c) \quad & y_t \geq 0, \quad \forall t, & (2c) \end{aligned}$$

where $c \in \mathbb{R}^N$, $x_t \in \mathbb{R}^N$, $A \in \mathbb{R}^{M \times N}$, $b_t \in \mathbb{R}^M$ and $y_t \in \mathbb{R}^M$. Without loss of generality, we assume that A and c are identical across all time steps. Moreover, we require that Problems (1) and (2) are feasible and that strong duality holds. Optimal solutions to decision variables are denoted by $\{\}^*$, e.g., $x_t \rightarrow x_t^*$.

Since Problem (1) is feasible, for every time step $t \in \mathcal{T}$, there is a set of active (also called binding) constraints $\mathcal{A}_t \subseteq \{1, 2, \dots, M\}$, that are satisfied with equality. Then, we define $i \in \mathcal{I}$ as the set of bases, i.e., unique active constraint sets, \mathcal{A}_i , such that $\mathcal{A}_i \neq \mathcal{A}_{i'}$ for any $i, i' \in \mathcal{I}$ with $i \neq i'$, and $\mathcal{T}_i \subseteq \mathcal{T}$, such that $\mathcal{A}_t = \mathcal{A}_i, \quad \forall t \in \mathcal{T}_i$. Note that since $\mathcal{T}_i \neq \emptyset, \forall i \in \mathcal{I}$,

³In related work [22], time-linking constraints such as ramping are included in the underlying data aggregation framework based on active constraint sets. Future work will address storage technologies. However, such time-linking constraints are out of the scope of this contribution.

and $\bigcap_{i \in \mathcal{I}} \mathcal{T}_i = \emptyset$, it always holds that $|\mathcal{I}| \leq |\mathcal{T}|$. However, ideally, we want $|\mathcal{I}| \ll |\mathcal{T}|$.

Using set \mathcal{I} related to the unique active constraint sets, we can aggregate Problems (1) and (2), respectively, to:

$$\begin{aligned} \min_{\bar{x}_i} \quad & \sum_{i \in \mathcal{I}} c^T \bar{x}_i W_i & (3a) \quad & \max_{\bar{y}_i} \quad \sum_{i \in \mathcal{I}} \bar{b}_i^T \bar{y}_i & (4a) \\ \text{s. t.} \quad & A \bar{x}_i \leq \bar{b}_i, \quad \forall i, & (3b) \quad & \text{s. t.} \quad A^T \bar{y}_i \geq c, \quad \forall i, & (4b) \\ & \bar{x}_i \geq 0, \quad \forall i, & (3c) \quad & \bar{y}_i \geq 0, \quad \forall i, & (4c) \end{aligned}$$

where $W_i = |\mathcal{T}_i|$ is the weight of basis i , i.e., the number of time steps with the same set of active constraints, and $\bar{b}_i = \frac{1}{W_i} \sum_{t \in \mathcal{T}_i} b_t$ the centroid of the data points in basis i .

As proven in [18], the optimal solutions to Problems (1) and (3) are identical in terms of objective function value and average of the decision variables $\frac{1}{W_i} \sum_{t \in \mathcal{T}_i} x_t^* = x_i^*, \forall i \in \mathcal{I}$. Considering the potentially much smaller dimension of Problem (3) compared to (1) when $|\mathcal{I}| \ll |\mathcal{T}|$, it is computationally preferable to solve Problem (3) when averages of the optimal decision variables suffice. Since we are assuming a constant cost function $c_t = c, \forall t \in \mathcal{T}$, all time steps \mathcal{T}_i associated with active constraint set i have the same optimal dual solution, such that $y_t^* = y_i^*, \forall t \in \mathcal{T}_i$. Therefore, when including weight W_i of basis i in Equation (3a), the dual variables of the aggregated Problem (4) change such that $\bar{y}_i^* = W_i y_i^*$.

B. Optimal transport problem: A case study

Let us introduce an example of a standard LP problem in energy markets: the optimal electricity transport problem. We define sets of generators $g \in \mathcal{G}$, nodes $n \in \mathcal{N}$, lines $l \in \mathcal{L}$ including one line per flow direction, and time steps $t \in \mathcal{T}$, with $|\mathcal{T}| = 8736$ hours corresponding to 52 weeks. The set \mathcal{G}_n and $\mathcal{L}_n^{\text{out/in}}$ collect the generators at and lines going in to (out of) node n . Let $p_{g,t}$ denote the power production of generator g in time step t , $f_{l,t}$ the power flow of line l in time step t , C_g the variable production cost of generator g , C_l^{TR} the transportation cost of line l , \bar{P}_g the installed capacity of generator g , $CF_{g,t}$ the capacity factor of generator g in time step t , \bar{F}_l the symmetric power flow limit of line l , and $D_{n,t}$ the power demand in node n in time step t . Then the optimal transport problem is given by:

$$\min_{p_{g,t}, f_{l,t}} \quad \sum_{g,t} C_g p_{g,t} + \sum_{l,t} C_l^{\text{TR}} f_{l,t} \quad (5a)$$

$$\text{s. t.} \quad 0 \leq p_{g,t} \leq \bar{P}_g CF_{g,t} \quad : \quad \underline{\mu}_{g,t}, \bar{\mu}_{g,t} \quad \forall g, t, \quad (5b)$$

$$0 \leq f_{l,t} \leq \bar{F}_l \quad : \quad \underline{\eta}_{l,t}, \bar{\eta}_{l,t} \quad \forall l, t, \quad (5c)$$

$$D_{n,t} = \sum_{l \in \mathcal{L}_n^{\text{in}}} f_{l,t} - \sum_{l \in \mathcal{L}_n^{\text{out}}} f_{l,t} + \sum_{g \in \mathcal{G}_n} p_{g,t} \quad : \quad \lambda_{n,t} \quad \forall n, t. \quad (5d)$$

