Chapter 10

Fast kernel methods for Data Quality Monitoring as a goodness-of-fit test

In this chapter, we address the data quality monitoring problem in high-energy physics domain outlined in Chapter 1. The objective is to evaluate the compatibility of incoming experimental data with a reference dataset obtained under normal circumstances. Additionally, given the high data volume generated by experimental devices used in high-energy physics experiments, we aim to design a procedure capable of detecting anomalies in real-time.

To accomplish this, we propose an accurate and efficient machine learning approach for real-time monitoring of particle detectors. Our method assesses the compatibility of experimental data with the reference dataset using a likelihood-ratio hypothesis test. The model is based on modern kernel methods, which are nonparametric algorithms capable of learning any continuous function given sufficient data.

The resulting approach is efficient and agnostic to the types of anomalies present in the data. Our study demonstrates the effectiveness of this strategy using multivariate data from drift tube chamber muon detectors.

Contributions.

- We introduce a new machine learning pipeline for monitoring real-time particle detectors.
- We verify the efficiency and the effectiveness of our pipeline on an empirical experiment.

Status.

This chapter is based on our paper: Grosso, G., Lai, N., Letizia, M., Pazzini, J., Rando, M., Rosasco, L., Wulzer, A. & Zanetti, M. (2023). **Fast kernel methods for data quality monitoring as a goodness-of-fit test**. Machine Learning: Science and Technology, 4(3), 035029. [GLL^{+}23].

10.1 Introduction

Modern high-energy physics experiments operating at colliders are extremely sophisticated devices consisting of millions of sensors sampled every few nanoseconds, producing an enormous throughput of complex data. Several types of technologies are employed for identifying and measuring the particles originated in the collisions; in all cases, the environmental conditions are severe, making the required performances challenging to achieve. Although the various subsystems are designed to offer redundancy, measurements can be undermined by malfunctions of parts of the experiment, either because of critical inefficiencies or because of possibly misinterpreted spurious signals. In addition to supervising the status (powering, electronic configuration, temperature, etc.) of the various hardware components, data from all sources must thus be monitored continuously to assess their quality and to promptly detect any faults, possibly providing indications about their causes. Given the rate of tens of MHz at which data is gathered and the number of sensors to be checked, the monitoring process needs to be as automated as possible: approaches based on Machine Learning (ML) techniques are particularly appealing for this task and have started being employed by the experimental collaborations [PCGP22, PCG$^+$19, AAC$^+$19, AAB$^+$17], complementing more traditional methods [Rov15, AvBB$^+$19, Mar19, Kau22, A$^+$20]. Data quality monitoring (DQM) consists, in essence, of comparing batches of data with corresponding reference samples gathered in nominal conditions; departures from the latter can then be analysed to identify their origin. The data processing must fit the computational constraints imposed by the frequency at which batches are delivered and by their size, with the latter depending on the granularity with which sensors are grouped and the statistical uncertainty aimed at.

In this work, we present the application of a methodology developed in the context of model-independent searches for new physics [DW19, DGP$^+$21, dGP$^+$22]—specifically of its kernel methods implementation [LLR$^+$22] based on the Falkon [MCRR20] library—as an efficient and effective DQM tool. The method (dubbed NPLM) implements a hypothesis test leveraging the ability of classifiers to infer the underlying data-generating distributions in order to estimate the likelihood ratio test statistic. The Falkon-based implementation of NPLM offers tremendous advantages in terms of training time compared to the one based on neural networks. It can thus be used for DQM.

Conventional DQM methods typically consider a number of one-dimensional distributions; a key feature of NPLM is the capability of examining the phase space as a whole, not depending critically on the choice of input variables and being sensitive to their correlation. It is then possible to provide low-level quantities to the algorithm that require limited pre-processing. This can be particularly advantageous for DQM, as it allows it to deal with almost raw data from the detectors' electronic front-ends, therefore limiting the bias introduced by further manipulations that could hide issues in the data.

To test the effectiveness of NPLM for DQM, we exploit an experimental setup which we have full control of, consisting of a reduced-size version of the muon chambers installed in the CMS experiment at the Large Hadron Collider (LHC). The setup is operated as a cosmic muon telescope. As explained later, scaling tests are performed to assess the performances of the DQM algorithm in view of its possible deployment during standard LHC operations.

The paper is organised as follows. In the next section, we introduce the experimental setup and the algorithm input variables. These include a reference data set collected under standard conditions and smaller samples with anomalous controlled behaviours. The ML model and our

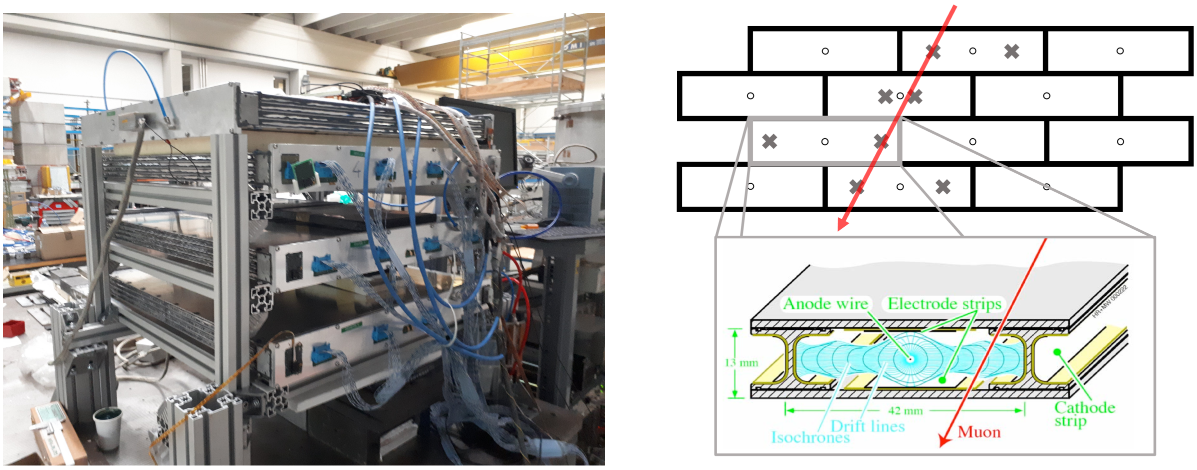

Figure 10.1: Left: the experimental apparatus at Legnaro Laboratory, with four drift-tube chambers, vertically stacked. Right: a schematic view of the cell (bottom) and an example of hit pattern left by a charged particle crossing a chamber (top).

core strategy are then described in Section 10.3, whereas an overview of the results is given in Section 10.4. Finally, the last section is devoted to conclusions and further developments.

10.2 Experimental setup

For this research, we exploited an experimental apparatus consisting of a set of Drift Tube (DT) chambers housed at the Legnaro INFN National Laboratory (Fig. 10.1, left). These chambers are a smaller in size copy of those deployed in the CMS experiment at the LHC [C$^+$08]. The basic element of a DT chamber is a 70 cm long tube with a cross section of 4×2.1 cm^2 (Fig. 10.1, bottom right). Inside each tube, an electric field is produced by an anodic wire laid in the centre and two cathodic strips (cathodes) on the sides; the former is set at a voltage of 3.6 kV, the latter at -1.2 kV. An additional pair of strips at 1.8 kV is placed above and below the wire to improve the homogeneity of the field. The tubes are filled with a mixture of argon and carbon dioxide gas (85%-15%) that gets ionised by charged particles passing through it. The produced electrons drift towards the wire at a constant velocity along the field lines, where they are collected. For each tube, the front-end electronics record the arrival time of the ions, amplify the signal, and filter out noise below a specific threshold (nominally 100 mV).

A drift tube chamber consists of 64 tubes arranged in four layers of 16 tubes each. The layers are staggered horizontally by half a cell. The setup at Legnaro records muons from cosmic rays, which occur at a rate of about 1 per minute per cm^2 at sea level. Data acquisition occurs continuously at a rate of 40 MHz, without the need for any trigger logic. An external time reference is provided by plastic scintillators placed in between the DT chambers; the corresponding information is added to the data stream and used in the following analysis steps.

Thanks to the homogeneity of the electric field, the particle's position within each tube (with a left-right ambiguity) is linearly dependent on the drift time. Namely, the distance of the muon track from the wire reads

$$x_{\pm} = \pm v_d \, (t_{hit} - t_0) \; = \pm v_d \, t \, , \tag{10.1}$$

with t_{hit} the time associated to each signal in a tube (called a hit). The two parameters are the drift velocity v_d, known by means of a calibration procedure (in our case, $v_d = 53\,\mu$m/ns), and the time pedestal t_0, which can be deduced from the timing information provided by the scintillators[1]. The drift time t is obtained by the difference between t_{hit} and the time pedestal.

The hits occurring in a time window of $90\,\mu$s centred around the signal provided by the scintillators are grouped in quadruplets (with one hit pertaining to each of the four layers as in Fig. 10.1, right top). Then, a linear fit is performed on each of the quadruplets and the candidate muon track is obtained from the combination yielding the best χ^2. In this way the trajectory of the muon in the plane transverse to the tubes is determined, with a precision on the position of about $180\,\mu$m and on the slope of about 1 mrad. Tracks from various DT chambers can be combined to determine the 3D muon trajectory; in the following we will however consider only the 2D measurement.

If the detector conditions are anomalous, the efficiency and accuracy of the muon track reconstruction may be compromised. Ensuring the proper operation of the detector thus requires monitoring the quality of the recorded data. In what follows, we consider six basic quantities related to the passage of a muon through a DT chamber:

- Drift times t_i: the four drift times associated with the muon track. The drift time distribution is displayed in Fig. 10.2 in different ranges for the muon track angle θ (or "slope", see the next item), showing the correlations between these two variables. The t_i distributions are also reported in Figs. 10.3 and 10.4.
- Slope θ: the angle formed by the muon track with the vertical axis. The chamber efficiency is expected to drop beyond $|\theta| \sim 40$ degrees as we see in Figs. 10.3 and 10.4.
- Number of hits n_{Hits}: the number of hits recorded in a time window of one second around the muon crossing time. Many spurious hits are present in addition to those due to the passage of a muon. The noise rate depends on the environmental conditions, with the one at the LHC orders of magnitude larger than that of our laboratory in Legnaro, but the recorded spurious hits rate can also be affected by issues related to the detector operation conditions.

The six variables $x = \{t_1, \dots, t_4, \theta, n_{Hits}\}$ will be the input features of the NPLM algorithm for DQM, described in the next sections. Notice that the data are gathered from the subset of tubes in a single chamber that geometrically matches the scintillators, i.e. about three tubes per layer.

We collected the data by artificially inducing possible issues that can occur during detector operations. Specifically, we reduced the voltage of the cathodic strips to 75%, 50%, and 25% of their nominal value (-900 V, -600 V, and -300 V, respectively), and we lowered the front-end thresholds to 75%, 50%, and 25% of their nominal value (75 mV, 50 mV, and 25 mV, respectively). The former action distorts the electric field shape, whereas the latter mimics the sudden contribution of noise sources. We conducted a dedicated data acquisition campaign in these six anomalous configurations, collecting around 10^4 events for each configuration. We also collected around 3×10^5 data points in the normal (or, reference) working conditions of the apparatus.[2] The distribution of the six input features for the reference data and the data

[1] In addition a mean-timer algorithm [MPT+22] is executed on the back-end board receiving the data. The timing information provided by that algorithm is currently not used in this analysis.

[2] Dataset available at `https://doi.org/10.5281/zenodo.7128223`.

