Skoltech
Skolkovo Institute of Science and Technology

# Thesis Changes Log

**Name of Candidate:** Mile Mitrovic

**PhD Program:** Engineering Systems

**Title of Thesis:** Data-Driven Stochastic AC-OPF using Gaussian Processes

**Supervisor:** Assistant Professor Elena Gryazina

**Co-supervisor:** Assistant Professor Petr Vorobev

*The thesis document includes the following changes in answer to the external review process.*

I am grateful to the jury members for their positive feedback and useful comments. I am happy to address the comments and questions in this document and in the revised version of the thesis.

**Prof. Henni Ouerdane**

1. Conclusion is quite short and rather looks like an abstract or a simple recap. In fact, while some limitations and challenges with the methods routinely used for optimal power flow calculations are mentioned in the Introduction, I see no research question(s) properly formulated there.
   I gratefully appreciate this comment. I completely agree with you that the General Introduction lacks properly formulated research questions. I added it in the General Introduction of the thesis. Accordingly, I expanded the conclusion. I would like to mention that each chapter has its own introduction and conclusion, so in the conclusion I did not go into detail to summarize each chapter.

**Prof. Andrei Osiptsov**

1. On p 73 RMSE as a metrics is proposed. What is the rationale behind this choice?
   I appreciate your pointing this out. We are considering a regression problem. The idea was to evaluate the prediction performance of the GP model. Accordingly, we use RMSE to perform it since we can interpret the error metric in the same units as the target variable. Thus, we express the average error on a more understandable scale.

2. User guide in Chapter 6 is quite brief yet clear for the reader.
   Here I just wanted to give general guidelines about the code how to use the code. In detail about the code, the user can find it in the code itself, as well as everything about what each function is for.

3. Conclusions could have been a bit more detailed to explain the impact of this work.
   I gratefully appreciate this comment. I fixed it in the thesis. Also, I would like to mention that each chapter has its own conclusion, so I did not repeat the conclusions per chapter,

but I defined the research questions more precisely in the introduction and expanded the conclusion based on that.

**Prof. Federico Martin Ibanez**

1. When the power is introduced, it is not clear if it is talking about the instantaneous power i.v or the average power V.I* if so, I guess the conjugate is missing.
   Thank you for pointing this out. You are right, * is missing. This is a complex power, and I is a complex conjugate. It is fixed in the thesis.

2. As the author mention there are many ways of improving the OPF, why the author is only focused in machine learning Gaussian Processes, it will be good to compare with other type of ML.
   I gratefully appreciate this comment. Comparing the Gaussian Process model with other ML models makes sense for the deterministic OPF problem. In general, all other ML models including Deep Learning models are deterministic. They can be implemented in a stochastic OPF way, but to realize the control action, we will have to iterate through the samples. One of the main advantages of Gaussian Processes is that we analytically reformulate the stochastic OPF problem and solve it in one sample iteration for given input random variables.

3. There are typos when you copy text from your own articles, like "this paper..." and some missing references in p. 67-68.
   Thank you for noticing these typos. I passed through the thesis and fixed it.

4. The technique is much faster for than for CC-OPF. My question is, which is the type of OPF used now in the real applications. What will be the benefits of using yours? Is it interesting to have the OPF once per minute?
   In the real application, they still use deterministic OPF. I do not have a lot of information about using stochastic OPF in real industry, but from my research, the industry is not yet prepared for stochastic models. Using our model can benefit from avoiding choosing scenarios (that made the industry suspicious in applying stochastic models since different scenarios can lead to different dispatches) and analytically solving the problem similarly to a deterministic problem with accounting for uncertainty. Accordingly, they will get a more robust solution, with less computational resources, as well as with providing satisfying security against uncertain injections.

   In general, OPF is static control also known as tertiary control. Depending on the system they are realized from 15min to several hours. One of the problems with having every 1 minute can be difficulty dispatching big generators in e.g. thermal power station. In general primary and secondary control, are much faster and they are done in several minutes.

5. Is it possible to run a IEEE-9 in the lab? What would be the benefits?
   With a small adaptation of the application, it could easily be released in the lab. In lab conditions, I can not say what the actual benefit will be other than just running the app and getting the faster solution with less computation resources for stochastic problems, but in real-life conditions, you should get a trade-off between security and cost, and thus provide enough secure system with optimal costs that can save billions of dollars avoiding choosing which scenarios to use.

**Prof. Haoran Zhao**

1. Solver: The thesis mentions the utilization of IPOPT, an NLP solver, for solving both the full GP CC-OPF and hybrid GP CC-OPF. However, it is important to acknowledge that IPOPT might not always provide a globally optimal solution. Therefore, it would be advantageous to explain the steps taken to mitigate any potential suboptimality resulting from this choice of solver.