The objective function in (5a) minimizes total system cost, which consists of generation and transmission costs. Constraints (5b) correspond to the lower and upper production limits of the generators. Constraints (5c) limit the line power flow in both directions. Constraint (5d) presents the nodal

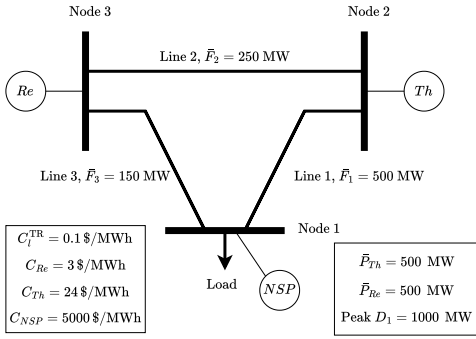


Fig. 1. Stylized 3-node case study diagram including line and generator parameters for the optimal transport problem (5).

power balance. Based on the dual variables, defined after the colon, we derive the dual objective function of Problem (5):

$$\max_{\Theta} \sum_{g,t} \bar{\mu}_{g,t} \bar{P}_g CF_{g,t} + \sum_{l,t} \bar{F}_l \bar{\eta}_{l,t} + \sum_{n,t} \lambda_{n,t} D_{n,t}, \quad (6)$$

where $\Theta = \{\bar{\mu}_{g,t}, \bar{\eta}_{l,t}, \lambda_{n,t}\}$. We present the numerical results for our case study based on Equation (6) in Section IV. We consider the 3-node system depicted in Figure 1 as a stylized but illustrative case study. In addition to generators *Renewable* (*Re*) and *Thermal* (*Th*), we also consider non-supplied power (*NSP*) as a generator⁴. There are two TS in the input data: the capacity factors of the wind generator, $CF_{Re,t}$; and the demand at Node 1, $D_{1,t}$.

After solving the optimal transport problem (5) for our case study, we observe eight unique sets of active constraints (bases), such that $\mathcal{I} = \{1, 2, \dots, 8\}$. We visualize the normalized time series input data for $CF_{Re,t}$ and $D_{1,t}$ in Figure 2, where every point represents a capacity factor-demand pair for a particular hour, colored according to their basis i . Table II in Appendix A provides intuitive descriptions for the corresponding active constraint sets. For example, for all hours in basis 2 (orange), the wind generator is the marginal generator (at node 1), and line 3 is congested.

III. BASES MERGING FRAMEWORK

In large-scale and complex modern energy systems, the cardinality of \mathcal{I} may be close to that of \mathcal{T} [22]. This compromises the efficiency of the aggregation and may, in the worst case, lead to an aggregated model that is still computationally intractable. Therefore, in this section, we extend the existing active constraint sets aggregation framework presented in the last section further to reduce the size of the aggregated optimization model. In particular, we describe in Section III-A how we can reduce the number of bases by merging data points from different active constraint sets and, in Section III-B, how this impacts the approximation accuracy compared to the full model output. Note that in this paper, we assume

⁴We assume *Th* and *NSP* generators to have a capacity factor equal to 1.

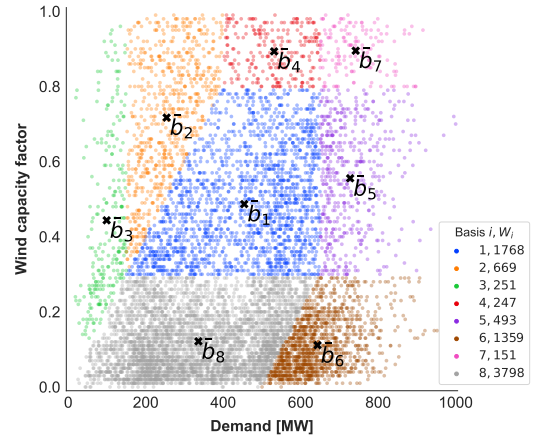


Fig. 2. Hourly demand $D_{1,t}$ and wind capacity factor $CF_{Re,t}$ time series, colored according to their basis i , relating input data space to model outputs. Crosses mark the centroid of each basis.

that the mapping of time steps to bases is known through past model runs. That is the only information required to run the aggregated Problem (3). In future research, we will investigate methods to acquire the mapping without solving the full optimization problem, e.g. parametric programming.

A. Definition of bases merging

The main idea of our proposed method is to reduce the number of bases in \mathcal{I} by merging data points from different bases and replacing them with a single centroid. With slight abuse of terminology, we refer to this as *bases merging*. In this paper, we only consider merging all data points of one basis with all data points of a single other basis. Consequently, when deciding which bases to merge, it is sufficient to only look at the centroids of the data points of each basis. While this may not necessarily be optimal, it helps us illustrate the general idea and derive some mathematical foundations of our proposed method. We will relax this assumption in future research.

