Accelerating the BSM interpretation of LHC data with machine learning

Gianfranco Bertone,1 Marc Peter Deisenroth,2 Jong Soo Kim,3 Sebastian Liem,4 Roberto Ruiz de Austri,4 and Max Welling5

1GRAPPA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands
2Department of Computing, Imperial College London, 180 Queen’s Gate, SW7 2AZ London, United Kingdom
3Center for Theoretical Physics of the Universe, Institute for Basic Science (IBS), Daejeon, 34051, Korea and Instituto de Física Teórica UAM/CSIC, Madrid, Spain
4Instituto de Física Corpuscular IFIC-UV/CSIC, Valencia, Spain
5Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands
(Dated: November 10, 2016)

The interpretation of Large Hadron Collider (LHC) data in the framework of Beyond the Standard Model (BSM) theories is hampered by the need to run computationally expensive event generators and detector simulators. Performing statistically convergent scans of high-dimensional BSM theories is consequently challenging, and in practice unfeasible for very high-dimensional BSM theories. We present here a new machine learning method that accelerates the interpretation of LHC data, by learning the relationship between BSM theory parameters and data. As a proof-of-concept, we demonstrate that this technique accurately predicts natural SUSY signal events in two signal regions at the High Luminosity LHC, up to four orders of magnitude faster than standard techniques. The new approach makes it possible to rapidly and accurately reconstruct the theory parameters of complex BSM theories, should an excess in the data be discovered at the LHC.

Introduction: A vast effort is currently in progress to discover physics Beyond the Standard Model (BSM) at the Large Hadron Collider (LHC), motivated in part by the possible connection between new particles at the weak scale and the dark matter problem in astrophysics and cosmology [1]–[3]. The absence of clear evidence for BSM physics in current LHC data has been interpreted in the context of simplified models [4,5] as well as of full models, such as various incarnations of the minimal Supersymmetric Standard Model (MSSM) [6]–[10].

Such studies, and even more the interpretation of a hypothetical excess in future data, are hampered by the computationally intensive task of sampling the high-dimensional parameter space of theoretical models, and comparing, for each sample, the predicted signal with actual data. For each set of input parameters one needs in fact to: (i) generate a Monte Carlo (MC) sample of collision events; (ii) run the sample through a detector simulation; and (iii) compare the predicted signal with data, often within signal regions (SRs) defined by experimental cuts on observable quantities, such as missing transverse energy, number of jets, momenta, and angles [7]. This procedure is computationally very expensive, and it constitutes the bottleneck for global analyses of BSM theories, especially for those with high-dimensional theory parameter spaces: in Ref. [6], for instance, it was estimated that \(400\) CPU-years would be needed to obtain a statistically convergent scan of a 15-dimensional supersymmetric model.

We demonstrate here that this bottleneck can be bypassed by introducing machine learning (ML) tools that can learn the mapping between theory and data, and then rapidly and accurately predict signal region efficiencies.

Gaussian processes: The number of events \(N_i\) in SR \(i\) can be written as \(N_i = L \sigma_i \epsilon_i\), where \(L\) is the integrated luminosity, \(\sigma\) the production cross-section of the relevant process(es), and \(\epsilon_i \in [0,1]\) is the SR efficiency (which is in turn the product of the detector efficiency times the acceptance, i.e. the fraction of events that passes analysis cuts). A classification ML method was introduced in Ref. [11] to predict whether or not a given point in the BSM theory parameter space is compatible with LHC data. Here, we are interested in the more general regression problem of estimating the continuous quantities \(\epsilon_i\) given the input BSM parameters \(\theta\), i.e. in modeling the relationship \(\epsilon_i = f_i(\theta)\).

We specifically implement here the Gaussian process (GP) regression model [12]. Instead of predicting a single value, a GP has the virtue of equipping predictions with consistent uncertainty estimates by means of a full posterior distribution. The crucial ingredient of GPs is the covariance function, which specifies the correlation structure between the function value at different points in the input parameter space. We use here for the covariance function an anisotropic squared exponential kernel [12]

\[
k(\theta, \theta') = \sigma_f^2 \exp \left( \sum_j \frac{(\theta_j - \theta'_j)^2}{2l_j^2} \right)
\]

where the sum is over the BSM theory parameters. \(\sigma_f\) and \(l_j\) are hyperparameters: \(\sigma_f^2\) encodes the intrinsic variance of the function we are modeling, and the \(l_j\) are characteristic length-scales which determine how quickly the function changes from point to point. Choosing the
optimal values of these hyperparameters to model our function is the learning task of GPs, and is done by the standard procedure of evidence maximization [12].

The major limitation of standard GPs is that training scales cubically with the size $n$ of the training data set as it involves computing the inverse of $n \times n$ matrices. Therefore, in practice, there is a limitation on the amount of training data that can be used. To eliminate this limitation we make use of distributed GPs (DGPs), specifically the robust Bayesian Committee Machine [13] algorithm, which avoids large matrices by partitioning the training data into smaller data sets and distributing the computation across independent computing nodes.

