Report of the ANR JCJC project "GAussian Processes for computer experiments and machine learning: more guarantees and broader applications" (GAP), 2021-2026

ANR-21-CE40-0007

Project coordinator: François Bachoc
(Université de Lille and Institut Universitaire de France (IUF)[1])

1 Quick summary of the project

- **Duration:** October 2021 to April 2026.
- **Budget:** 205 000 euros.
- **Location:** Université de Toulouse with collaborations with Sorbonne Université and Imperial College London.
- **Members:** François Bachoc (PI), Ismaël Castillo, Mathis Deronzier (PhD student), Nicolas Durrande, Agnès Lagnoux, Andrés Felipe Lopez Lopera, Pierre Neuvial, Thibault Randrianarisoa (PhD student when project started), Olivier Roustant, Elie Odin (PhD student when project started), Mark van der Wilk.
- **Defended PhD theses:** 2 + 1 forthcoming in July 2026.
- **Publications:** 8 journal articles, 4 conference proceedings, 4 preprints.

2 State of the art at the start of the project

2.1 Gaussian processes for computer experiments and machine learning

Gaussian processes are a class of stochastic processes, that provide a Bayesian non-parametric prior on unknown functions. They enable to model the knowledge that is a priori available on these functions, through their mean and covariance functions. They also enable to update this knowledge with observation data, through the posterior distribution they provide. This posterior distribution is very valuable for uncertainty quantification purposes, as it provides optimal predictions of the function values, together with confidence intervals on them, and with the possibility to generate conditional sample functions. Because of the diverse uncertainty quantification tools they provide, Gaussian processes are used in many application fields, in particular computer experiments and machine learning.

In the context of computer experiments, a run of a costly numerical model corresponds to a function evaluation, where the input parameters are the simulation

[1] During most of the project duration, François Bachoc was affiliated to the University of Toulouse.

conditions (for instance the geometry parameters of the simulation domain) and where the output value is a quantity of interest of the simulation result. Gaussian processes are thus used to provide a prior distribution on these computer model functions and to predict simulation results for new input parameters.

In machine learning, Gaussian processes are applied to perform various standard tasks, including regression, single or multi-class classification and point process modeling. In regression, they provide a prior distribution on the conditional mean of outputs given inputs. In classification, they provide a prior distribution on the probabilities of the different classes, given the inputs. When modeling point processes, Gaussian processes provide a prior on the event intensity, as a function of the spatial location.

2.2 Main challenges at the start of the project

Gaussian processes provide competitive practical results in a large number of tasks, for computer experiments and machine learning. Nevertheless, compared to a profusion of empirical contributions, the theoretical guarantees for Gaussian process methods are scarce. For instance, there are many settings involving Gaussian processes where no rates of convergence are proved, regarding prediction errors. An important such setting is that of *deep Gaussian processes*. Another setting where open problems remain is that of *Gaussian process-based dimension reduction.*

Gaussian processes are applied on a regular basis to a set of standard tasks, for instance regression, classification and point process modeling as discussed above. There exist refinements of Gaussian processes, for instance *constrained Gaussian processes*, that have applicative restrictions, typically on the dimension of the input space of the modeled functions. There is a need to make these refinements available to a broader class of applications.

Furthermore, there are settings where Gaussian processes would be relevant but are not exploited, which is here a loss of opportunity. One such setting is *post-selection inference*.

At last, an important challenge for Gaussian processes is to extend their definition and use to complex inputs. An important example is *distributional inputs*.

3 Scientific contributions of the project

The project directly led to **16 articles**, listed at the end of this document, and presented in what follows.

3.1 Constrained Gaussian processes

Constrained Gaussian processes, in particular monotonic Gaussian processes, are very relevant in applications. We tackled a class of constrained Gaussian processes that guarantees that the constraints are satisfied *everywhere on the input space*.

In [15], we focused on scalability to high-dimensional inputs. We suggested additive monotonic Gaussian processes. We defined the model and provided methods for computing the posterior. Then we addressed the automatic selection of two parameters of the model: the active variables and the numbers and positions of knots of basis functions per active variables. This selection was made by a MaxMod procedure, sequentially maximizing the gain of information brought by adding variables or knots to the model. The numerical results were successful in input dimension up to 1,000.