We define the set of *bases mergers* or clusters $k \in \mathcal{K}$, such that \mathcal{K} can contain elements from \mathcal{I} or combinations of elements from \mathcal{I} , but it always holds that $|\mathcal{K}| \leq |\mathcal{I}|$. Let us consider in the following, that we want to merge data points from, e.g., two bases $i, j \in \mathcal{I}$ with centroids \bar{b}_i and \bar{b}_j . Then, the new centroid is given by $\bar{b}_{ij} = \frac{1}{W_i + W_j} (W_i \bar{b}_i + W_j \bar{b}_j)$. For example, when merging all the data points from bases 2 and 3 in Figure 2, the new centroid will fall into basis 2.⁵ Consequently, all data points of basis 3 will be treated as if they would fall into basis 2. This is illustrated in Figure 3. Therefore, the points initially in basis 3 will be assigned to a false set of active constraints, incurring an output error, which we discuss next.

B. Cost of misclassification

The aggregated LP associated with the merge of all points in bases i and j is defined as:

⁵Note that when merging data points from two bases not next to each other, e.g., 3 and 5, the new centroid may even fall outside these two bases, e.g., into basis 1. We will elaborate on this possibility in Section III-B.

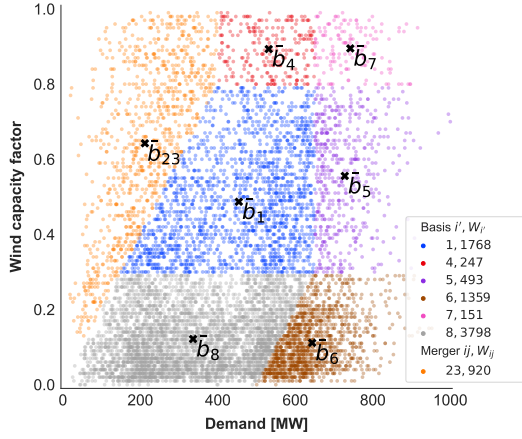


Fig. 3. Basis membership of time series tuples after merging (data points from bases 2 and 3).

$$\min_{\bar{x}_{i'}, \bar{x}_{ij}} \sum_{i' \in \mathcal{I}'} c^T \bar{x}_{i'} W_{i'} + c^T \bar{x}_{ij} W_{ij} \quad (7a)$$

$$\text{s. t. } A \bar{x}_{i'} \leq \bar{b}_{i'}, \quad \forall i' \in \mathcal{I}', \quad (7b)$$

$$A \bar{x}_{ij} \leq \bar{b}_{ij}, \quad (7c)$$

$$\bar{x}_{i'} \geq 0, \quad \forall i' \in \mathcal{I}', \quad (7d)$$

$$\bar{x}_{ij} \geq 0, \quad (7e)$$

where $\mathcal{I}' = \mathcal{I} \setminus \{i, j\}$, $\mathcal{K} = \mathcal{I}' \cup \{ij\}$, and $W_{ij} = W_i + W_j$ as the total weight of the new cluster, resulting from the merge of all data points from bases $i, j \in \mathcal{I}$. Since some data points in \mathcal{T}_{ij} are assigned to a false set of active constraints, the optimal objective function values of Problems (3) and (7) will not coincide. Hence, merging data points from bases i and j incurs an output error in the total system cost for which we derive an analytical formulation in the following. We will term this error *cost of misclassification* (CoM).

Let $OV^{\mathcal{I}}$ and $OV^{\mathcal{K}}$ denote the objective function values of Problems (3) and (7) at optimum, respectively. Then, we define the cost of misclassification associated with merging points from bases i and $j \in \mathcal{I}$, assuming the new centroid \bar{b}_{ij} falls into basis $h \in \mathcal{I}$, as:

$$\text{CoM}^{ij|h} = OV^{\mathcal{I}} - OV^{\mathcal{K}} = \bar{b}_i^T (\bar{y}_i^* - \frac{W_i}{W_h} \bar{y}_h^*) + \bar{b}_j^T (\bar{y}_j^* - \frac{W_j}{W_h} \bar{y}_h^*). \quad (8)$$

The derivation of the expression in Equation (8) can be found in Appendix B, with a generalized expression of the CoM.

We can calculate the centroid as shown, but it is not clear a priori in which basis the centroid will fall. In order to find out, we could, for example, check which set of active constraints is a feasible solution for the centroid. However, we have observed a favorable property of the Cost of Misclassification, which we want to mathematically prove in future research. When computing the Cost of Misclassification for all new possible centroids of a merge, say $\text{CoM}^{ij|i}$, $\text{CoM}^{ij|j}$ and $\text{CoM}^{ij|h}$ in our previous example, we have observed that it always holds that the new centroid of the merge coincides with the one incurring the least CoM, e.g., $\text{CoM}^{ij|h} = \min_{i' \in \mathcal{I}'} \text{CoM}^{ij|i'}$. Therefore, we conjecture that there is no need to bound the error when

we cannot know the location of the new merge, since we can know the exact error that the merge would incur.

This gives modelers control over the TSA procedure, as they can exactly quantify, without needing to solve any aggregated model – assuming the mapping of time steps to bases is known – the error that they are inducing in their models for any desired level of aggregation, as opposed to traditional TSA methods, that are unaware of the error that they introduce in the optimization model.