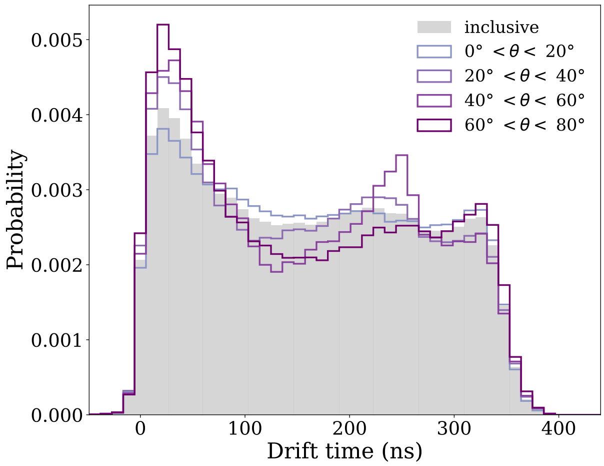

Figure 10.2: Drift time distribution in different θ ranges.

collected under the different anomalous conditions are shown in Fig. 10.3 (variation of the cathodes voltages) and Fig. 10.4 (variation of the thresholds). These data will be used to design and calibrate the DQM algorithm, as described in the following section.

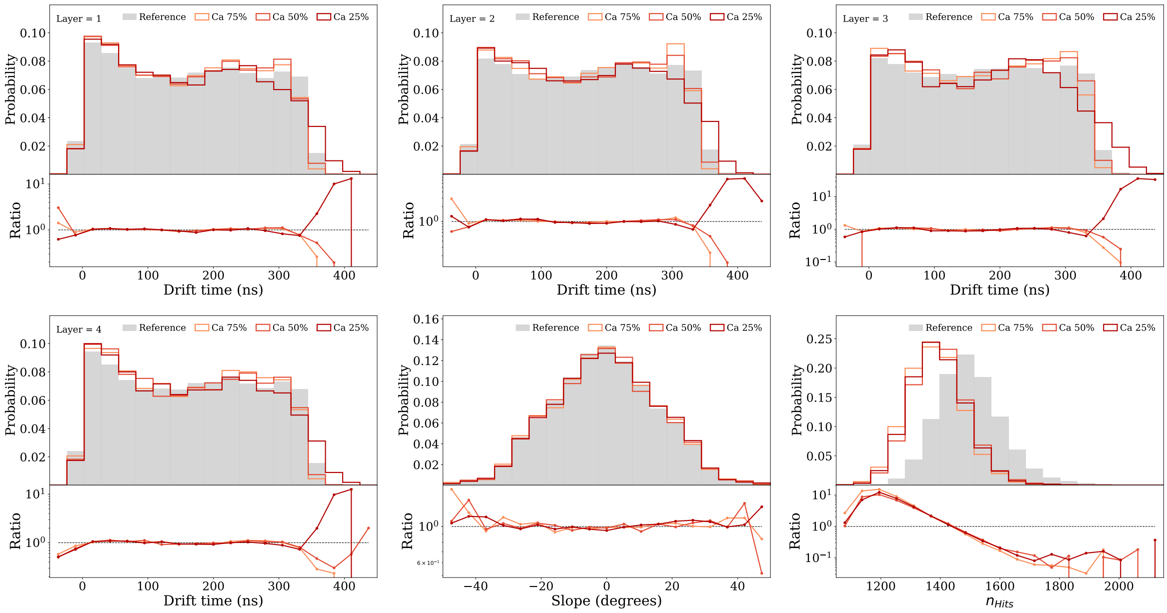

Figure 10.3: The distribution of the input features in the reference and in three anomalous working conditions of the cathodes voltages

10.3 Methodology

In the setup described in the previous section, we are interested in assessing the quality of individual batches of data collected by the apparatus, each of which denoted as $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}$. Namely, we ask whether the statistical distribution of the data points in $\mathcal{D}$ coincides or not with the one expected under *reference* working conditions, $p(x|\mathrm{R})$. We thus aim at performing what is known in statistics as a *goodness-of-fit test*. See [Cou] for references and a concise overview.

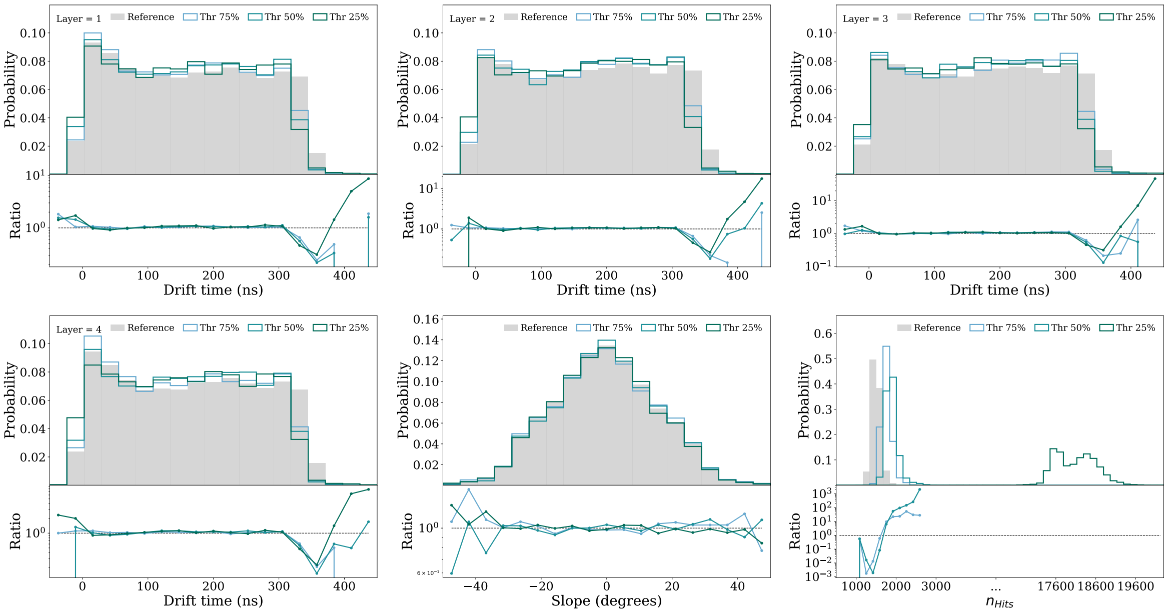

Figure 10.4: The distribution of the input features in the reference and in three anomalous working conditions of the thresholds.

The reference distribution $p(x|\text{R})$ is not available in closed form. What is available is instead a second dataset $\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}$ collected by the same apparatus when operated in the reference working conditions, such that the data in $\mathcal{R}$ do follow the $p(x|\text{R})$ distribution. Our goodness-of-fit test is thus carried out by comparing the two datasets $\mathcal{D}$ and $\mathcal{R}$, asking whether they are sampled from the same statistical distribution. The problem can then be formulated as a *two-sample test*, in which, however, $\mathcal{D}$ and $\mathcal{R}$ play asymmetric roles.

The data batch $\mathcal{D}$ is what needs to be tested. Therefore its composition and its size, $N_\mathcal{D}$, are among the specification requirements of the DQM methodology we are developing. $N_\mathcal{D} \sim 1000$ is in the ballpark of what is typically considered by DQM applications deployed at CMS.

The reference dataset $\mathcal{R}$ is instead created within the methodology design, with mild or no limitation on its size, $N_\mathcal{R}$. A larger $\mathcal{R}$ dataset offers a more faithful representation of the underlying reference statistical distribution and therefore a more accurate test. Furthermore, taking $N_\mathcal{R}$ larger than $N_\mathcal{D}$ reduces the effect of the $\mathcal{R}$ dataset statistical fluctuation on the outcome of the test, leaving only those inherently due to the fluctuations of $\mathcal{D}$. This makes the outcome for a given data batch $\mathcal{D}$ nearly independent on the specific instance of the set $\mathcal{R}$ that is employed for the test, making the result more robust. In what follows, we will thus preferentially consider an unbalanced setup for the two datasets, with $N_\mathcal{R} > N_\mathcal{D}$. We will further exploit the availability of a relatively large volume of data collected under the reference working conditions for calibrating the test statistics variable and for selecting the hyperparameters, as discussed in the following.

The availability of a large set of data that are accurately labelled as having been collected under the reference detector conditions deserves further comments. These data are routinely available, in particular in high-energy physics experiments, and are in fact used for the design and calibration of regular DQM methods [Rov15, AvBB$^+$19, Mar19, Kau22, A$^+$20]. They are validated by a careful offline inspection, which typically requires human intervention. This validation process is way too demanding and slow to be employed as a DQM algorithm. The purpose of DQM is in fact to monitor the data quality online, i.e. while they are being collected.

The offline validation is instead straightforwardly capable of producing labelled reference data samples that are way larger than individual data batches.

10.3.1 The NPLM method

We employ the "New Physics Learning Machine" (NPLM) method, which was proposed and developed by some of us [DW19, DGP$^+$21, dGP$^+$22, LLR$^+$22] to address a similar problem in the different context of searches for new physical laws at collider experiments. The search for *new physics* is performed by comparing the measured data with a reference dataset whose statistical distribution is the one predicted by a *standard* set of physical laws that supposedly describe the experimental setup. The purpose of the comparison is not to assess the quality of the data like in DQM, but the quality of the distribution prediction and in turn to check whether the standard laws are adequate or, instead, new physical laws are needed to model the experimental setup. However, this conceptual difference does not have practical consequences. The NPLM setup of $\mathcal{D}$ versus $\mathcal{R}$ data comparison is straightforwardly portable to DQM problems.

The NPLM method design is inspired by the classical approach to hypothesis testing based on the likelihood ratio [NP33]. A model $f_{\mathbf{w}}(x)$ acting on the space of data x, with trainable parameters $\mathbf{w}$, is employed to define a set of alternatives to $p(x|\mathrm{R})$ for the distribution of the data points in $\mathcal{D}$. Since the alternative hypothesis depends on $\mathbf{w}$, we denote it as $\mathrm{H}_{\mathbf{w}}$ and $p(x|\mathrm{H}_{\mathbf{w}})$ is the alternative distribution of x. In particular, $f_{\mathbf{w}}(x)$ directly parametrises the logarithm of the ratio between $p(x|\mathrm{H}_{\mathbf{w}})$ and $p(x|\mathrm{R})$. The model $f_{\mathbf{w}}(x)$ could be a neural network as in [DW19, DGP$^+$21, dGP$^+$22], or it could be built with kernel methods [LLR$^+$22]. We will employ the latter option for reasons that will become clear soon. The model is trained by adjusting its parameters to best accommodate the observed data. Consequently, the trained parameters $\hat{w}$ define the best-fit hypothesis $H_{\hat{w}}$. Following [NP33], the test statistic variable to be employed for the assessment of the quality of the data $\mathcal{D}$ is [3]

$$t_{\hat{w}}(\mathcal{D}) = 2\sum_{x\in\mathcal{D}} \log\frac{p(x|H_{\hat{w}})}{p(x|\mathrm{R})} = 2\sum_{x\in\mathcal{D}} f_{\hat{w}}(x)\,. \tag{10.2}$$

In order to train the model we exploit a classical result of statistical learning: a continuous-output classifier trained to tell apart two datasets approximates —possibly up to a given monotonic transformation— the log ratio between the probability distribution of the two training sets. This property is proven explicitly in e.g. [DW19, LLR$^+$22] for the weighted logistic loss

$$\ell(y, f_{\mathbf{w}}(x)) = (1-y)(1+N_1/N_0)\log\left(1+e^{f_{\mathbf{w}}(x)}\right) + y\,(1+N_0/N_1)\log\left(1+e^{-f_{\mathbf{w}}(x)}\right)\,. \tag{10.3}$$

By assigning label $y=0$ to the data in $\mathcal{R}$, and $y=1$ to those in $\mathcal{D}$, the model $f_{\hat{w}}(x)$ trained with the loss in Eq. (10.3) approaches the logarithm of $p(x|H_{\hat{w}})/p(x|R)$ as it was needed in Eq. (10.2). The weight factors in Eq. (10.3), which depend on $N_1/N_0 = N_{\mathcal{D}}/N_{\mathcal{R}}$, are included because the two training datasets are unbalanced as previously explained.