In [12], we extended the previous work by suggesting block-additive constrained Gaussian processes. This made the new model suited to a larger class of problems, since non-additivity is often present in practice, but typically only between groups of variables of small size. We provided methods to compute the posterior, with different variants tailored to the quantities at hand (number of observations and of blocks). We also provided a MaxMod procedure to select the blocks and the numbers and positions of knots per active variable. The numerical results were successful, both on simulated data and on a real coastal flooding application.

Finally [14] addressed the theoretical foundations of constrained Gaussian processes. The article studied the convergence of the mode (maximum of posterior density) for a constrained Gaussian process, as the number of basis functions goes to infinity, for a fixed data set. We provided a quantitative bound for this convergence, that depends on the kernel regularity. This bound and the convergence also hold when the set of knots for the basis functions is not space-filling, which is relevant to study the convergence of the MaxMod procedures discussed above.

3.2 Deep Gaussian processes and Gaussian-process based dimension reduction

In [7] we provided a Bayesian non-parametric analysis of deep Gaussian processes. We considered a composition of multivariate Gaussian processes, where the composition structure and the kernels of the Gaussian processes were fixed in the Bayesian prior. In the setting of regression or classification, with iid random observations stemming from a fixed unknown function with the same compositional structure as the deep Gaussian process, we provided *posterior contraction rates* to the function as the number of observations goes to infinity. These rates highlight that having sparse compositional structures counteracts the curse of dimensionality of the input space. These rates also show the benefit of a higher smoothness of the function and of the Gaussian processes in the composition, when these smoothnesses are matched.

In [10], we improved the previous article in several ways. The compositional structure of the deep Gaussian process, and the smoothnesses of the composed Gaussian processes, needed not be fixed, but could be a component of the prior. This resulted in a statistical adaptation to the compositional structure and smoothness of the unknown function, where the obtained posterior contraction rates matched known minimax rates.

In [16], we studied a prior composed on a random active low-dimensional linear subspace of a high-dimensional input space, combined with a Gaussian process operating on the subspace. When the unknown function was also following a similar structure, we obtained posterior contraction rates to it. Our prior also consisted in random length scales and a random subspace dimension, enabling us to be adaptive to the true unknown subspace dimension and to smoothness. Furthermore, we could allow both the ambient dimension and the subspace dimension to increase with the number of observations. Intuitive arguments suggested that our obtained rates are optimal. Furthermore, we also showed that the posterior on the subspace could concentrate to subspaces containing the true unknown subspace. Hence, not only consistent function estimation was possible, but also consistent estimation of the function structure.

3.3 Gaussian processes for inference-post-region-selection

In [9], we considered a region-selection statistical procedure, typically in the context of detecting regions of interest for genetic data or brain imaging. We suggested confidence intervals for the average signal of the selected region, while allowing this selected region to be arbitrary. This is particularly important for robustness to the practice of the application fields. Obtaining a coverage guarantee for our confidence intervals necessitated to control for the supremum of a diverging number of averages indexed by locations on the domain of interest. We proved a functional convergence to a Gaussian process on this domain, which enabled us to optimally calibrate our confidence intervals and to obtain this coverage guarantee. We provided applications to simulated data and to the analysis of DNA copy number data in cancerology.

3.4 Gaussian processes (and kernels) indexed by distributions

In [1], we proposed a kernel for Gaussian processes indexed by distributions. The kernel was based on computing the Sinkhorn potentials of the entropy-regularized optimal transport problems to a common reference measure. These Sinkhorn potentials, belonging to a Hilbert space, could thus be given as inputs to any positive-definite kernel on a Hilbert space, resulting in a positive-definite kernel on the space of multi-dimensional distributions. We proved several useful theoretical properties of statistical consistency, both in the *one-stage sampling*, when the distributions are completely known, and in the *two-stage sampling*, where only samples from the distributions are observed. We obtained competitive numerical results in several standard benchmarks. Regarding the resulting distribution-indexed Gaussian process, we proved the existence of continuous versions, which is in particular relevant to study the convergence of goal-oriented sequential design of experiments.