Natural supersymmetry: As a proof-of-concept, we apply this new technique to the natural supersymmetry (SUSY) scenario, in which fine-tuning is low, and the electroweak scale is stabilized by a small subset of light SUSY states (e.g., $\mu$, $\mu_c$). We focus in particular on the minimal natural SUSY scenario of Refs. [19, 20], a realistic, yet low-dimensional theory, in which the gluinos, both stops, the left handed sbottom, and the higgsinos all have masses at TeV scale while the remaining states are decoupled. The six parameters of minimal natural SUSY are: the supersymmetric Higgs mixing parameter $\mu$, the gluino mass parameter $M_3$, the ratio of the two Higgs vacuum expectation values $\tan \beta$, the third generation SU(2)-doublet squark soft-breaking parameter $m_{Q_3}$, the third generation SU(2)-singlet soft-breaking parameter $m_{\ell s}$, and the top trilinear soft-breaking term $A_t$.

Data: The experimental scenario we consider is the planned high luminosity upgrade of the LHC (HL-LHC) [21] with 3000 fb$^{-1}$ worth of data collected at 14 TeV center-of-mass energy. We focus on two mutually exclusive SRs defined in Ref. [22], for which the ATLAS collaboration provides background estimates. These SRs are optimized for direct production of stops, the most relevant production channel for natural SUSY. The typical decay channels for the stop are: top or bottom quarks, W/Z/Higgs bosons, and the lightest neutralino. The detector signature is the presence of several jets (including b-jets), large missing transverse energy, and possibly leptons. We refer to the ATLAS note for the full definitions of the SRs, and we focus here on the 0-lepton and a 1-lepton SR.

Training and testing: For training and test data we analyzed 18647 samples generated in Ref. [20], for which SR efficiencies were calculated using SPheno 3.2.4 [23], Pythia 8.210 [24, 25] with default parton distribution function set [26], NLLFast 3.1 [27] [28], and Check-MATE 1.2.1 [29, 30] with Delphes 3.10 [31] and Fast-Jet 3.0.6 [32, 33].

We used 16647 of these samples to train DGPs for the two SRs, with one single level architecture with an ensemble of 256 GP. Training was fast due to the use of the DGP algorithm and took approximately 15 minutes on a desktop computer with a 4.0 GHz Intel 4790K processor. We then tested the predictions of the trained DGPs on the remaining 2000 points. In Fig. 1 we show the efficiency predicted by the DGP model in the 1-lepton SR, $\epsilon_{1\ell}^{\text{dgp}}$, versus the values calculated with the full MC calculation, $\epsilon_{1\ell}^{\text{mc}}$, for these 2000 test points. The DGP model accurately predicts the efficiencies, which cluster around the orange line defined by $\epsilon_{1\ell}^{\text{dgp}} = \epsilon_{1\ell}^{\text{mc}}$, with a spread consistent with the DGP error estimate, $\sigma_{1\ell}^{\text{dgp}}$. We can quantify the agreement by calculating the $\chi^2$; for both the 0-lepton and 1-lepton SRs we get $\chi^2 \approx 1300$, while naively expecting $\chi^2 = 2000 \pm 64$ given the 2000 degrees of freedom. The reason for these low values of $\chi^2$ is that the DGP model slightly overestimates its error. We visualize this in the inset of Fig. 1 where we see that the distribution of $(\epsilon_{1\ell}^{\text{dgp}} - \epsilon_{1\ell}^{\text{mc}})/\sigma_{1\ell}^{\text{dgp}}$ is more peaked than the standard normal distribution $N(0,1)$.

Reconstruction: The DGP model, thus, effectively acts as a surrogate model for the full simulation chain, opening up new opportunities in the interpretation of LHC data. For example, the DGP model can rapidly reconstruct the theory parameters of a BSM model in the case where an excess is observed on top of the Standard

\[1 \text{ In Ref. [22] they vary the cuts as a function of the stop mass. We use the cuts optimized for 1.1 TeV.} \]
We further restrict the Higgs boson mass to the range $121 < m < 129$ GeV and the chargino mass $> 103.5$ GeV. The uncertainty on the mock signal includes a contribution of the likelihoods for the two signal regions.

The orange histograms show the marginal posteriors when we also fit to the mock excess. We scan over both positive and negative $\mu$ but show only results for $|\mu|$ as they are symmetric.