A direct application of the classical theory of hypothesis testing [NP33] would actually suggest to employ a different loss function. In fact, the best-fit parameters $\hat{w}$ to be used in the definition

[3] Unlike in NPLM applications to new physics searches, the total number of data points in $\mathcal{D}$ is not a random variable, but is fixed to the data batch size. The regular likelihood for i.i.d. data is thus employed rather than the extended likelihood. Correspondingly, the test statistic contains one term less than in [DW19, DGP$^+$21, dGP$^+$22, LLR$^+$22].

of the test statistic (10.2) should be those that maximise the likelihood function. Minimising the logistic loss produces instead an estimate of the best-fit parameters that is different, a priori, from the maximum likelihood estimate. This can be remedied by employing a special loss function called "maximum likelihood loss", whose minimisation is equivalent to maximising the likelihood [DW19]. The maximum likelihood loss is not used in the kernel-based implementation of NPLM [LLR$^+$22] and the logistic loss (10.3) is preferred for practical reasons. No strong performance degradation has been observed using the logistic loss in place of the maximum likelihood loss in the tests of the NPLM method performed so far.

Using the elements above, the design of the NPLM method for DQM works as follows. We first pick up a model for $f_{\mathbf{w}}(x)$ and select its hyperparameters. The hyperparameters selection strategy is described in the next section for the kernel-based implementation of NPLM. Next, we need to calibrate the test statistics variable (10.2) in order to be able to associate its value $t(\mathcal{D})$ to a probability $\mathrm{p}[t(\mathcal{D})]$, the *p-value*. This probability will be the output of the DQM algorithm. Based on its value, the analyser will eventually judge the quality of each data batch $\mathcal{D}$. For instance, the analyser might define a probability threshold, below which the data batch is discarded or set apart for further analyses. Above the threshold the batch could be retained as a good batch.

It should be noted that the selected hyperparameters and the p-value do depend on the detailed setup of the DQM problem under consideration. For instance, different hyperparameters will be used in Section 10.4 for the setup with 5 input features and data batch size $N_{\mathcal{D}} = 1000$ than in the case of 6 features and $N_{\mathcal{D}} = 500$. The p-value calibration function $\mathrm{p}[t]$ will be also different. However, once these elements are made available for a given setup, they can be used to evaluate the quality of all the $\mathcal{D}$ batches in that setup. The only operation that the DQM algorithm has to perform at run-time is one single training of $\mathcal{D}$ against $\mathcal{R}$, out of which $t(\mathcal{D})$ is obtained and in turn $\mathrm{p}[t(\mathcal{D})]$.

Calibration is performed as follows. The test statistics (10.2) is preferentially large and positive if the best-fit alternative distribution $p(x|H_{\hat{w}})$ accommodates the data better than the reference distribution $p(x|\mathrm{R})$ does, signalling that the data batch is likely not drawn from $p(x|\mathrm{R})$. Large $t(\mathcal{D})$ should thus correspond to a small probability. The precise correspondence is established by comparison with the typical values that t attains when the data batch is instead a good batch. We thus compute the distribution, $p(t|\mathrm{R})$, that the t variable possesses when the data follow the reference statistical distribution and the p-value is defined as

$$\mathrm{p}[t] = \int_t^{\infty} dt'\, p(t'|\mathrm{R})\,. \tag{10.4}$$

The physical meaning of $\mathrm{p}[t(\mathcal{D})]$ is the probability that a good data batch gives a value of t that is more unlikely (i.e., larger) than the value $t(\mathcal{D})$ produced by the batch $\mathcal{D}$. If a threshold is set on p, this threshold measures the frequency at which good data batches are not recognised as such by the algorithm.

The $p(t|\mathrm{R})$ distribution is straightforwardly estimated empirically, thanks to the availability of reference-distributed labelled data points. We create several artificial data batches—called *Toy* datasets— of the same size $N_{\mathcal{D}}$ as the true batches. We run the training and compute t on each of them. Each Toy dataset should be statistically independent, and independent from the reference dataset $\mathcal{R}$ that is employed for training. A very large sample of reference-distributed data is thus used in order to produce both the Toy batches and the reference dataset. By histogramming the values of t computed on the Toys we could easily obtain an estimate of $p(t|R)$ and hence of $\mathrm{p}[t]$. A different procedure is adopted here, exploiting the empirical observation [LLR$^+$22] that $p(t|\mathrm{R})$ is well approximated by a chi-squared (χ^2) distribution. The number of degrees of freedom of the χ^2 depends on the setup but can be determined by fitting to the empirical distribution of the t values computed on the Toys. The survival function (one minus the cumulative) of the corresponding χ^2 distribution will be used as an estimate of $\mathrm{p}[t]$. It should be noted that by proceeding in this way we will be formally able to compute very small p-values that correspond to highly-discrepant data batches with very large $t(\mathcal{D})$. However, the agreement of $p(t|\mathrm{R})$ with the χ^2 cannot be verified in the high t region, which the Toys do not populate, and there is no theoretical reason to expect that this agreement will persist in that region. Our quantification of the p-value is thus only accurate in the region that the Toys statistically populate. For instance, if 300 Toys are thrown, only p-values larger than around $1/300$ are accurately computed. If $t(\mathcal{D})$ falls in a region where our determination of p is much smaller than that, ours should be regarded as a reasonable estimate that is particularly useful to compare the level of discrepancy of different batches, but it cannot be directly validated. However, in those cases we will be able to ensure that $\mathrm{p}[t(\mathcal{D})] \lesssim 1/300$ by directly comparing with the t values on the Toys.

Another feature of the NPLM approach is the possibility of exploiting the function $f_{\hat{w}}$ learned during the training task to characterise anomalous batches of data. The function $f_{\hat{w}}$ represents the log-ratio between $p(x|H_{\hat{w}})$ and $p(x|\mathrm{R})$ and, hence, can be used to deform and adapt the reference distribution to the data by reweighting, according to the following expression

$$p(x|H_{\hat{w}}) = e^{f_{\hat{w}}(x)} p(x|\mathrm{R}). \tag{10.5}$$

The function $\exp(f_{\hat{w}}(x))$ will be close to one if the data are well-described by the reference distribution, while it will depart from it otherwise. One should therefore be able to gain additional information about the anomalous batch by inspecting this quantity as a function of the input variables, or any combination of them, even when not explicitly provided as an input feature for the training. Having access to this kind of information is a valuable element in the context of the search for new physics [DW19, DGP$^+$21, LLR$^+$22], since the physics-motivated variables that one might want to inspect to explain a potential anomalous score could be some type of nontrivial combination of the input features with a clear physical meaning, such as the invariant mass of a many-body final state. For DQM applications, this analysis is less relevant since a direct visual inspection of the ratio between the binned data and reference marginal distributions is already quite informative and the user might not be not interested in exploring specific high-level features in the first place. On the other hand, one can still exploit the possibility of reconstructing the data distribution using $f_{\hat{w}}$ as a debugging tool, namely to check whether the learning model correctly recognises if the data deviates from the reference and how.

Moreover, somewhat aside from the main goal of the present article, the output of the NPLM-DQM application could be exploited to study those data batches that display significant deviations from the reference and, depending on the characteristic of the departures, to classifying them into different anomalous categories. Further investigations on a possible extension of the application in this respect are left for future work.

10.3.2 Falkon-based NPLM

Applying NPLM to the DQM problem is simpler than using it for new physics searches. For new physics searches one needs to worry about imperfections in the reference data that stem from the mismodelling of the reference distribution based on the underlying standard physical laws. Including these effects in NPLM is possible but requires dedicated work and domain-specific expertise [dGP+22]. Mismodelling is not a concern in DQM problems because no modelling is required at all: the reference-distributed data are merely collected from the same experimental apparatus and not simulated. NPLM algorithms for DQM can thus be designed more easily and systematically without the need for extremely specialised domain knowledge.

DQM applications are, however, much more computationally demanding than new physics searches. For new physics searches there is typically only one dataset $\mathcal{D}$ to be analysed. For DQM, a large flow of data batches needs to be analysed online. We will see in Section 10.5 that, for instance, order 10 seconds are needed to the CMS muon system to collect one data batch. Our DQM algorithm must respond on a competitive timescale in order to be applicable to that problem. The relevant operation time is the one needed for a single training, as previously explained. The original implementation of NPLM based on neural networks is vastly incompatible with this requirement. On the other hand, the one based on kernel methods is much faster to train on problems of comparable scale [LLR+22]. It could thus match the specification requirements for applications to LHC detectors.

The performance of the kernel-based version of NPLM stems from those of the Falkon [MCRR20] library, the core algorithm powering our implementation. A sketch of the basic theoretical and algorithmic ideas implemented in Falkon, developed in Ref. [RCR17, MFBR19, MOBR19], are reported below.

With kernel methods, one learns functions of the following form

$$f_{\mathbf{w}}(x) = \sum_{i=1}^{N} w_i k_\sigma(x, x_i)\,, \tag{10.6}$$

with $N = N_0 + N_1$ the total size of the training dataset. Here $k_\sigma(x, x_i)$ is the kernel function and σ some hyperparameter. We consider the Gaussian kernel

$$k_\sigma(x, x') = e^{-\|x-x'\|^2/2\sigma^2}\,, \tag{10.7}$$

so that $f_{\mathbf{w}}$ is a linear combination of Gaussians of fixed width σ, centred at the training data points. The optimisation of the model parameters $\mathbf{w}$ is achieved by minimising the empirical risk $\hat{L}(f_{\mathbf{w}})$, plus a regularisation term

$$\hat{L}_\lambda(f_{\mathbf{w}}) = \hat{L}(f_{\mathbf{w}}) + \lambda R(f_{\mathbf{w}})\,. \tag{10.8}$$

The empirical risk in our case is the one associated with the logistic loss (10.3)

$$\hat{L}(f_{\mathbf{w}}) = \sum_{i=1}^{N} \ell(y_i, f_{\mathbf{w}}(x_i))\,. \tag{10.9}$$

The regularisation term is given by

$$R(f_{\mathbf{w}}) = \sum_{ij} w_i w_j k_\sigma(x_i, x_j)\,. \tag{10.10}$$

Its relative importance in the optimisation target (10.8) is controlled by the hyperparameter λ.

Kernel methods are non-parametric approaches, in the sense that the number of parameters $\mathbf{w}$ in Eq. (10.6) increases automatically with the total number of data points. Gaussian kernel methods are universal, meaning that they can recover any continuous function in the large sample limit [MXZ06, CS08]. However, optimising the function in Eq. (10.6), with the target in Eq. (10.8), requires handling an $N \times N$ matrix—the *kernel matrix*—with entries $k_\sigma(x_i, x_j)$. The computational complexity of the optimisation thus scales cubically in time and quadratically space with respect to the number of training points N [RCR17, MCRR20]. These costs prevent the application to large-scale settings, and some approximation is needed.

Within the Falkon library, the problem of minimising Eq. (10.8) is formulated in terms of an approximate Newton method (see Algorithm 2 of [MCRR20]). The algorithm is based on the Nyström approximation, which is used twice. First, to reduce the size of the problem, by considering solutions of the form

$$f_{\mathbf{w}}(x) = \sum_{i=1}^{M} w_i k_\sigma(x, \tilde{x}_i), \tag{10.11}$$

where $\{\tilde{x}_1, ..., \tilde{x}_M\} \subset \{x_1, ..., x_N\}$ are called Nyström centres and are sampled uniformly at random from the input data. The number of centres $M \leq N$ is a hyperparameter to be chosen. Then, Nyström approximation is again used to derive an approximate Hessian matrix

$$\tilde{\mathbf{H}} = \frac{1}{M} T \tilde{D} T^\intercal + \lambda I. \tag{10.12}$$

Here, T is such that $T^\intercal T = \tilde{K}$ (Cholesky decomposition), with $\tilde{K} \in \mathbb{R}^{M \times M}$ the kernel matrix subsampled with respect to both rows and columns. $\tilde{D} \in \mathbb{R}^{M \times M}$ is a diagonal matrix s.t. the i-th element is the second derivative of the loss $\ell''(y_i, f_{\mathbf{w}}(x_i),)$ with respect to its first variable. Eq. (10.12) is then used as a preconditioner to perform conjugate gradient descent. With this strategy, the overall computational cost to achieve optimal statistical bounds is $\mathcal{O}(N)$ in memory and, of particular importance for our scope, $\mathcal{O}(N\sqrt{N} \log N)$ in time. The reader can find more details in Ref. [MCRR20].