This issue also affects a-priori TSA methods. For example, when employing k-means, the obtained cluster centroids would also fall into one of the 8 bases that we have shown before. That means that all the hours that belong to that cluster, would be represented by one operational state of the power system. Just that, when doing a-priori TSA methods, one is usually ignorant of the error and misclassification one is committing.

IV. RESULTS

We now apply our new bases-merging methodology to the previously introduced transport problem (5) using the stylized case study in Figure 1. In Section IV-A, we identify optimal bases mergers among all potential ones. Afterward, in Section IV-B, we discuss the computational complexity of finding optimal bases mergers. We use the Pyomo package v6.5.0 [23], [24] for Python and the Gurobi Solver v11.0.0 [25] for our numerical analysis. The model implementation and case study data are provided in [26].

A. Exhaustive enumeration and optimal bases mergers

Available bases mergers can be mathematically interpreted as partitions of a set. A partition of a set \mathcal{S} is defined as a family of nonempty, pairwise disjoint subsets of \mathcal{S} whose union is \mathcal{S} .⁶ Generally, the total number of partitions of a set of size $|\mathcal{S}| = n$ is given by the Bell Number B_n , which is defined recursively as $B_{n+1} = \sum_{k=0}^n \binom{n}{k} B_k$ [27].

In the context of bases merging, we are looking for all partitions of the set of bases \mathcal{I} . In our numerical example of the transport problem, we have $|\mathcal{I}| = 8$ bases. We exhaustively enumerate all $B_8 = 4140$ possible bases mergers and solve the corresponding aggregated problems. Then, we can easily analyze how different aggregations of the input data, i.e., different bases mergers, affect the output of the aggregated model compared to the full model. Thereby, aiming to understand good or even optimal mergers from a structural point of view. We measure the quality of a merger in terms of the output error related to the objective function value ε_{OV} (with terms OV^i and OV^k defined as in Equation (15), in Appendix B) and decision variables $\varepsilon_g, \forall g \in \mathcal{G}$ as:

$$\varepsilon_{OV} = \frac{\sum_{i \in \mathcal{I}} OV^i - \sum_{k \in \mathcal{K}} OV^k}{\sum_{i \in \mathcal{I}} OV^i}, \quad (9)$$

$$\varepsilon_g = \frac{\sum_{i \in \mathcal{I}} P_{g,i} - \sum_{k \in \mathcal{K}} P_{g,k}}{\sum_{i \in \mathcal{I}} P_{g,i}}. \quad (10)$$

⁶For example, the set $\{1, 2, 3\}$ has these 5 partitions: $\{\{1\}, \{2\}, \{3\}\}$, $\{\{1, 2\}, \{3\}\}$, $\{\{1, 3\}, \{2\}\}$, $\{\{1\}, \{2, 3\}\}$ and $\{\{1, 2, 3\}\}$.

TABLE I
RESULTS OF OPTIMAL BASES MERGING. RELATIVE ERRORS ε ARE SHOWN IN %. VALUES ARE ROUNDED TO TWO DECIMALS.

$ \mathcal{K} $	Optimal bases mergers	ε_{OV}	ε_{Re}	ε_{Th}	ε_{NSP}
8	{1},{2},{3},{4},{5},{6},{7},{8}	0.00	0.00	0.00	0.00
7	{1},{2,3},{4},{5},{6},{7},{8}	0.00	0.00	0.00	0.00
6	{1,8},{2,3},{4},{5},{6},{7}	0.00	0.00	0.00	0.00
5	{1,8},{2,3},{4},{5,7},{6}	0.02	-0.59	0.31	0.00
4	{1,4,8},{2,3},{5,7},{6}	0.05	-1.52	0.80	0.00
3	{1,2,3,4,8},{5,7},{6}	0.28	-9.74	5.08	0.00
2	{1,2,3,4,5,7,8},{6}	28.53	-9.74	2.86	30.39
1	{1,2,3,4,5,6,7,8}	93.24	-9.74	-2.22	100

The optimal bases mergers and the associated output errors⁷ with respect to the full hourly model are shown in Table I. We observe when reducing the number of clusters, e.g., from 8 to 7, it is optimal to merge data points from bases 2 and 3. In bases 2 and 3, the wind generator is the marginal unit. But they differ in the congestion of line 3, resulting in a difference in their locational marginal prices of 0.1 \$/MWh due to the transport cost C_i^{TR} (see Table II in Appendix A). As the transport cost is comparably small, the error in objective function value induced by the aggregation is $\varepsilon_{OV} \approx 0$. Notably, in our example, it is possible to reduce the number of clusters down to four, without incurring an absolute relative error of above 0.05% in the optimal objective function value and 2% in the decision variable values. This shows the potential of the proposed bases merging framework to choose input data aggregations that reduce model complexity without losing a lot of accuracy. While this accuracy can potentially also be achieved with a-priori methods, there is no way of controlling the output error during the aggregation.

Note that it is not needed for quantifying the error, to actually re-solve the aggregated optimization problems after merging bases, as the CoM (15) given in Appendix B, which can be computed as shown in Section III-B, are equal to ε_{OV} in absolute terms. However, when the number of bases \mathcal{I} is large this may still be computationally challenging. We tackle this issue in the next Section.

B. Computational complexity of finding optimal bases mergers

The Bell Number is growing faster than exponentially with the size of the set, in our case the bases \mathcal{I} , to be partitioned. While for $n = 8$ there are 4140 partitions, for $n = 15$ it would already be 1.3 Billion. In our numerical example, however, we observed that the optimal bases mergers share some common properties drastically reducing the number of partitions to consider. We explain both properties in the following. Due to the scope of this paper, we must postpone mathematical proofs for their validity to future research.