   I gratefully appreciate this comment. I completely agree that IPOPT might not always provide a globally optimal solution, which can be attributed to the considerable non-convexity of the AC-PF function. Because of that one of the advantages of the GP approach that we proposed is some kind of relaxation where we approximate real function making the problem more tractable than the AC-PF equations by avoiding some local suboptimal. This approximation simplifies the optimization problem significantly and it is discussed in Chapter 4.3.1 and illustrated in Figure 4-1.

2. Case Study: The proposed CC-OPF methods are assessed using the IEEE9, IEEE39, and IEEE118 bus systems, which serve as standard benchmarks. However, it is worth noting that real-world OPF scenarios often involve systems with hundreds or even thousands of nodes. Therefore, providing insights into the performance of these methods when applied to larger and more realistic benchmark systems would be highly valuable.

   I appreciate your pointing this out. These statements are mainly covered in subsections 4.4.6 and 5.5 where large-scale problems and their solutions are discussed. One of the main problems of GP is scalability, but with appropriate approximation techniques and tricks, some of which we have already implemented, the GP approach can be efficiently applied to large-scale power systems.

3. Results Comparison: The manuscript compares the data-driven GP CC-OPF approach with state-of-the-art sample-based CC-OPF methods. However, to provide a more comprehensive analysis, it would be beneficial to compare the results with non-sample-based CC-OPF methodologies as well.

   I thank the reviewer for this comment. I thank the reviewer for this comment. Actually, the proposed GP approach can be considered a non-sample-based CC-OPF since chance constraints are analytically reformulated without a number of scenarios. Also, we consider so-called case B where we consider different worst cases for the given uncertant input. Yes, we sample through Monte Carlo samples to see results for the different worst cases with different std, but in general, any worst-case sample can be considered as non-sample-based CC-OPF.

4. Model Generalization: The primary idea of the thesis is to replace the standard AC-OPF equation and security constraints with a GP regression model. However, concerns arise regarding the generalization capability of this model. It is important to investigate how the model performs when exposed to data that deviates from the assumed distribution. Are there any provisions or assessments in place to ensure robustness under such conditions? By addressing these concerns and evaluating the model's performance under different data scenarios, the thesis can provide a comprehensive understanding of the model's capabilities and limitations, ultimately enhancing its applicability and usefulness in real-world situations

   One of the reasons for the developing robust hybrid GP approach is the problem of the data deviation from the assumed distribution. This problem is discussed in subsection 5.4.6. Also, the models are evaluated on different distribution gaps in Table 5.3.

**Prof. Ashok Kumar Pradhan**

1. Page 17, last paragraph- here issues in today's power systems related to renewables integration are mentioned. It may be good to add issues related to inertia, frequency and voltage of the system.
   I appreciate your pointing this out. The primary focus of this thesis research was to develop a novel approach for addressing stochastic OPF problems, particularly those arising from the integration of renewable energy sources. In modern power grids, the integration of renewables introduces significant uncertainties, which can manifest in various ways, including challenges related to inertia, frequency regulation, and voltage stability. Therefore, we wanted to focus only on the problems in today's power systems related to the integration of renewable energy sources that affect the efficient solution of the stochastic OPF problem.

2. Page 18, paragraph-1, [Ullah 2022], here it would be better to state on- 'What is chance-constrained (CC)? can not should be a single word.
   I gratefully appreciate this comment. I completely agree that the chance-constrained approach should be explained in more detail in this paragraph. I added an additional explanation of the chance-constrained approach on Page 18, Paragraph-1.

3. Paragraph-2, line2, what is Alternating Current here? You may provide a statement here.
   I completely agree with you. An additional statement about alternating current has been added on Page 18, Paragraph-2.

4. Page 25, last line, low should be law.
   I thank you for noticing. It is in the thesis.

5. Page 27, equation (2.6). We know S=VI*. Is the equation ok?
   I appreciate you pointing this out. You are right, the current phasor is a complex conjugate and * is missed to describe it. I fixed it in the thesis.

6. Page 41, last line, does it mean the load power factor is fixed?
   Yes, the power factor is constant, which meens that the ratio of active and reactive power injections at each node remains unchanged during fluctuations. The common approach, that RES reactive power generation changes following the deviation of the active power output, is based on the fact that different grid operators have different requirements on the reactive power control from renewable generators (more in [1] and [2]) . However, the different types of control can be included in the reactive power formulation without any conceptual changes in the method and without affecting the data-driven approach developed in this paper

   [1] L. Roald, G. Andersson, Chance-constrained AC optimal power flow: Reformulations and efficient algorithms, IEEE Transactions on Power Systems 33 (3) (2017) 2906–2918.
   [2] M. Lubin, Y. Dvorkin, L. Roald, Chance constraints for improving the security of ac optimal power flow, IEEE Transactions on Power Systems 34 (3) (2019) 1908–1917.