Hyperparameters selection

The selection of the three Falkon hyperparameters M, σ and λ follows the prescriptions of Ref. [LLR$^+$22], with one minor modification described below. The hyperparameters selection employs data collected under the reference working condition, and proceeds as follows.

The number of centres M controls the expressive power of the model and therefore it should be as large as possible not to compromise the sensitivity to anomalous distributions with intricate shapes. It must also be at least as large as $\sqrt{N}$ in order to achieve statistically optimal bounds of the training convergence. At the same time, training is faster if M is smaller. The experiments performed in Ref. [LLR$^+$22] show that any value of M above around the data batch size $N_{\mathcal{D}}$ does not compromise sensitivity.

The Gaussian width σ is selected as the 90th percentile of the pairwise distance between reference-distributed data points. Notice that the model (10.11) acts on an input vector x whose input features are standardised to have zero mean and unit variance on reference-distributed data. The same standardisation is applied before computing the distances.

The regularisation parameter λ is kept as small as possible while keeping training stable, i.e. avoiding large training times or non-numerical outputs. A number of reference-distributed Toy data batches is employed for this study, each trained against the reference sample $\mathcal{R}$. Some of the experiments performed in this paper employ quite smaller data batches (e.g., $N_{\mathcal{D}} = 250$) than those considered in Ref. [LLR+22]. In these new conditions we observe that the compatibility of the test statistic distribution with a χ^2 (see the end of Section 10.3.1) is violated for very small λ. In these cases, we raise λ until when the agreement with the χ^2 is restored.

The hyperparameters selected with the above criteria, in the different setups for DQM considered in this paper, are reported in Table 10.1.

	$N_{\mathcal{R}}$	$N_{\mathcal{D}}$	M	σ	λ	dof
5D	2000	250 500 1000	2000	4.5	10^{-6} 10^{-7} 10^{-8}	40 83 171
6D	2000	250 500 1000	2000	4.8	10^{-6}	58 78 109

Table 10.1: NPLM algorithm parameters configuration for the five-dimensional and six-dimensional experiments considered in this work. The numbers of degrees of freedom of the χ^2 that best approximates $p(t|\mathrm{R})$ is reported in the last column.

10.3.3 Alternative approaches

Goodness-of-fit and two-sample test problems are of interest in several domains of science. Many approaches exist, and developing new strategies is an active area of research. One heuristic reason to choose NPLM for DQM, among the many different options, is that it has been developed in the challenging context of new physics searches. Prior experimental and theoretical knowledge suggests that new physics is elusive. The target for new physics searches is thus to spot out minor departures of the actual data from the reference distribution. These departures could emerge either as small corrections to the distribution shape or as relatively large corrections like sharp peaks, which however only account for a very small fraction of the experimental data. Detecting such small effects requires precisely comparing the reference distribution with large datasets, which NPLM is designed to perform. Using NPLM for DQM could thus enable a more accurate monitoring of the data offering sensitivity to more subtle failures of the apparatus. The number of input features in the data that are typically relevant for new physics searches ranges from few to tens, which is an adequate number also for the monitoring of individual detectors and detector systems fully exploiting the correlations among the variables. For comparison, methods to assess the quality of generated images target instead order thousand-dimensional input data. They could be less performant for DQM as they are designed to address a radically different problem.

These heuristic considerations suggest that NPLM is a reasonable starting point for the development of novel DQM algorithms based on advanced multivariate goodness-of-fit or two-sample test methods, which we advocate in this paper. On the other hand, no comprehensive comparative study of the NPLM performances is currently available. Such comparison is beyond the scope of this paper. However, the DQM problems and datasets we study will be useful benchmarks for future work in this direction.

Work has initiated [CKLW21, GLPW] to compare NPLM with a certain class of methods, called "classifier-based" methods. The classifier-based approaches [Fri03] are all those that entail training a classifier to tell apart $\mathcal{D}$ from $\mathcal{R}$ and using the trained classifier to construct a test statistic for the hypothesis test. A simple implementation [LO17] employs the classification accuracy as test statistics. Following the standard pipeline for classifiers, the model is trained on part of the $\mathcal{D}$ and $\mathcal{R}$ datasets (the training set), while the accuracy is evaluated on the remaining data (the test set). The idea is that while the accuracy will be poor (around random guess) if $\mathcal{D}$ and $\mathcal{R}$ follow the same distribution, it will be higher if their distributions differ.

NPLM is technically a classifier-based method. Its major peculiarities are the choice of the likelihood ratio test statistic in Eq. (10.2) and the fact that the entire datasets are employed both for training and for the evaluation of the test statistics. None of these choices is motivated from the viewpoint of the theory of classification, while they are both natural or in fact required from the perspective of the theory of hypothesis testing that underlies the NPLM approach. Performance studies in [GLPW] show that these choices are beneficial for the sensitivity. These results partly contradict Ref. [CKLW21], which however employs different classification models, different criteria for hyperparameters selection and uses permutation tests for the estimate of the sensitivity rather than computing it empirically as in NPLM. These differences are evidently responsible for the different findings and more work is needed for a conclusive assessment.

10.4 Results

In this section, we present the application of the NPLM strategy for DQM to the DT chambers data described in Section 10.2.[4] We will consider monitoring data batches of variable size $\mathrm{N}_{\mathcal{D}} = 250$, 500 and 1000, by employing a reference dataset of fixed size $\mathrm{N}_{\mathcal{R}} = 2000$.

The input data consists of six features: the four drift times, the muon angle and the number of hits. As shown in the bottom-right plots of Figures 10.3 and 10.4, the number of hits, n_{Hits}, is highly discriminant for the anomalies we considered in our study, and in particular for the ones affecting the thresholds (the lower the threshold, the higher the noise). At the LHC, however, that quantity also depends on the luminosity delivered to the experiment, which could vary greatly even during a single run. Not being necessarily a proxy to a detector issue, it is worth considering also the case where only the other five variables are provided to the algorithm; as an additional benefit, this will allow assessing the ability of the NPLM DQM approach to exploit correlations between variables and detect anomalies even when their effect is unexpected and not straightforwardly evident.

The left and middle panels of Figure 10.5 show the test statistics distribution in the five-dimensional problem, for data batches size $\mathrm{N}_{\mathcal{D}} = 500$. The grey histograms display the distribution of t in the reference working conditions, $P(t|\mathrm{R})$. This is obtained empirically by processing reference-distributed Toy data batches, and fitted to a χ^2 distribution as explained in Section 10.3.1. The different distributions of the test statistic associated with the anomalous batches shown in the coloured histograms are very well separated from the reference distribution, meaning that anomalous data are very likely to be identified as such by the algorithm. This is quantified by the median p-value of the anomalous batches, reported in the central column of Table 10.2. The table also reports the median p-value for larger ($\mathrm{N}_{\mathcal{D}} = 1000$) and smaller ($\mathrm{N}_{\mathcal{D}} = 250$) batches. The sensitivity to the anomaly increases with $\mathrm{N}_{\mathcal{D}}$, as expected.

[4] Code available at `https://github.com/FalkonHEP`.

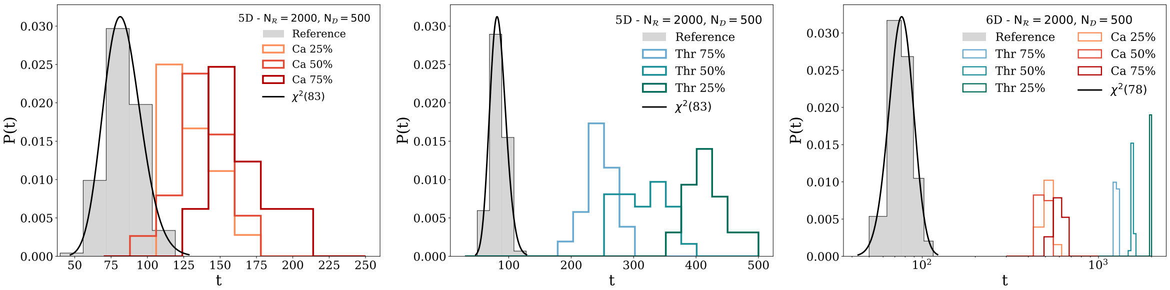

Figure 10.5: Distribution of the test statistics in the scenario $N_{\mathcal{D}} = 500$. The plot displays the distribution of the test statistic t on reference-distributed Toys and on the data collected under anomalous detector conditions.

Anomaly	$N_{\mathcal{D}} = 250$	$N_{\mathcal{D}} = 500$	$N_{\mathcal{D}} = 1000$
Cathode 75%	0.0034	1.1×10^{-6}	$< 10^{-7}$
Cathode 50%	0.029	3.4×10^{-4}	$< 10^{-7}$
Cathode 25%	0.14	0.0019	$< 10^{-7}$
Threshold 75%	2.8×10^{-7}	$< 10^{-7}$	$< 10^{-7}$
Threshold 50%	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$
Threshold 25%	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$

Table 10.2: Median p-values for different anomalies and data batches size. Five input features are considered, excluding n_{hits}.

For a comparative assessment of the performance, we computed a Kolmogorov–Smirnov (KS) test on each individual feature for the same data used to train the NPLM model. The KS median p-values are reported in Table 10.3 and compared with the ones obtained with the five-dimensional NPLM test. We see that individual variables have a very limited power to discriminate the anomalous batches. The NPLM method instead is sensitive to correlated discrepancies in the different distributions and discriminates the anomalies effectively. For illustrative purposes, we show in the left and middle panels of Figure 10.6 the distribution of the one-dimensional KS statistic computed on the drift time of the first layer (t_1) for reference and anomalous batches. By comparison with Figure 10.5, it is easy to recognise the advantage of the NPLM strategy.

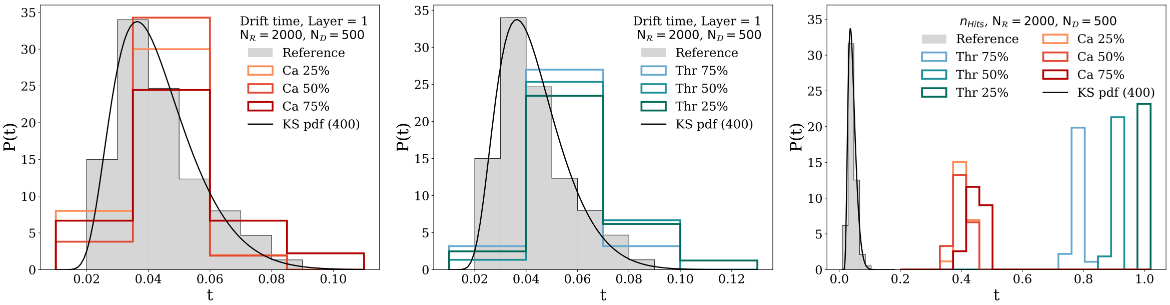

Figure 10.6: Distribution of the test statistic for the KS test.