⁷The mergers in Table I are optimal in terms of the error in the objective function value ε_{OV} . Depending on the model's purpose, the metric of interest may be a different variable. For example, if the modeler is a wind power investor, the metric of interest may be the error in wind production ε_{Re} , while for a grid operator, it might be error in non-supplied power ε_{NSP} . In summary, the modeler's point of view determines the optimal mergers.

a) *Greediness*: Looking at Table I, it can be noted that for any number of bases in the merged set, the optimal bases mergers follow a hierarchical pattern. For example, merging data points from bases 2 and 3 is not only optimal when reducing the number of clusters by 1, but also for all further reductions. We therefore hypothesize that a greedy strategy will generally deliver good, if not optimal, bases mergers for all types of LPs.

b) *Adjacency*: We have further observed that all optimal bases mergers shown in Table I include bases that are *adjacent*, meaning they share only one edge in the input space shown in Figure 2. For instance, bases 2 and 3 are adjacent as opposed to bases 2 and 6. We hypothesize that adjacency in input space relates to the adjacency of basic solutions in the Simplex method. This may be the foundation for proving it as a necessary condition for optimal bases mergers.

Table III in Appendix C shows the number of potential optimal bases mergers based on the exhaustive enumeration baseline, a greedy strategy, and the latter with an adjacency requirement.

Our results suggest both properties, greediness and adjacency, significantly reduce potential optimal bases mergers compared to the Bell Number, possibly overcoming its faster-than-exponential growth and achieving polynomial complexity. In a non-degenerate LP with m constraints and n variables, the number of basic feasible solutions is at most $\binom{n}{m}$, which grows as $\mathcal{O}(2^n)$ in the worst case. On the other hand, each basic feasible solution has at most $m(n-m)$ adjacent basic feasible solutions, assuming there is no degeneracy and all pivots lead to feasible solutions. This grows as $\mathcal{O}(n^2)$ in the worst case. Therefore, we can expect greater reductions in larger problems as the difference of both terms grows exponentially.

V. CONCLUSIONS

Recently, it was proposed in [18] to use TSA based on active constraint sets (or bases). This aggregation becomes inefficient when the number of active constraint sets becomes large. We extend this framework by introducing the concept of merging data belonging to different active constraint sets, which we term bases merging. While this reduces model size it incurs an output error. However, we demonstrate that optimal mergers may be systematically identified and derive an exact analytical formulation to compute the corresponding output error that does not require re-solving the optimization model. This enables modelers to balance model accuracy and computational complexity efficiently.

Future research should derive mathematical proofs or guarantees for choosing optimal bases mergers using a greedy strategy or requiring adjacency. Furthermore, the proposed bases merging method should be extended to account for time-linking constraints. Finally, it should be investigated how the assignment of data points to bases can be derived or approximated without solving the full model.

ACKNOWLEDGMENT

Funded by the European Union (ERC, NetZero-Opt, 101116212). Views and opinions expressed are however those