In [2], we did not address Gaussian processes *per se*, but the very related topic of kernel ridge regression on distributional spaces. We considered a more general setting than previously, by studying kernels operating on Hilbertian embeddings of the distributions. Thanks to an original mathematical analysis, we obtained improved rates of convergence for the necessary number of samples per distributions, for the convergence rates in the two-stage sampling setting, to match the minimax ones for the one-stage sampling setting. We also provided a numerical application of kernel distribution regression to ecological inference, with the aim of inferring county-level votes in a US presidential election.

3.5 Other contributions

3.5.1 Post-selection inference

In [8], we tackled a problem of post-selection inference related to the region-selection one. Here, a multi-dimensional clustering was selected by aggregating one-dimensional ones, that were themselves obtained by convex optimization. The goal was to test for the difference of average signal, between different clusters, for specific variables. We first characterized the event of obtaining given clusters by a set of linear constraints on the Gaussian observation vector. This enabled us to obtain a test procedure with a theoretical guarantee for its level. The test was successfully applied to simulated data. In passing, we obtained a numerically competitive solution path algorithm for one-dimensional convex clustering.

In, [13] we also studied inference post-clustering, but for other clustering procedures, notably hierarchical agglomerative clustering algorithms with several types of linkages, and the k-means algorithm. We obtained test procedures with theoretical guarantees of their levels, for Gaussian data with a general dependence structure. We provided successful applications to synthetic data and real data of protein structures.

3.5.2 Online learning for bilateral trade

In [4], we studied sequential strategies for bilateral trade, where the goal is to sequentially allocate prices to maximize a fair gain from trade of two random buyer and sellers. While this problem does not directly involve Gaussian processes, the challenges and proof methods can be closely related to the analysis of sequential designs of experiments for Gaussian processes. We obtained matching lower and upper bounds for the *cumulated regret* for various versions of this problem.

In [3], we studied a similar problem as before, with the difference that the buyer and seller were replaced by two agents with exchangeable roles. This yielded a more favorable setting, where often one could reduce the problem to the sequential estimation of a mean. We obtained matching lower and upper bounds for the *cumulated regret* for various versions of this problem.

In [5], we studied an extension of the problem of the previous article, where now for each price to be allocated, a context vector was available. We used an online ridge regression to obtain new algorithms together with their upper bounds on the cumulated regret. We also provided lower bounds, that matched our upper bounds, both with respect to the ambient dimension and the time horizon of the sequential problem.

3.5.3 Distributional data and problems: statistical depth and fairness of push-forward measures

In [6], related to our work on Gaussian processes indexed by distributions, we suggested a notion of spatial depth for distributions. Our proposal naturally extends the Euclidean spatial depth and keeps its main theoretical benefits. Furthermore, we showed how consistent estimation, with rates of convergence, is possible both in the one-stage and two-stage sampling. We demonstrated the benefit of the spatial depth, in particular for outlier detection, both with simulated data and a real data set related to the evolution of European meteorological temperatures.

Finally, in [11], also motivated by our above works related to optimal transport, we studied the problem of deterministically pushing forward two probability measure with one transport map, with the constraint of obtaining the same image probability measure. This problem is important for the field of statistical fairness. We provided general conditions under which the set of maps, such that indeed the images are equal, is non-convex. We then studied the impact of these negative results to optimization problems encountered in statistical fairness.

4 Benefits of the projects to the participants and community

4.1 PhD theses

- The project GAP funded the PhD thesis of Mathis Deronzier (2023-2026). The aimed date for the defense is July 2026. Mathis Deronzier published two journal articles ([11, 12]), wrote one preprint [13], and has two articles in preparation.
- Elie Odin was a PhD student participating to the project GAP (2022-2025, not funded by the project). He defended his PhD thesis in December 2025 with one article [16], and he is currently preparing two preprints. His next goal is to obtain a postdoctoral or temporary teaching and research (ATER) position.
- Thibault Randrianarisoa was a PhD student participating to the project GAP (2019-2022, not funded by the project). He wrote the article [10] on a core topic of the ANR GAP project. After two post-doctoral positions following his PhD thesis, Thibault Randrianarisoa is now Assistant Professor at the University of Toronto.