We now turn to the study of the complete six-dimensional problem, including the variable

Anomaly	NPLM (5D)	KS (t_1)	KS (t_2)	KS (t_3)	KS (t_4)	KS (ϕ)
Cathode 75%	1.1×10^{-6}	0.50	0.41	0.43	0.40	0.42
Cathode 50%	3.4×10^{-4}	0.47	0.27	0.47	0.37	0.41
Cathode 25%	0.0019	0.45	0.44	0.13	0.45	0.50
Threshold 75%	$< 10^{-7}$	0.23	0.14	0.16	0.14	0.48
Threshold 50%	$< 10^{-7}$	0.09	0.10	0.06	0.17	0.42
Threshold 25%	$< 10^{-7}$	0.11	0.07	0.04	0.11	0.66

Table 10.3: Median p-values in the setup $\mathrm{N}_{\mathcal{D}} = 500$.

n_{hits}. The reference and anomalous test statistic distributions are shown on the right panel of Figure 10.5. By comparing with the other panels of the figure we can appreciate the tremendous discriminating power of the n_{hits} variable: including n_{hits} all the anomalies can be detected with very high significance. Therefore, using this variable alone for the NPLM DQM test, or running a regular KS test (as shown in the right panel of Figure 10.6), is sufficient to identify the anomalies, as previously mentioned.

We conclude this section by showing some examples of the data marginal distribution reconstructed by the model. The three plots reported in Figure 10.7 are produced by reweighting each event of the reference sample used for the training by an exponential factor $e^{f_{\hat{w}}(x)}$, as explained in Eq. 10.5; both the reweighted reference and the data samples are binned, and their ratio with respect to the original reference sample is shown in the bottom panels. By comparing the data-versus-reference ratio (labelled as "true") with the reconstructed one ("learned") we can appreciate the correctness of the model in understanding the nature of the anomaly and, hence, trust the results of the machine learning task.

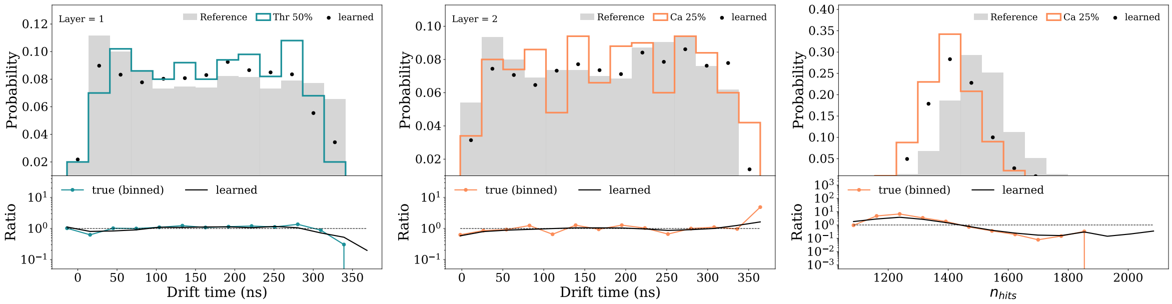

Figure 10.7: Examples of input data and respective learned likelihood ratios with sample size $N_R = 2000$ and $N_D = 500$.

All the numerical experiments presented in this paper have been performed on a single machine equipped with a NVIDIA Titan Xp GPU with 12 GB of VRAM. We tested the performances of the algorithm in terms of execution time; the training time for a single five-dimensional classification task is approximately 0.5 seconds, with no significant dependency on the nature of the data and the size of the sample.

10.5 Conclusions and outlook

We presented the test of a powerful ML-based algorithm, NPLM, as a tool to monitor the quality of the data originated by a typical detector used for measuring particles at high energy colliders. NPLM compares collected measurements with a reference dataset describing the standard detector readout, performing a multidimensional likelihood-ratio hypothesis test.

The study demonstrated the capability of the algorithm to detect anomalous detector conditions, with a much greater discriminating power than simpler traditional methods, like Kolmogorov–Smirnov test.

Although conducted on simplified experimental conditions, the test presents figures appropriate for a typical monitoring system of a detector operating at the LHC; in particular, the number of channels and the size of the datasets are of the same order of magnitude as the corresponding CMS DQM application. The amount of data we consider for each batch can be gathered much more quickly at the LHC than in a cosmic stand like the one used here, anyhow the rate at which possible issues should be detected is not larger than one in a minute[5]; the time requested by NPLM to run —less than a second— makes the algorithms suitable to be executed online.

Acknowledgements: L.R., M.L. and M.R. acknowledge the financial support of the European Research Council (grant SLING 819789). L.R. acknowledges the financial support of the AFOSR project FA9550-18-1-7009 and the EU H2020-MSCA-RISE project NoMADS - DLV-777826. G.G. is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no 772369). A.W. is supported by the grant PID2020-115845GB-I00/AEI/10.13039/501100011033.

[5] Failures potentially leading to catastrophic consequences that requires a much prompt reaction are typically controlled by hardware interlock systems

Chapter 11

Conclusions

In this thesis, we addressed both black-box optimization problems and practical machine-learning applications by designing, analyzing, and implementing various solutions.

In the first part of the manuscript, we tackled the black-box optimization problem facing the challenges of the structured finite-difference approach and Bayesian Optimization framework. To this aims, we introduced two algorithms: S-SZD (Stochastic Structured Zeroth-order Descent) and O-ZD (Orthogonal Zeroth-order Descent). Through the analysis of these algorithms, we derived convergence rates in the smooth stochastic and non-smooth settings, respectively.

For S-SZD, we observed that the convergence rate in the stochastic convex case approaches $1/\sqrt{k}$, analogous to SGD in the same setting. Additionally, for the λ-smooth stochastic non-convex γ-PL setting, we derived a convergence rate close to $1/k$, matching the rate of SGD in the strongly convex case. However, the convergence rate deteriorates for the limit choice of the stepsize $\alpha_k = 1/k$, depending on the function's condition number γ/λ and the ratio d/ℓ, where d is the dimension of the input space and ℓ is the number of directions. A similar behavior can be observed in SGD applied to strongly convex functions. Empirical comparisons with a variety of zeroth-order methods suggest that our algorithm outperforms direct search methods in different settings. With O-ZD, we addressed the non-smooth setting by introducing the first Smoothing Lemma for structured gradient approximation. We derived convergence rates for convex and non-convex functions in both smooth and non-smooth settings. For convex non-smooth functions, we achieved a convergence rate similar to the subgradient method in terms of iterations k. The complexity depends on the choice of stepsize, and with an appropriate choice, our method achieves optimal dependence on the dimension. Similarly, for non-convex non-smooth functions, we provided rates on the expected norm of the smoothed gradient and assessed approximate stationarity through the concept of Goldstein stationarity. For convex smooth settings, our method achieved a rate of $\mathcal{O}(1/k)$, matching Gradient Descent, with optimal dependence on dimension in complexity. Similarly, for smooth non-convex functions, we derived a rate of $\mathcal{O}(1/k)$ with complexity $\mathcal{O}(d\varepsilon^{-1})$. These works open up several research directions, including the development of adaptive strategies to choose direction matrices along iterations, and extending the algorithms to include inertia or variance reduction.

Next, we addressed the scalability limitations of Bayesian Optimization on continuous domains by introducing Ada-BKB, an algorithm that combines ideas from BKB [CCL$^+$19] and optimistic optimization. The proposed approach is analyzed theoretically in terms of regret guarantees, demonstrating improved efficiency without sacrificing accuracy. Through computational cost analysis, we observed that Ada-BKB has the smallest provable computational complexity among methods with adaptive discretization capable of handling noisy observation cases. Furthermore, note that the GP-UCB algorithm (explained in Chapter 2) has a computational cost of $\mathcal{O}(T^3A)$ with A representing the size of the discretization of the search space X, and it has been proven in [SKKS12] that achieving low regret requires the cardinality of the discretization to grow exponentially with the dimension of X. Moreover, our algorithm is faster than other scalable methods, including BKB ($\mathcal{O}(TAd_{\text{eff}}^2)$) and TS-QFF ($\tilde{\mathcal{O}}(TA2^p d_{\text{eff}})$) [MK19], in the same setting. Empirically, we observed very good performances on both simulated data and a hyperparameter tuning task. This work opens up several research directions. For example, efficiency could be further improved using experimentation batching, as discussed in [CCL$^+$20]. Another interesting question could be to extend the ideas presented in Chapter 6 to explore alternative methods for defining upper function estimates, such as those based on expected improvements [QKR17].

Then, in the second part of the manuscript, we tackled the applicative challenges proposing and implementing different methods.

We tackled the plankton monitoring problem by introducing an efficient unsupervised learning pipeline for plankton image clustering. Our approach is composed of three steps. In the first step, input images are pre-processed and fed to a neural network (DenseNet201 in our experiments) pre-trained on ImageNet, without fine-tuning. In the second step, the set of features extracted are used as inputs to train an variational encoder-decoder neural network and the resulting latent space representations of the inputs are used as a lower dimensional set of embedded features. In the third step, the embedded features are passed to a clustering algorithm (a fuzzy k-means in our experiments). We showed that our approach outperforms state-of-the-art unsupervised learning methods [PZBB20] where hand-crafted features are engineered and used for clustering. Furthermore, we empirically proved the high quality of the embedded features produced by our pipeline using a supervised classification framework (in terms of test accuracy). Precisely, we showed that our embedded features coupled to a ridge regression classifier outperforms state-of-the-art classifiers where hand-crafted features are used as input for SVM [SO07, ZWY$^+$17], fully connected neural networks and random forests [PZBB20]. As a further development, the implementation of an end-to-end solution would be crucial for an easy deployment in real-life scenarios. Additionally, it would be interesting and useful to test the approach for anomaly detection. Moreover, since our pipeline is general with respect to the source of input data, it could be interesting to perform a complete analysis of the performances on other kind of data. These aspects are currently under study.

For the olfactory navigation problem, we demonstrated that agents exposed to a turbulent odor plume learn to associate key features of the odor time trace (the olfactory state) with optimal moves that guide them towards the odor source. By responding solely to odor cues, the agent operates without a spatial map or prior information about the odor plume, thereby avoiding significant computational burden. However, in our stimulus-response algorithm, agents must start from within the plume, even if it is sparse and fragmented. When far enough from the source, Q-learning agents primarily exist in a 'void' state, and they can only recover the plume if they previously detected the odor or are right outside the plume. In contrast, agents using a map of space can navigate from larger distances than those reachable by responding directly to odor cues. In a map-based Partially Observable Markov Decision Process (POMDP) setting, the absence of odor detection is still informative, enabling agents to first find the plume and then refine the search to localize the target within the plume [RRSV22]. We addressed the challenge of handling both the absence and presence of odor stimuli by alternating between two strategies:

(i) The prolonged absence of odor triggers entry into the void state, prompting a recovery strategy to make contact with the plume again. We explored two heuristic recovery methods and found that back-tracking to the last odor detection point is more efficient than Brownian recovery. An even more efficient recovery mimics cross-wind casting, limiting the void state to a narrow region just outside the plume. Casting, a well-studied computational strategy [BI02], is also observed in animal behavior, notably in flying insects [DKL83].

(ii) Odor detections prompt entry into non-void olfactory states, primarily resulting in upwind surges. Short blanks in odor detection, typical of turbulence, are ignored, allowing agents to respond to stimuli experienced prior to the blank.

Further optimization of these non-void olfactory states may involve feature engineering, such as testing different discretizations to reduce redundancy or screening a large feature library using supervised learning methods [RMRS22]. Alternatively, recurrent neural networks (RNNs) could bypass feature engineering altogether, as proposed in [SvBRB23], potentially sacrificing interpretability. A systematic comparison using a common dataset is necessary to understand how other heuristic and normative model-free algorithms handle odor presence versus absence.