of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

- [1] E. Ridha, L. Nolting, and A. Praktijnjo, “Complexity profiles: A large-scale review of energy system models in terms of complexity,” *Energy Strategy Rev.*, vol. 30, p. 100 515, 2020.
- [2] B. Kroposki, “Integrating high levels of variable renewable energy into electric power systems,” *J. Mod. Power Syst. Clean Energy*, vol. 5, no. 6, pp. 831–837, 2017.
- [3] T. Sousa, T. Soares, P. Pinson, *et al.*, “Peer-to-peer and community-based markets: A comprehensive review,” *Renew. Sustain. Energy Rev.*, vol. 104, pp. 367–378, 2019.
- [4] D. Ilic, P. G. Da Silva, S. Karnouskos, *et al.*, “An energy market for trading electricity in smart grid neighbourhoods,” in *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, 2012, pp. 1–6.
- [5] A. Mohammadi and M. Mehrpooya, “A comprehensive review on coupling different types of electrolyzer to renewable energy sources,” *Energy*, vol. 158, pp. 632–655, 2018.
- [6] M. Robinius, T. Raje, S. Nykamp, *et al.*, “Power-to-Gas: Electrolyzers as an alternative to network expansion – An example from a distribution system operator,” *Appl. Energy*, vol. 210, pp. 182–197, 2018.
- [7] A. S. Gaur, D. Z. Fitiwi, and J. Curtis, “Heat pumps and our low-carbon future: A comprehensive review,” *Energy Res. Soc. Sci.*, vol. 71, p. 101 764, 2021.
- [8] S. Abd Alla, V. Bianco, F. Scarpa, *et al.*, “Electrification of the residential heat demand: An analysis of the power market potential to accommodate heat pumps,” *Therm. Sci. Eng. Prog.*, vol. 27, p. 101 173, 2022.
- [9] M. Bröchin, B. Pickering, T. Tröndle, *et al.*, “Harder, better, faster, stronger: Understanding and improving the tractability of large energy system models,” in *Energy Sustain. Soc.*, vol. 14, no. 1, p. 27, 2024.
- [10] L. Kotzur, L. Nolting, M. Hoffmann, *et al.*, “A modeler’s guide to handle complexity in energy systems optimization,” *Adv. Appl. Energy*, vol. 4, p. 100 063, 2021.
- [11] M. Hoffmann, L. Kotzur, D. Stolten, *et al.*, “A Review on Time Series Aggregation Methods for Energy System Models,” in *Energies*, vol. 13, no. 3, p. 641, 2020.
- [12] S. Lloyd, “Least Squares Quantization in PCM,” *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [13] S. Pfenninger, “Dealing with multiple decades of hourly wind and PV time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability,” *Appl. Energy*, vol. 197, pp. 1–13, 2017.
- [14] L. Kotzur, P. Markewitz, M. Robinius, *et al.*, “Impact of different time series aggregation methods on optimal energy system design,” *Renew. Energy*, vol. 117, pp. 474–487, 2018.
- [15] K. Poncelet, H. Hoshle, E. Delarue, *et al.*, “Selecting Representative Days for Capturing the Implications of Integrating Intermittent Renewables in Generation Expansion Planning Problems,” in *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 1936–1948, 2017.
- [16] G. Mavrotas, D. Diakoulaki, K. Florios, *et al.*, “A mathematical programming framework for energy planning in services’ sector buildings under uncertainty in load demand: The case of a hospital in Athens,” *Energy Policy*, vol. 36, no. 7, pp. 2415–2429, 2008.
- [17] R. Green, I. Staffell, and N. Vasilakos, “Divide and Conquer? k-Means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System,” *IEEE Trans. Eng. Manag.*, vol. 61, no. 2, pp. 251–260, 2014.
- [18] S. Wogrin, “Time Series Aggregation for Optimization: One-Size-Fits-All?” *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 2489–2492, 2023.
- [19] B. Bahl, A. Kümpel, H. Seele, *et al.*, “Time-series aggregation for synthesis problems by bounding error in the objective function,” *Energy*, vol. 135, pp. 900–912, 2017.
- [20] M. Sun, F. Teng, X. Zhang, *et al.*, “Data-Driven Representative Day Selection for Investment Decisions: A Cost-Oriented Approach,” *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 2925–2936, 2019.
- [21] A. P. Hilbers, D. J. Brayshaw, and A. Gandy, “Reducing climate risk in energy system planning: A posteriori time series aggregation for models with storage,” *Appl. Energy*, vol. 334, p. 120 624, 2023.
- [22] D. Cardona-Vasquez, T. Klatzer, B. Klinz, *et al.*, “Enhancing time series aggregation for power system optimization models: Incorporating network and ramping constraints,” *Electr. Power Syst. Res.*, vol. 230, p. 110 267, 2024.
- [23] W. E. Hart, J.-P. Watson, and D. L. Woodruff, “Pyomo: Modeling and solving mathematical programs in python,” *Mathematical Programming Computation*, vol. 3, no. 3, pp. 219–260, 2011.
- [24] M. L. Bynum, G. A. Hackebeil, W. E. Hart, *et al.*, *Pyomo—optimization modeling in python*, Third. Springer Science & Business Media, 2021, vol. 67.
- [25] Gurobi Optimization, LLC, *Gurobi Optimizer Reference Manual*, 2024. [Online]. Available: <https://www.gurobi.com>.
- [26] *TSA: Basis Merging Methodology*, https://anonymous.4open.science/t/TSA_BasisMerging-B054, 2025.
- [27] J. Pitman, “Some Probabilistic Aspects of Set Partitions,” *The American Mathematical Monthly*, vol. 104, no. 3, pp. 201–209, 1997.

- [28] K. N. B. and, “Close encounters with the stirling numbers of the second kind,” *Mathematics Magazine*, vol. 85, no. 4, pp. 252–266, 2012.

APPENDIX A

BASES OF THE OPTIMAL TRANSPORT PROBLEM

In Table II, we intuitively describe the eight bases of the optimal transport problem applied to the case study in Section II-B. For every basis $i \in \mathcal{I}$, we present its color used in Figure 2, the number of data points that share this basis (weight), which network lines are congested, the generators operating at full load, and, lastly, the marginal generator (i.e., most expensive generator needed to balance supply and demand) and locational marginal price (LMP) at Node 1, where all the demand is.

TABLE II
DESCRIPTION OF SET OF BASES \mathcal{I} FROM TRANSPORT PROBLEM.

Basis i	Color	$ \mathcal{T}_i $	Congestion	Full load	Marginal gen.	LMP
1	Blue	1768	Line 3	<i>Re</i>	<i>Th</i>	24.1
2	Orange	669	Line 3	-	<i>Re</i>	3.2
3	Green	251	-	-	<i>Re</i>	3.1
4	Red	247	Lines 2, 3	-	<i>Th</i>	24.1
5	Purple	493	Lines 1, 3	<i>Re</i>	<i>NSP</i>	5000
6	Brown	1359	Line 1	<i>Re, Th</i>	<i>NSP</i>	5000
7	Pink	151	Lines 1, 2, 3	-	<i>NSP</i>	5000
8	Gray	3798	-	<i>Re</i>	<i>Th</i>	24.1

APPENDIX B

COST OF MISCLASSIFICATION

Let $OV^{ij|h}$ in Equation (11) denote the part of the dual objective function related to Constraint (7c) of Problem (7) at optimum. Since we are assuming that the centroid b_{ij} falls into basis $h \in \mathcal{I}$, the associated dual variables $\bar{y}_{ij|h}^*$ correspond to \bar{y}_h^* , associated to basis h , re-weighted according to the weight W_{ij} of cluster ij .