4.2 Other project members

The project ANR GAP was very beneficial to the scientific development of its other members. There were two internal workshops per year, that enabled all the participants to develop new scientific contacts and collaborations. In 2023, the ANR project BACKUP was funded, with a participant to the project GAP as PI, and several participants to the project GAP as participants, or in charge of the Toulouse team of the BACKUP project. In 2024, several of the participants of the ANR GAP project became PI or co-chairs of research Chairs of the Toulouse ANITI AI cluster. Two participants of ANR GAP were promoted from Tenured Assistant Professors to Professors. Finally, the PI of ANR GAP became Junior Member of Institut Universitaire de France in 2024.

4.3 Benefits to the community

The outcomes of the project were presented in multiple conferences. In particular, nationally there were presentations at the annual workshop of the French thematic network on uncertainty quantification, and at the annual conference of the French statistical society. Internationally, there were presentations at the International Society of Bayesian Analysis.

There was a closing conference of the project, entitled *Gaussian processes and related topics*, in Toulouse in July 2025, of European breadth, with about 70 participants, in particular from France, Italy, the UK and Switzerland:

https://indico.math.cnrs.fr/event/13402/.

Publications of the ANR JCJC GAP project

[1] F. Bachoc, L. Béthune, A. Gonzalez-Sanz, and J.-M. Loubes. Gaussian processes on distributions based on regularized optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 4986–5010. PMLR, 2023.

[2] F. Bachoc, L. Béthune, A. González-Sanz, and J.-M. Loubes. Improved learning theory for kernel distribution regression with two-stage sampling. *Annals of Statistics*, 53(4), 2025.

[3] F. Bachoc, N. Bolić, T. Cesari, and R. Colomboni. An online learning theory of brokerage. *https://hal.science/hal-04987905/*, 2025.

[4] F. Bachoc, N. Cesa-Bianchi, T. Cesari, and R. Colomboni. Fair online bilateral trade. *Advances in Neural Information Processing Systems*, 37:37241–37263, 2024.

[5] F. Bachoc, T. Cesari, and R. Colomboni. A parametric contextual online learning theory of brokerage. In *Forty-second International Conference on Machine Learning*.

[6] F. Bachoc, A. González-Sanz, J.-M. Loubes, and Y. Yao. Wasserstein spatial depth. *arXiv preprint arXiv:2411.10646*, 2024.

[7] F. Bachoc and A. Lagnoux. Posterior contraction rates for constrained deep Gaussian processes in density estimation and classification. *Communications in Statistics-Theory and Methods*, 54(3):774–811, 2025.

[8] F. Bachoc, C. Maugis, and P. Neuvial. Selective inference after convex clustering with ℓ_1 penalization. *ESAIM: Probability and Statistics*, 29:204–242, 2025.

[9] D. Bontemps, F. Bachoc, and P. Neuvial. Inference post region selection. *arXiv preprint arXiv:2506.11564*, 2025.

[10] I. Castillo and T. Randrianarisoa. Deep horseshoe Gaussian processes. *The Annals of Statistics*, 53(5):1886–1912, 2025.

[11] L. De Lara, M. Deronzier, A. González-Sanz, and V. Foy. On the nonconvexity of push-forward constraints and its consequences in machine learning. *SIAM Journal on Mathematics of Data Science*, 7(2):597–620, 2025.

[12] M. Deronzier, A. F. López-Lopera, F. Bachoc, O. Roustant, and J. Rohmer. Block-additive Gaussian processes under monotonicity constraints. *Statistics and Computing*, 36(1):42, 2026.

[13] J. González-Delgado, M. Deronzier, J. Cortés, and P. Neuvial. Post-clustering inference under dependence. https://arxiv.org/abs/2310.11822v2, Oct. 2023.

[14] L. Grammont, F. Bachoc, and A. F. López-Lopera. Error bounds for a kernel-based constrained optimal smoothing approximation. *SIAM/ASA Journal on Uncertainty Quantification*, 13(3):1145–1173, 2025.

[15] A. López-Lopera, F. Bachoc, and O. Roustant. High-dimensional additive Gaussian processes under monotonicity constraints. *Advances in Neural Information Processing Systems*, 35:8041–8053, 2022.

[16] E. Odin, F. Bachoc, and A. Lagnoux. Contraction rates and projection subspace estimation with Gaussian process priors in high dimension. *Bernoulli*, 32(2):1645–1664, 2026.