We addressed the new physiscs learning problem by presenting a machine-learning approach for model-independent searches using kernel-based machine-learning models. Our approach is powered by Falkon, a recent library developed for large-scale applications of kernel methods, and builds on the original ideas of [DW19, DGP$^+$21]. The main focus of our contribution in this field is computational efficiency. The original neural network proposal suffers from long training times, which, combined with a toy-based hypothesis testing framework, makes the algorithm challenging to use in high-dimensional cases. In contrast, our model delivers comparable performance with a dramatic reduction in training times. As a consequence, the model can be efficiently trained on single GPU machines while possessing high scalability for multi-GPU systems [MCRR20]. However, similar to [DGP$^+$21], the applicability of our proposed method relies on a heuristic procedure to tune the algorithm hyperparameters. A deeper understanding of the interplay between the expressibility of the model, its complexity, and the structure of the input dataset could lead to more performant alternatives for hyperparameter selection. A possible research direction would be to find a more principled way to relate Falkon hyperparameters to physical quantities. This could also allow the introduction of explicit quantities to be optimized, opening the possibility of applying modern optimization techniques for hyperparameter selection. Besides the challenges related to optimization and regularization, an essential development for the application of realistic data analysis concerns the treatment of systematic uncertainties, which has not been considered in the present work. This aspect was successfully addressed in a recent work [dGP$^+$22] in the context of the neural network implementation. Finally, the boost in efficiency provided by the model developed in this part of the work could extend the applicability of this analysis strategy to other use cases beyond the search for new physics and to other domains

For data quality monitoring application, we developed and tested a robust machine-learning-based algorithm to monitor the quality of the data originated by a typical detector used in high-energy collider experiments. Our method compares collected measurements with a reference dataset that describes the standard detector readout, employing a multidimensional likelihood-ratio hypothesis test. The study demonstrated the algorithm's effectiveness in detecting anomalous detector conditions, showcasing significantly greater discriminating power compared to simpler traditional methods, such as the Kolmogorov-Smirnov test. While conducted under simplified experimental conditions, the test's parameters align well with those of a typical detector monitoring system at the LHC. Although the data batch sizes considered in our study can be collected more rapidly at the LHC compared to a cosmic stand like the one used here, the rate at which potential issues should be detected remains relatively low (not exceeding one per minute). However, our approach's quick execution time (less than a second) renders it suitable for online execution.

Bibliography

[A$^+$00] B. Abbott et al. Search for new physics in eμX data at DØ using SLEUTH: A quasi-model-independent search strategy for new physics. *Phys. Rev. D*, 62:092004, 2000. arXiv:hep-ex/0006011.

[A$^+$01] V. M. Abazov et al. A Quasi model independent search for new physics at large transverse momentum. *Phys. Rev. D*, 64:012004, 2001. arXiv:hep-ex/0011067.

[A$^+$04] A. Aktas et al. A General search for new phenomena in ep scattering at HERA. *Phys. Lett. B*, 602:14–30, 2004. arXiv:hep-ex/0408044.

[A$^+$08] T. Aaltonen et al. Model-Independent and Quasi-Model-Independent Search for New Physics at CDF. *Phys. Rev. D*, 78:012002, 2008. arXiv:0712.1311 [hep-ex].

[A$^+$09a] T. Aaltonen et al. Global Search for New Physics with 2.0 fb^{-1} at CDF. *Phys. Rev. D*, 79:011101, 2009. arXiv:0809.3781 [hep-ex].

[A$^+$09b] F. D. Aaron et al. A General Search for New Phenomena at HERA. *Phys. Lett. B*, 674:257–268, 2009. arXiv:0901.0507 [hep-ex].

[A$^+$19] Morad Aaboud et al. A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment. *Eur. Phys. J. C*, 79(2):120, 2019. arXiv:1807.07447 [hep-ex].

[A$^+$20] Georges Aad et al. ATLAS data quality operations and performance for 2015–2018 data-taking. *JINST*, 15(04):P04003, 2020.

[A$^+$22] Thea Aarrestad et al. The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider. *SciPost Phys.*, 12(1):043, 2022. arXiv:2105.14027 [hep-ph].

[AAB$^+$15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[AAB$^+$17] M Adinolfi, F Archilli, W Baldini, A Baranov, D Derkach, A Panin, A Pearce, and A Ustyuzhanin. LHCb data quality monitoring. *Journal of Physics: Conference Series*, 898(9):092027, oct 2017.

[AAC$^+$19] Virginia Azzolini, Michael Andrews, Gianluca Cerminara, Nabarun Dev, Colin Jessop, Nancy Marinelli, Tanmay Mudholkar, Maurizio Pierini, Adrian Pol, and Jean-Roch Vlimant. Improving data quality monitoring via a partnership of technologies and resources between the CMS experiment at CERN and industry. *EPJ Web Conf.*, 214:01007, 2019.

[ABD$^+$17] Pouya Asadi, Matthew R. Buckley, Anthony DiFranzo, Angelo Monteux, and David Shih. Digging Deeper for New Physics in the LHC Data. *JHEP*, 11:194, 2017. arXiv:1707.05783 [hep-ph].

[ADX10] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Colt*, pages 28–40. Citeseer, 2010.

[AF01] Edward J. Anderson and Michael C. Ferris. A direct search algorithm for optimization with noisy function evaluations. *SIAM Journal on Optimization*, 11(3):837–857, 2001.

[Alt16] Hans Wilhelm Alt. *Linear Functional Analysis: An Application-Oriented Introduction*. Springer London, London, 2016.

[And14] Sigrún Andradóttir. A review of random search methods. *Handbook of Simulation Optimization*, pages 277–292, 2014.

[ANS20] Anders Andreassen, Benjamin Nachman, and David Shih. Simulation Assisted Likelihood-free Anomaly Detection. *Phys. Rev. D*, 101(9):095004, 2020. arXiv:2001.05001 [hep-ph].

[AOU87] T. W. Anderson, I. Olkin, and L. G. Underhill. Generation of random orthogonal matrices. *SIAM Journal on Scientific and Statistical Computing*, 8(4):625–629, 1987.

[ARL$^+$22] Paolo Didier Alfano, Marco Rando, Marco Letizia, Francesca Odone, Lorenzo Rosasco, and Vito Paolo Pastore. Efficient unsupervised learning for plankton images. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1314–1321. IEEE, 2022.

[AS21] Oz Amram and Cristina Mantilla Suarez. Tag N' Train: a technique to train improved classifiers on unlabeled data. *JHEP*, 01:153, 2021. arXiv:2002.12376 [hep-ph].

[Ate96] J Atema. Eddy chemotaxis and odor landscapes: exploration of nature with animal sensors. *Biol. Bull.*, 191:129, 1996.

[ATL14] A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV. Technical report, CERN, Geneva, Mar 2014. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2014-006.

[AvBB$^+$19] Virginia Azzolini, Broen van Besien, Dmitrijus Bugelskis, Tomas Hreus, Kaori Maeshima, Javier Fernandez Menendez, Antanas Norkus, James Fraser Patrick, Marco Rovere, and Marcel Andre Schneider. The Data Quality Monitoring software for the CMS experiment at the LHC: past, present and future. *EPJ Web Conf.*, 214:02003. 8 p, 2019.

[Bac13] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on learning theory*, pages 185–209. PMLR, 2013. arXiv:1208.2015 [cs.LG].

[Bar05] A. Barvinok. Approximating orthogonal matrices by permutation matrices. *Pure and applied mathematics quarterly*, 2, 11 2005.

[BBBK11] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[BCCS22] Albert S Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, 22(2):507–560, 2022.

[BDKS21] Blaž Bortolato, Barry M. Dillon, Jernej F. Kamenik, and Aleks Smolkovič. Bump Hunting in Latent Space. 3 2021. arXiv:2103.06595 [hep-ph].

[BDLM09] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity. *Transactions of The American Mathematical Society - TRANS AMER MATH SOC*, 362:3319–3363, 06 2009.

[Ber73] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, Aug 1973.

[Ber75] H C Berg. Chemotaxis in bacteria. *Annual Review of Biophysics and Bioengineering*, 4(1):119–136, 1975.

[Ber97] D. P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, Mar 1997.

[Bes10] Michael J Best. *Portfolio optimization*. CRC Press, 2010.

[BG13] C. Boutsidis and A. Gittens. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1301–1340, 2013.

[BG18] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. *Advances in Neural Information Processing Systems*, 31, 2018.

[BGC$^+$07] Mark C Benfield, Philippe Grosjean, Phil F Culverhouse, Xabier Irigoien, Michael E Sieracki, Angel Lopez-Urrutia, Hans G Dam, Qiao Hu, Cabell S Davis, Allen Hansen, et al. Rapid: research on automated plankton identification. *Oceanography*, 20(2):172–187, 2007.

[BGR20] El Houcine Bergou, Eduard Gorbunov, and Peter Richtárik. Stochastic three points method for unconstrained smooth minimization. *SIAM Journal on Optimization*, 30(4):2726–2749, 2020.

[BH77] J.B. Baillon and G. Haddad. Quelques propriétés des opérateurs angle-bornés etn-cycliquement monotones. *Israel Journal of Mathematics*, 26(2):137–150, Jun 1977.

[BHM$^+$05] Matthew B. Blaschko, Gary Holness, Marwan A. Mattar, Dimitri Lisin, Paul E. Utgoff, Allen R. Hanson, Howard Schultz, Edward M. Riseman, Michael E. Sieracki, William M. Balch, and Ben Tupper. Automatic in situ identification of plankton. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, volume 1, pages 79–86, 2005.

[BI02] E Balkovsky and Shraiman B. I. Olfactory search at high reynolds number. *Proc Nat Acad Sci*, 99(20):12589–93, 2002.

[Bil17] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

[Bjo94] A. Bjorck. Numerics of gram-schmidt orthogonalization. *Linear Algebra and its Applications*, 197-198:297–316, 1994.

[BK10] Rémi Bardenet and Balázs Kégl. Surrogating the surrogate: accelerating gaussian-process-based global optimization with a mixture cross-entropy algorithm. In *27th International Conference on Machine Learning (ICML 2010)*, pages 55–62. Omnipress, 2010.

[BLB$^+$13] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[BLW10] Daniel Boyce, Marlon Lewis, and Boris Worm. Global phytoplankton decline over the past century. *Nature*, 466:591–6, 07 2010.

[BMP23] Pourya Behmandpoor, Marc Moonen, and Panagiotis Patrinos. Zeroth-order asynchronous learning with bounded delays with a use-case in resource allocation in communication networks. *arXiv preprint arXiv:2311.04604*, 2023.

[BMSS11] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

[BRM$^+$01] Michael J Behrenfeld, James T Randerson, Charles R McClain, Gene C Feldman, Sietse O Los, Compton J Tucker, Paul G Falkowski, Christopher B Field, Robert Frouin, Wayne E Esaias, et al. Biospheric primary production during an enso transition. *Science*, 291(5513):2594–2597, 2001.

[BRVDW19] David Burt, Carl Edward Rasmussen, and Mark Van Der Wilk. Rates of convergence for sparse variational Gaussian process regression. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 862–871. PMLR, 09–15 Jun 2019.

[BSW14] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Commun.*, 5:4308, 2014. arXiv:1402.4735 [hep-ph].

[BSW19] Andrew Blance, Michael Spannowsky, and Philip Waite. Adversarially-trained autoencoders for robust unsupervised new physics searches. *JHEP*, 10:047, 2019. arXiv:1905.10384 [hep-ph].

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(4):231–357, January 2015.

[BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BZM$^+$19] Sujoy Kumar Biswas, Thomas Zimmerman, Lucrezia Maini, Aminat Adebiyi, Luisa Bozano, Cecelia Brown, Vito Paolo Pastore, and Simone Bianco. High throughput analysis of plankton morphology and dynamic. In *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVII*, volume 10881, page 1088109. International Society for Optics and Photonics, 2019.

[C$^+$08] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[CAB$^+$20] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat K. Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey, 2020.

[CALM$^+$20] Taoli Cheng, Jean-Franacois Arguin, Julien Leissner-Martin, Jacinthe Pilette, and Tobias Golling. Variational Autoencoders for Anomalous Jet Tagging. 7 2020. arXiv:2007.01850 [hep-ph].

[CCGV11] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C*, 71:1554, 2011. [Erratum: Eur.Phys.J.C 73, 2501 (2013)]. arXiv:1007.1727 [physics.data-an].