Model background. To demonstrate the feasibility of a full reconstruction procedure, we perform it on a mock data set generated assuming a future excess will be detected. Our benchmark is the following point in the natural SUSY parameter space: $\mu = 254.6$ GeV, $\tan \beta = 20$, $M_3 = 2000$ GeV, $m_{Q_3} = 1280.5$ GeV, $m_{t_R} = 1333.6$ GeV, $A_t = -2000$ GeV. The physical masses are $1145$ and $1413$ GeV for the stops, $1259$ GeV for the sbottom, and $254$ GeV for the neutralino mass. This benchmark point is not excluded by current searches [41–44], but should lead to a detectable signal in our signal regions: $33$ events. The posterior probability distributions after taking the excess in the mock data into consideration. As one can see, the mock data in the two signal regions have a very limited impact on the determination of $\mu$, but they lead to a measurement on $m_{Q_3}$, and to a more stringent lower limit on $m_{t_R}$. The posterior slightly disfavors larger values of $|\mu|$ because it determines the neutralino mass, and the missing energy cut requires that the mass difference between the neutralino and the produced squark is large. The parameters $m_{Q_3}$ and $m_{t_R}$, together with the off-diagonal entries of the stop mass matrix, determine the mass of the three light squarks; $b_1$, $t_1$, and $\tilde{t}_2$. The analysis is sensitive to these masses through the production cross-sections, as the total cross-section must be large enough to produce the measured number of events. The total production cross-section is dominated by two contributions: one that depends on both $m_{Q_3}$ and $m_{t_R}$ (via the mass of $\tilde{t}_1$) and one that depends on $m_{Q_3}$ only (via the mass of $b_1$). The constraint on the total production cross-section therefore translates more directly into a constraint on $m_{Q_3}$, explaining why we are able to reconstruct $m_{Q_3}$ better than $m_{t_R}$.

The core of SuperBayeS is the MultiNest scanning algorithm [24,25] which is interfaced with PHENO 3.2.4 [23] and NLLFAST 3.1 [27,29] to calculate observables. The likelihood is implemented in the ROOT framework [30] with the RooFit [31] and RooStats [32] packages.

FIG. 2. Reconstruction of the natural SUSY parameters $\mu$, $m_{Q_3}$, and $m_{t_R}$ using a mock excess generated from the benchmark point. The DGP was used to calculate the SR efficiencies; no event generation was performed. Gray histograms are the marginal priors where we only apply bounds on the Higgs boson mass $\in [121, 129]$ GeV and the chargino mass $> 103.5$ GeV. The orange histograms show the marginal posteriors when we also fit to the mock excess. We scan over both positive and negative $\mu$ but show only results for $|\mu|$ as they are symmetric.
In reconstructing the benchmark parameters from mock data we used 4 000 live points in MultiNest, which required $10^8$ likelihood evaluations before converging. The estimate of the two SR efficiencies with the DGP model took 0.06 seconds per evaluation on a single 4 GHz Intel 4790K core, a factor $\sim 10^4$ faster than the $O(10)$ minutes per evaluation required to generate the training set. The whole scan took 66 hours on six CPUs.

**Discussion and Conclusions:** In this letter, we have introduced the use of Gaussian processes to accelerate the interpretation of LHC data in the framework of BSM theories. Their ability to estimate an error on their predictions makes them ideal for fast and robust approximate calculations. We have specifically demonstrated that the estimate of SR efficiencies can be accelerated by a factor $10^4$, making it possible to rapidly and accurately reconstruct the natural SUSY theory parameters, should an excess in the data be discovered at the HL-LHC. The method can be generalized to any BSM theory, and it can be in principle extended to accelerate other time-consuming tasks, such as the calculation of the cross-sections, or of the likelihood itself.

Gaussian processes are currently a very active area of research and it is likely that the method will be further improved and refined. One particularly interesting extension to our implementation is multi-output prediction: instead of training one (D)GP for each SR separately, one can train a single (D)GP that predicts all the SR efficiencies simultaneously. The correlation between the SR efficiencies could then be used to make more precise predictions. Another direction is to move away from Gaussianity and use Student-t processes [56] which might model the underlying noise for the Monte Carlo generators better. Another intriguing development is Bayesian optimization [57], a form of active learning [58, 59], which aims to minimize the amount of training data needed, by letting the Gaussian process itself specify where to sample the theory parameter space next in an iterative fashion.

In the traditional approach to global analyses, new samples have to be generated in the parameter space of a BSM theory, and new simulations performed every time new data become available, as the sampling is driven by the likelihood. An important aspect of our new method is that detector simulations need to be performed only once for each BSM theory, to generate the training sample. Once this is done, the surrogate model can be reused by anyone, and applied to any data set.

**Acknowledgments:** We thank S. Caron and C. McCabe for their useful comments. G.B. (P.I) and S.L. acknowledge support from the European Research Council through the ERC starting grant WIMPs Kairos. M.P.D. acknowledges support from a Google Faculty Research Award. The work of R. RdA was supported by the Ramón y Cajal program of the Spanish MINECO and also thanks the support of the grants FPA2014-57816-P and FPA2013-44773, and the Severo Ochoa MINECO project SEV-2014-0398. He also acknowledge specially the support of the Spanish MINECO’s Consolider-Ingenio 2010 Programme under grant MultiDark CSD2009-00064. JSK wants to thank T. Weber for his help with the efficiency files. The work of JSK was supported by IBS under the project code, IBS-R018-D1 and was partially supported by the MINECO, Spain, under contract FPA2013-44773-P; Consolider-Ingenio CPAN CSD2007-00042 and the Spanish MINECO Centro de excelencia Severo Ochoa Program under grant SEV-2012-0249.