$$OV^{ij|h} = \bar{b}_{ij}^\top \bar{y}_{ij|h}^* = \quad (11a)$$

$$= \left(\frac{W_i}{W_i + W_j} \bar{b}_i^\top + \frac{W_j}{W_i + W_j} \bar{b}_j^\top \right) (W_i + W_j) \frac{\bar{y}_h^*}{W_h} = \quad (11b)$$

$$= \left(\frac{W_i}{W_h} \bar{b}_i^\top + \frac{W_j}{W_h} \bar{b}_j^\top \right) \bar{y}_h^*. \quad (11c)$$

Further, define $OV^{i'} = c^\top \bar{x}_{i'}^* W_{i'} = \bar{b}_{i'}^\top \bar{y}_{i'}^*$ as the part of the primal and dual objective functions (3a) and (4a), respectively, attributed to basis $\forall i' \in \mathcal{I}$. We derive the expression of the CoM from Equation (8) as:

$$CoM^{ij|h} = OV^{\mathcal{I}} - OV^K = \quad (12a)$$

$$= OV^i + OV^j - OV^{ij|h} = \quad (12b)$$

$$= \bar{b}_i^\top \bar{y}_i^* + \bar{b}_j^\top \bar{y}_j^* - \left(\frac{W_i}{W_h} \bar{b}_i^\top + \frac{W_j}{W_h} \bar{b}_j^\top \right) \bar{y}_h^* = \quad (12c)$$

$$= \bar{b}_i^\top \left(\bar{y}_i^* - \frac{W_i}{W_h} \bar{y}_h^* \right) + \bar{b}_j^\top \left(\bar{y}_j^* - \frac{W_j}{W_h} \bar{y}_h^* \right). \quad (12d)$$

In order to define a generalized expression of the CoM, let us consider the primal and dual LPs associated to the aggregation of data points with clusters $k \in \mathcal{K}$ as:

TABLE III

NUMBER OF POTENTIAL MERGERS FOR DIFFERENT STRATEGIES AS A FUNCTION OF THE CARDINALITY OF \mathcal{K} .

	$ \mathcal{K} $							
	8	7	6	5	4	3	2	1
Exhaustive	1	28	266	1050	1701	966	127	1
Greedy (G.)	1	28	21	15	10	6	3	1
G. & Adjacent	1	11	10	8	7	4	3	1

$$\min_{\bar{x}_k} \sum_{k \in \mathcal{K}} c^\top \bar{x}_k W_k \quad (13a) \quad \max_{\bar{y}_k} \sum_{k \in \mathcal{K}} \bar{b}_k^\top \bar{y}_k \quad (14a)$$

$$\text{s. t. } A \bar{x}_k \leq \bar{b}_k, \quad \forall k, \quad (13b) \quad \text{s. t. } A^\top \bar{y}_k \geq c, \quad \forall k, \quad (14b)$$

$$\bar{x}_k \geq 0, \quad \forall k, \quad (13c) \quad \bar{y}_k \geq 0, \quad \forall k. \quad (14c)$$

Note that, in the case of perfect aggregation, $\mathcal{K} = \{\{i\} | i \in \mathcal{I}\}$, where every cluster contains points from only one basis. In the case from LP (7), $\mathcal{K} = \{\{ij\}\} \cup \{\{i'\} | i' \in \mathcal{I} \setminus \{i, j\}\}$, thus points from all bases except for i and j are represented by their original basis' centroid. It can be noted that LP (3) and LP (13), as well as LP (4) and LP (14), are equivalent from an optimization point of view. They differ in the way the input data centroids \bar{b}_i and \bar{b}_k are computed.

Now, let $OV^{\mathcal{I}}$ and OV^K denote the objective function values of Problems (3) and (13) at optimum, respectively. Define $OV^i = c^\top \bar{x}_i^* W_i = \bar{b}_i^\top \bar{y}_i^*$ as the part of the primal and dual objective functions (3a) and (4a), respectively, attributed to basis $\forall i \in \mathcal{I}$. Additionally, define $OV^k = c^\top \bar{x}_k^* W_k = \bar{b}_k^\top \bar{y}_k^*$ as the part of the primal and dual objective functions (13a) and (14a), respectively, attributed to cluster $\forall k \in \mathcal{K}$. Then, let us define the CoM for a general case where we merge bases from $i \in \mathcal{I}$ into clusters $k \in \mathcal{K}$ as the difference between $OV^{\mathcal{I}}$ and OV^K :

$$CoM = OV^{\mathcal{I}} - OV^K = \quad (15a)$$

$$= \sum_{i \in \mathcal{I}} OV^i - \sum_{k \in \mathcal{K}} OV^k = \quad (15b)$$

$$= \sum_{i \in \mathcal{I}} \bar{b}_i^\top \bar{y}_i^* - \sum_{k \in \mathcal{K}} \bar{b}_k^\top \bar{y}_k^*. \quad (15c)$$

APPENDIX C

BASES MERGING STRATEGIES

In Table III, we observe the number of potential bases mergers that are evaluated by three merging strategies (Exhaustive, Greedy, Greedy & Adjacent) as a function of the cardinality of the set of clusters \mathcal{K} . Note that i, i' denote bases $\in \mathcal{I}$, and $k, k' \in \mathcal{K}$ denote clusters. A cluster can contain only one basis or more than one merged bases.