[CCL$^+$19] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, pages 533–557. PMLR, 2019.

[CCL$^+$20] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Near-linear time gaussian process optimization with adaptive batching and resparsification. In *International Conference on Machine Learning*, pages 1295–1305. PMLR, 2020.

[CCL$^+$22] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Scaling Gaussian process optimization by evaluating a few unique candidates multiple times. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2523–2541. PMLR, 17–23 Jul 2022.

[CES$^+$94] PF Culverhouse, R Ellis, RG Simpson, R Williams, RW Pierce, and JT Turner. Automatic categorisation of five species of cymatocylis (protozoa, tintinnida) by artificial neural network. *Marine Ecology Progress Series*, pages 273–280, 1994.

[Chi03] Yasuko Chikuse. *Statistics on special manifolds*, volume 174. Springer Science & Business Media, 2003.

[Chi10] Rui Chibante. *Simulated annealing: theory with applications*. BoD–Books on Demand, 2010.

[CHN18] Jack H. Collins, Kiel Howe, and Benjamin Nachman. Anomaly Detection for Resonant New Physics with Machine Learning. *Phys. Rev. Lett.*, 121(24):241803, 2018. arXiv:1805.02664 [hep-ph].

[CHN19] Jack H. Collins, Kiel Howe, and Benjamin Nachman. Extending the search for new resonances with machine learning. *Phys. Rev. D*, 99(1):014038, 2019. arXiv:1902.02634 [hep-ph].

[Cho11] Georgios Choudalakis. On hypothesis testing, trials factor, hypertests and the BumpHunter. In *PHYSTAT 2011*, 1 2011. arXiv:1101.0390 [physics.data-an].

[Chu54] K. L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.

[CKLW21] Purvasha Chakravarti, Mikael Kuusela, Jing Lei, and Larry Wasserman. Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests. 2 2021.

[Cla90] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

[CLMY21] Hanqin Cai, Yuchen Lou, Daniel Mckenzie, and Wotao Yin. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1193–1203. PMLR, 18–24 Jul 2021.

[CM06] Bishop CM. Pattern recognition and machine learning,‖ springer, 2006.

[CMYZ22] H.Q. Cai, D. McKenzie, W. Yin, and Z. Zhang. Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. *SIAM Journal on Optimization*, 32(2):687–714, 2022.

[CNP$^+$19] Olmo Cerri, Thong Q. Nguyen, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant. Variational Autoencoders for New Physics Mining at the Large Hadron Collider. *JHEP*, 05:036, 2019. arXiv:1811.10276 [hep-ex].

[Col21] CMS Collaboration. Music: a model-unspecific search for new physics in proton–proton collisions at. *Eur. Phys. J. C*, 81:629, 2021.

[Cou] Robert D. Cousins. On goodness-of-fit tests. `https://www.physics.ucla.edu/~cousins/stats/ongoodness6march2016.pdf`.

[CP15] Patrick L. Combettes and Jean-Christophe Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

[CR18] Daniele Calandriello and Lorenzo Rosasco. Statistical and computational trade-offs in kernel k-means. *Advances in neural information processing systems*, 31, 2018. arXiv:1908.10284 [stat.ML].

[CRCW19] K. Choromanski, M. Rowland, W. Chen, and A. Weller. Unifying orthogonal monte carlo methods. In *International Conference on Machine Learning*, pages 1203–1212. PMLR, 2019.

[CRS$^+$18] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 970–978. PMLR, 10–15 Jul 2018.

[CS08] Andreas Christmann and Ingo Steinwart. *Support vector machines*. Springer, 2008.

[CSV09] Andrew R. Conn, Katya Scheinberg, and Luís Nunes Vicente. Introduction to derivative-free optimization. In *MPS-SIAM series on optimization*, 2009.

[CV16] Emile Contal and Nicolas Vayatis. Stochastic process bandits: Upper confidence bounds algorithms via generic chaining. *arXiv preprint arXiv:1602.04976*, 2016.

[CVV14] A. Celani, E. Villermaux, and M. Vergassola. Odor landscapes in turbulent environments. *Phys. Rev. X*, 4:041015, 2014.

[CW15] Ruobing Chen and Stefan Wild. Randomized Derivative-Free Optimization of Noisy Convex Functions. *arXiv e-prints*, page arXiv:1507.03332, July 2015.

[DBW12] John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[DDL$^+$22] D. Davis, D. Drusvyatskiy, Y. T. Lee, S. Padmanabhan, and G. Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6692–6703. Curran Associates, Inc., 2022.

[DDS$^+$09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DFK16] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks, 2016.

[DFKS20] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc. Learning the latent structure of collider events. *JHEP*, 10:206, 2020. arXiv:2005.12319 [hep-ph].

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[DGP$^+$21] Raffaele Tito D'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning multivariate new physics. *Eur. Phys. J. C*, 81(1):89, 2021.

[dGP$^+$22] Raffaele Tito d'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning new physics from an imperfect machine. *Eur. Phys. J. C*, 82(3):275, 2022.

[DJWW15] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[DKA$^+$20] Mahmut Demir, Nirag Kadakia, Hope D Anderson, Damon A Clark, and Thierry Emonet. Walking *Drosophila* navigate complex plumes using stochastic decisions biased by the timing of odor encounters. *eLife*, 9:e57524, 2020.

[DKL83] C. T. David, J. S. Kennedy, and A. R. Ludlow. Finding of a sex pheromone source by gypsy moths released in the field. *Nature*, 303:804–806, 1983.

[DMC05] Petros Drineas, Michael W Mahoney, and Nello Cristianini. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(12):2153–2175, 2005.

[DSJ19] Andrea De Simone and Thomas Jacques. Guiding New Physics Searches with Unsupervised Learning. *Eur. Phys. J. C*, 79(4):289, 2019. arXiv:1807.06038 [hep-ph].

[dT21] Ian du Toit. Enhanced Deep Learning Feature Extraction for Plankton Taxonomy. In *Proceedings of the International Conference on Artificial Intelligence and its Applications*, number 7, pages 1–8. Association for Computing Machinery, New York, NY, USA, 2021.

[Dun73] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[DV16] M. Dodangeh and L. N. Vicente. Worst case complexity of direct search under convexity. *Mathematical Programming*, 155(1):307–332, Jan 2016.

[DVZ16] M. Dodangeh, L. N. Vicente, and Zaikun Zhang. On the optimal order of worst case complexity of direct search. *Optimization Letters*, 10(4):699–708, April 2016.

[DW19] Raffaele Tito D'Agnolo and Andrea Wulzer. Learning New Physics from a Machine. *Phys. Rev. D*, 99(1):015014, 2019.

[DWZ+16] Jialun Dai, Ruchen Wang, Haiyong Zheng, Guangrong Ji, and Xiaoyan Qiao. Zooplanktonet: Deep convolutional network for zooplankton classification. In *OCEANS 2016 - Shanghai*, pages 1–6, 2016.

[EFG07] Brochu Eric, Nando Freitas, and Abhijeet Ghosh. Active preference learning with discrete choice data. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[EGO19] Jeffrey S Ellen, Casey A Graff, and Mark D Ohman. Improving plankton image classification using context metadata. *Limnology and Oceanography: Methods*, 17(8):439–461, 2019.

[EJ04] Ariane S. Etienne and Kathryn J. Jeffery. Path integration in mammals. *Hippocampus*, 14(2):180–192, 2004.

[Elk01] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[EMS96] Ariane S. Etienne, Roland Maurer, and Valérie Séguinot. Path Integration in Mammals and its Interaction With Visual Landmarks. *Journal of Experimental Biology*, 199(1):201–209, 01 1996.

[FFD+19] Trygve O. Fossum, Glaucia M. Fragoso, Emlyn J. Davies, Jenny E. Ullgren, Renato Mendes, Geir Johnsen, Ingrid Ellingsen, Jo Eidsvik, Martin Ludvigsen, and Kanna Rajan. Toward adaptive robotic sampling of phytoplankton in the coastal ocean. *Science Robotics*, 4(27):eaav3041, 2019.

[FGV01] G. Falkovich, K. Gawedzki, and M. Vergassola. Particles and fields in fluid turbulence. *Rev. Mod. Phys.*, 73:913, 2001.

[FKM05] A. Flaxman, A. Tauman Kalai, and B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *SODA '05 Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, January 2005.

[FKM+21] Thorben Finke, Michael Krämer, Alessandro Morandini, Alexander Mück, and Ivan Oleksiyuk. Autoencoders for unsupervised anomaly detection in high energy physics. *JHEP*, 06:161, 2021. arXiv:2104.09051 [hep-ph].

[FKR21] M. Fornasier, T. Klock, and K. Riedl. Consensus-based optimization methods converge globally in mean-field law. *arXiv 2103.15130*, 2021.

[FNS20] Marco Farina, Yuichiro Nakai, and David Shih. Searching for New Physics with Deep Autoencoders. *Phys. Rev. D*, 101(7):075021, 2020. arXiv:1808.08992 [hep-ph].

[Fra18a] P. I. Frazier. *Bayesian Optimization*, chapter 11, pages 255–278. INFORMS, 2018.

[Fra18b] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[Fri03] Jerome H. Friedman. On multivariate goodness of fit and two sample testing. *eConf*, C030908:THPD002, 2003.

[FW23] Yasong Feng and Tianyu Wang. Stochastic zeroth-order gradient and hessian estimators: variance reduction and refined bias bounds. *Information and Inference: A Journal of the IMA*, 12(3):iaad014, 2023.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[Gen00] A. Genz. Methods for generating random orthogonal matrices. In *Monte-Carlo and Quasi-Monte Carlo Methods 1998: Proceedings of a Conference held at the Claremont Graduate University, Claremont, California, USA, June 22–26, 1998*, pages 199–213. Springer, 2000.

[GG23] Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

[GJZ18] X. Gao, B. Jiang, and S. Zhang. On the information-adaptive variants of the admm: An iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, Jul 2018.

[GL13] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[GLL$^+$23] Gaia Grosso, Nicolò Lai, Marco Letizia, Jacopo Pazzini, Marco Rando, Lorenzo Rosasco, Andrea Wulzer, and Marco Zanetti. Fast kernel methods for data quality monitoring as a goodness-of-fit test. *Machine Learning: Science and Technology*, 4(3):035029, aug 2023.

[GLNO21] Julia Gonski, Jerry Lai, Benjamin Nachman, and Inês Ochoa. High-dimensional Anomaly Detection with Radiative Return in e^+e^- Collisions. 8 2021. arXiv:2108.13451 [hep-ph].

[GLPW] Gaia Grosso, Marco Letizia, Maurizio Pierini, and Andrea Wulzer. The new-physics-learning machine as a classifier-based goodness-of-fit test.

[GMSB$^+$15] Ruben Gepner, Mirna Mihovilovic Skanata, Natalie M Bernat, Margarita Kaplow, and Marc Gershow. Computations underlying *Drosophila* photo-taxis, odor-taxis, and multi-sensory integration. *eLife*, 4:e06229, may 2015.

[GNN$^+$22] Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takac, Pavel Dvurechensky, and Bin Gu. The power of first-order smooth optimization for black-box non-smooth problems. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7241–7265. PMLR, 17–23 Jul 2022.

[GPW^{+}18] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[Gra22] Geovani Nunes Grapiglia. Worst-case evaluation complexity of a derivative-free quadratic regularization method. 2022.

[Gra23] G. N. Grapiglia. Worst-case evaluation complexity of a derivative-free quadratic regularization method. *Optimization Letters*, Feb 2023.

[GRVZ15] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Direct search based on probabilistic descent. *SIAM Journal on Optimization*, 25(3):1515–1541, 2015.

[GV13] R. Garmanjani and L. N. Vicente. Smoothing and worst-case complexity for direct-search methods in nonsmooth optimization. *IMA Journal of Numerical Analysis*, 33(3):1008–1028, 2013.