7. Page 42, the last statement, perhaps it outlies also- active and reactive powers at renewables, load and other form of generations.
   You are right, but with the emphasis that the chapter outlines a process for generating the synthetic dataset we mean the sampling of active and reactive powers at renewables, load, and other forms of generations.

8. Page 51, 3.3.2, line-3, training data (not tada).
   Thank you for noticing this typo. It is fixed in the thesis.

9. Page 73, In Chapter-5 number of training cases are mentioned. It would be better to include that here also.
   The number of training cases is mentioned in Chapter 4 too. They are mentioned on Page 74, in subsection: 4.4.3 Models Performance

10. Page 79, Fig 4.5 is for IEEE bus and Fig 4.6 is for 39 bus. In both the results, frequency is scaled to around 200. What does it imply? What are the number of simulation cases for different systems?
    I will try to explain the idea of these figures in another way. We wanted to graphically represent the predicted output values as a random variable and how they behave near certain constraints. For example, in our proposed GP approach the output voltage and current are random variables represented with mean and variance. For this mean and variance, we take a random sample with a specific number of samples to create a distribution. For other approaches and output values from different IEEE systems, we take the same specific number of samples to create the distribution as in these figures. This is why we have a similar frequency for different IEEE systems. Note that this is not the number of training data, but only the number of samples to visually create random outputs.

11. Page 81, What is the general guideline for training data preparation? What coverage it should have in the operating space of a power system?
    At best, we should have data from the real system. Due to the lack of data from real power systems, we had to create synthetic data to work on this problem. General guidelines for preparing training data are presented in Chapter 2, subsection 2.4 Synthetic dataset generation. This guideline follows [1], where the authors work with real data sets and try to derive an appropriate approach for creating synthetic data sets. In terms of operating space coverage, in data science, it is better to have data available for all possible operating spaces. Of course, sometimes it is impossible to have it, especially in high-risk systems like power systems. One of the reasons to propose a hybrid GP-CCOPF approach is related to the data space coverage problem.

    [1] B. Donnot, Deep learning methods for predicting flows in power grids: novel architectures and algorithms, Ph.D. thesis, Universite Paris Saclay (CO-mUE) (2019).

**Prof. Alexander Nazin**

1. How big is the typical range of uncertainties in the CC-OPF problem?
   I am not sure that I correctly understand this question, but I assume that this question points to the acceptable violation probabilities. If we talk about it, then it is very changeable depending on the variables that we constrain. For example, for variables produced from generators that technically can exceed some limit we need to constrain these variations with so small probabilities. However other variables such as voltages and currents can be constrained with more violation probability. There is no exact definition of it, but in our cases, we consider a 2.5% violation to provide a more secure system.

2. How well is your assumption about log-normal distribution for power fluctuations justified?

I appreciate your pointing this out. In general, we should work with real-life data. Due to the lack of these data, we were forced to generate synthetic data. The idea was to simulate a similar time series of real input data and thus obtain corresponding output data. To do this we follow the work [1] where the authors try to create synthetic data following real data. Consequently, they proposed a log-normal distribution to simulate real input variables. Therefore, we follow this assumption to generate synthetic data.

[1] B. Donnot, Deep learning methods for predicting flows in power grids: novel architectures and algorithms, Ph.D. thesis, Universite Paris Saclay (CO-mUE) (2019).

3. Does your method suffer from the "curse of dimensionality" with the increase in the grid size?
   I gratefully appreciate this question. In general, the GP model suffers from dimensionality. Consequently, our first approach, the so-called full GP-CCOPF, suffers from dimensionality. To reduce this impact, we proposed a so-called hybrid GP-CCOPF approach where we use some techniques such as sparse approximation. The results showed that this approximation greatly reduces the impact of dimensionality without greatly affecting the level of accuracy.

4. Are there instances when scenario-based approach is better than your method?
   One of the problems with the scenario approach is the use of a large number of scenarios that still cannot guarantee an accurate solution. More scenarios increase the probability of finding the optimal solution but notoriously increase computational resources. In the case that we can be better in accuracy and optimality with a scenario-based approach, we will need huge computational resources. The advantage of our approach is to find this trade-off.

5. How would you find the level of uncertainty for real-life power grids?
   The level of uncertainty in power networks is not easy to define, because there can be many factors of uncertainty. But if we only focus on renewable energy sources as the main factor of uncertainty in power grids, as in this thesis, we can define the level of uncertainty as the ratio of energy produced from renewable sources to the total energy produced. Of course, that definition will be a rough estimation.