All three strategies initialize \mathcal{K} by assigning each of the eight bases from \mathcal{I} into its own cluster. Then, they iteratively reduce the cardinality of \mathcal{K} by 1, as they attempt to find at each iteration the bases merger that minimizes a given metric, i.e., the CoM in our case. For the case study, all three strategies found the same optimal bases mergers presented in Table I.

The Exhaustive strategy is presented in Algorithm 1. At each iteration $iter$, the number of bases mergers evaluated by the strategy is given by the Stirling number of the second kind $S_{|\mathcal{I}|, iter}$, which describe the number of partitions of size $iter$ of the set of bases \mathcal{I} , as $S_{|\mathcal{I}|, iter} = \frac{1}{iter!} \sum_{n=0}^{iter} \binom{iter}{n} (-1)^{iter-n} n^{|\mathcal{I}|}$ [28].

In this strategy, the iterations can be interpreted as independent from each other, in the sense that the merger chosen at a given iteration does not affect the merger to be chosen in the subsequent ones, as all potential mergers are evaluated.

Since a priori, we do not know what is the optimal number of clusters for a particular problem, we try out all options as calculated in Algorithm 1, which yields the Bell number, e.g. $B_8 = \sum_{iter=1}^8 S_{8, iter}$ in our case.

This strategy serves as a baseline to the Greedy and Greedy & Adjacent strategies, as it provides an upper bound of the number of potential bases mergers.

Algorithm 1 Exhaustive strategy

```

for  $iter = |\mathcal{I}|, |\mathcal{I}| - 1, \dots, 1$  do
   $mergers \leftarrow partitions(\mathcal{I}, size = iter)$ 
  Choose merger  $\in mergers$  with min. CoM
end for

```

The Greedy strategy, presented in Algorithm 2, can be seen as an improvement with respect to the Exhaustive strategy. In this strategy, we keep a memory of the optimal merger found in the previous step. Therefore, the number of bases mergers to evaluate at iteration $iter$ depends on the decision previously taken at $iter + 1$.

At each iteration, Algorithm 2 evaluates all possible mergers given by the pairwise combinations of the elements in \mathcal{K} . The merger that minimizes the CoM, we refer to it as $\{kk'\}$, is chosen and \mathcal{K} is updated with this choice. This means that the original separate clusters $\{k\}$ and $\{k'\}$ are removed from \mathcal{K} and replace by the merged cluster $\{kk'\}$. Once two clusters have been merged, this merge cannot be undone in subsequent iterations.

The number of bases mergers evaluated by this strategy at iteration $iter$ is equal to $\binom{iter+1}{2}$ if $iter < |\mathcal{I}|$ else 1.

Algorithm 2 Greedy strategy

```

 $\mathcal{K} \leftarrow \{\{i\} | i \in \mathcal{I}\}$ 
for  $iter = |\mathcal{I}|, |\mathcal{I}| - 1, \dots, 1$  do
   $mergers \leftarrow pairs(\mathcal{K})$ 
  Choose merger  $\{kk'\} \in mergers$  with min. CoM
  Update  $\mathcal{K} \leftarrow \mathcal{K} \setminus \{\{k\}, \{k'\}\} \cup \{kk'\}$ 
end for

```

The Greedy & Adjacent strategy, described in Algorithm 3, can be understood as an improvement with respect to the Greedy strategy. Here, on top of the memory of optimal mergers from previous iterations, we only evaluate mergers that satisfy an adjacency condition. In the case study, we have identified the following pairs of bases as adjacent: $adj = [(1,$

2), (1, 4), (1, 5), (1, 8), (2, 3), (2, 4), (3, 8), (4, 7), (5, 6), (5, 7), (6, 8)] from Figure 2.

For a pair of clusters $k, k' \in \mathcal{K}$ to be considered adjacent:

- If each of the two clusters coincides with a basis (i.e. $k = i$ and $k' = i'$), they are adjacent if they are one of the identified pairs of adjacent bases in adj .
- If any of the two clusters k, k' that are considered for a merge, contain more than one basis (e.g. $k = \{ii'\}$), they are adjacent if the resulting merged cluster can be obtained from the union of pairs of adjacent bases. As an example, clusters $\{2, 3\}$ and $\{4\}$ are adjacent since bases 2 and 3, as well as 3 and 4, are adjacent. As opposed to, for example, clusters $\{2, 3\}$ and $\{6\}$, which do not satisfy the adjacency condition since basis 6 is neither adjacent to bases 2 or 3.

Since adjacency is problem-dependent, it is not trivial to derive a closed-form expression of the number of mergers that are evaluated by this strategy. However, the Greedy strategy serves as an upper bound to it.

Algorithm 3 Greedy & Adjacent strategy

```

 $\mathcal{K} \leftarrow \{\{i\} | i \in \mathcal{I}\}$ 
 $adj \leftarrow$  List of adjacent bases
for  $iter = |\mathcal{I}|, |\mathcal{I}| - 1, \dots, 1$  do
   $mergers \leftarrow pairs(\mathcal{K})$ 
   $adjmergers \leftarrow filter(mergers, adj)$ 
  Choose merger  $\{kk'\} \in adjmergers$  with min. CoM
  Update  $\mathcal{K} \leftarrow \mathcal{K} \setminus \{\{k\}, \{k'\}\} \cup \{kk'\}$ 
end for

```