[Han06] N. Hansen. *The CMA Evolution Strategy: A Comparing Review*, pages 75–102. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[Han07] Nikolaus Hansen. *The CMA Evolution Strategy: A Comparing Review*, volume 192, pages 75–102. 06 2007.

[Hau00] M Hauskrecht. Value-function approximations for partially observable markov decision processes. *J. Artif. Intell. Res.*, 13:33, 2000.

[HBCV23] R. A. Heinonen, L. Biferale, A. Celani, and M. Vergassola. Optimal policies for bayesian olfactory search in turbulent flows. *Phys. Rev. E*, 107:055105, May 2023.

[HC21] Jordan R. Hall and Varis Carey. Accelerating Derivative-Free Optimization with Dimension Reduction and Hyperparameter Learning. *arXiv e-prints*, page arXiv:2101.07444, January 2021.

[HCKB19] Jonathan H Huggins, Trevor Campbell, Mikolaj Kasprzak, and Tamara Broderick. Scalable gaussian process inference with finite-data mean and variance guarantees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 796–805. PMLR, 2019.

[HD05] Qiao Hu and Cabell Davis. Automatic plankton image recognition with co-occurrence matrices and support vector machine. *Marine Ecology Progress Series*, 295:21–31, 2005.

[Heg99] Tarek Hegazy. Optimization of resource allocation and leveling using genetic algorithms. *Journal of construction engineering and management*, 125(3):167–175, 1999.

[HFM^{+}17] T. Hafting, M. Fyhn, S. Molden, M. B. Moser, and E. I. Moser. Scalar turbulence. *Nat. Neurosci.*, 20:1448–1464, 2017.

[HHLB11] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In Carlos A. Coello Coello, editor, *Learning and Intelligent Optimization*, pages 507–523, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[HIK^{+}21] Anna Hallin, Joshua Isaacson, Gregor Kasieczka, Claudius Krause, Benjamin Nachman, Tobias Quadfasel, Matthias Schlaffer, David Shih, and Manuel Sommerhalder. Classifying Anomalies THrough Outer Density Estimation (CATHODE). 9 2021. arXiv:2109.00546 [hep-ph].

[HKPT19] Theo Heimel, Gregor Kasieczka, Tilman Plehn, and Jennifer M. Thompson. QCD or What? *SciPost Phys.*, 6(3):030, 2019. arXiv:1808.08979 [hep-ph].

[HLLW20] Jan Hajer, Ying-Ying Li, Tao Liu, and He Wang. Novelty Detection Meets Collider Physics. *Phys. Rev. D*, 101(7):076015, 2020. arXiv:1807.10261 [hep-ph].

[HLvdMW18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[HMB^{+}17] Alex J. Hughes, Joseph D. Mornin, Sujoy K. Biswas, David P. Bauer, Simone Bianco, and Zev J. Gartner. Quantius: Generic, high-fidelity human annotation of scientific images at 105-clicks-per-hour. *bioRxiv*, 2017.

[HMG15] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA, 09–12 May 2015. PMLR.

[HMvdW^{+}20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[HNBK^{+}15] Luis Hernandez-Nunez, Jonas Belina, Mason Klein, Guangwei Si, Lindsey Claus, John R Carlson, and Aravinthan DT Samuel. Reverse-correlation analysis of navigation dynamics in *Drosophila* larva using optogenetics. *eLife*, 4:e06225, may 2015.

[HNT13] Warren Hare, Julie Nutini, and Solomon Tesfamariam. A survey of non-gradient optimization methods in structural engineering. *Advances in Engineering Software*, 59:19–28, 2013.

[HO01] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[HS06] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[HTF09] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer series in statistics. Springer, 2009.

[Hun07] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[HW78] A. Hedayat and W. D. Wallis. Hadamard matrices and their applications. *The Annals of Statistics*, 6(6):1184–1238, 1978.

[HX08] X. Huang and Y. M. Xie. Optimal design of periodic structures using evolutionary topology optimization. *Structural and Multidisciplinary Optimization*, 36(6):597–606, Nov 2008.

[JK52] J. Wolfowitz J. Kiefer. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462—466, 1952.

[JNJ18] N. K. Jain, U. Nangia, and J. Jain. A review of particle swarm optimization. *Journal of The Institution of Engineers (India): Series B*, 99(4):407–411, Aug 2018.

[JPS93] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, Oct 1993.

[JWZL19] Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3100–3109. PMLR, 09–15 Jun 2019.

[K$^+$21] Gregor Kasieczka et al. The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics. *Rept. Prog. Phys.*, 84(12):124201, 2021. arXiv:2101.08320 [hep-ph].

[Kau22] Sandeep Kaur. Online Data Monitoring of the ATLAS Muon System and Commissioning of the New Small Wheel (NSW) Data Quality System. *PoS*, ICHEP2022:1013, 11 2022.

[KBDT21] David Kozák, Stephen Becker, Alireza Doostan, and Luis Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Comput. Optim. Appl.*, 79:339–368, 2021.

[KBS$^+$20] Mohammad Khosravi, Varsha Behrunani, Roy S. Smith, Alisa Rupenyan, and John Lygeros. Cascade control: Data-driven tuning approach based on bayesian optimization. *IFAC-PapersOnLine*, 53(2):382–387, 2020. 21st IFAC World Congress.

[KCMY21] Bumsu Kim, HanQin Cai, Daniel McKenzie, and Wotao Yin. Curvature-aware derivative-free optimization. *arXiv preprint arXiv:2109.13391*, 2021.

[KCT$^+$11] Mrinal Kalakrishnan, Sachin Chitta, Evangelos Theodorou, Peter Pastor, and Stefan Schaal. Stomp: Stochastic trajectory optimization for motion planning. In *2011 IEEE international conference on robotics and automation*, pages 4569–4574. IEEE, 2011.

[KEH$^+$13] Karl G. Kempf, Feryal Erhun, Erik F. Hertzler, Timothy R. Rosenberg, and Chen Peng. Optimizing capital investment decisions at intel corporation. *Interfaces*, 43(1):62–78, 2013.

[KK13] Oliver Kramer and Oliver Kramer. K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors*, pages 13–23, 2013.

[KK17] Oliver Kramer and Oliver Kramer. *Genetic algorithms*. Springer, 2017.

[KKM$^+$23] Mohammad Khosravi, Christopher König, Markus Maier, Roy S. Smith, John Lygeros, and Alisa Rupenyan. Safety-aware cascade controller tuning using constrained bayesian optimization. *IEEE Transactions on Industrial Electronics*, 70(2):2128–2138, 2023.

[KLT03] Tamara G. Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45(3):385–482, 2003.

[KM98] Syrous K. Kooros and Bruce L. Mcmanis. A multiattribute optimization model for strategic investment decisions. *Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration*, 15(2):152–164, 1998.

[KMH$^+$19] Johannes Kirschner, Mojmir Mutny, Nicole Hiller, Rasmus Ischebeck, and Andreas Krause. Adaptive and safe bayesian optimization in high dimensions via one-dimensional subspaces. In *International Conference on Machine Learning*, pages 3429–3438. PMLR, 2019.

[KMR$^+$21] D. Kozak, C. Molinari, L. Rosasco, L. Tenorio, and S. Villa. Zeroth order optimization with orthogonal random directions, 2021.

[KR14] Jakub Konečný and Peter Richtárik. Simple complexity analysis of simplified direct search. Workingpaper, ArXiv, October 2014. 21 pages, 5 algorithms, 1 table.

[KR20] Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

[Kra91] Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

[Kri16] V Krishnamurthy. *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.

[KS20] Charanjit K. Khosa and Veronica Sanz. Anomaly Awareness. 7 2020. arXiv:2007.14462 [cs.LG].

[KSU08] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.

[KSU13] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277*, 2013.

[Kun14] S. Y. Kung. *Kernel Methods and Machine Learning*. Cambridge University Press, 2014.

[KW52] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

[KW14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[KW19] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[KZ10] Sujin Kim and Dali Zhang. Convergence properties of direct search methods for stochastic optimization. pages 1003–1011, 12 2010.

[LCK$^+$20] S. Liu, P. Y. Chen, B. Kailkhura, G. Zhang, A. O. Hero III, and P. K. Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

[LGS12] Daniel J. Lizotte, Russell Greiner, and Dale Schuurmans. An experimental methodology for response surface optimization methods. *Journal of Global Optimization*, 53(4):699–736, Aug 2012.

[LH23] Aurore Loisy and Robin A. Heinonen. Deep reinforcement learning for the olfactory search pomdp: a quantitative benchmark. *Cereb CortexThe European Physical Journal E*, 46:17, 2023.

[LIG$^+$18] Jessica Y. Luo, Jean-Olivier Irisson, Benjamin Graham, Cedric Guigand, Amin Sarafraz, Christopher Mader, and Robert K. Cowen. Automated plankton image analysis using convolutional neural networks. *Limnology and Oceanography: Methods*, 16(12):814–827, 2018.

[LKC$^+$18] Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[LLR$^+$22] Marco Letizia, Gianvito Losapio, Marco Rando, Gaia Grosso, Andrea Wulzer, Maurizio Pierini, Marco Zanetti, and Lorenzo Rosasco. Learning new physics efficiently with nonparametric methods. *Eur. Phys. J. C*, 82(10):879, 2022.

[LN19] Alessandra Lumini and Loris Nanni. Deep learning and transfer learning features for plankton classification. *Ecological informatics*, 51:33–43, 2019.

[LO17] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[Loj63] Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

[LOSC19] Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaussian process meets big data: A review of scalable gps, 2019.

[LPK16] Hansang Lee, Minseok Park, and Junmo Kim. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In *2016 IEEE international conference on image processing (ICIP)*, pages 3713–3717. IEEE, 2016.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[LSC$^+$16] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1):1104–1123, 04 2016.

[LTOS19] Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. In *International conference on machine learning*, pages 3905–3914. PMLR, 2019. arXiv:1806.09178 [stat.ML].

[LTT00] R. M. Lewis, V. Torczon, and M. W. Trosset. Direct search methods: then and now. *Journal of Computational and Applied Mathematics*, 124(1):191–207, 2000. Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations.

[LZJ22] T. Lin, Z. Zheng, and M. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26160–26175. Curran Associates, Inc., 2022.

[Mar19] Alexandros Marantis. The ATLAS Fast TracKer—Architecture, Status and High-Level Data Quality Monitoring Framework. *Universe*, 5(1):32, 2019.

[Mat95] P. Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

[MB11] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[Mc10] Arnd Meyer and CMS collaboration. Music—an automated scan for deviations between data and monte carlo simulation. In *AIP Conference Proceedings*, volume 1200, pages 293–296. American Institute of Physics, 2010.

[MCdFDC07] R. Martinez-Cantin, N. de Freitas, A. Doucet, and J. Castellanos. Active policy learning for robot planning and exploration under uncertainty. In *Proceedings of Robotics: Science and Systems*, Atlanta, GA, USA, June 2007.

[MCDVR22] Giacomo Meanti, Luigi Carratino, Ernesto De Vito, and Lorenzo Rosasco. Efficient hyperparameter tuning for large scale kernel ridge regression. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

[MCRR20] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: Handling billions of points efficiently. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14410–14422. Curran Associates, Inc., 2020.

[MDFF$^+$21] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Robustness of classifiers to universal perturbations: a geometric perspective, 2021.

[Mez06] Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54:592–604, 10 2006.

[MFBR19] Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Globally convergent newton methods for ill-conditioned generalized self-concordant losses. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[MGR18] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1805–1814, Red Hook, NY, USA, 2018. Curran Associates Inc.

[MHES19] Ketil Malde, Nils Olav Handegard, Line Eikvil, and Arnt-Børre Salberg. Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 77(4):1274–1285, 04 2019.

[MK19] Mojmír Mutnỳ and Andreas Krause. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. *Advances in Neural Information Processing Systems 31*, pages 9005–9016, 2019.

[MLB$^+$20] BT Michaelis, KW Leathers, YV Bobkov, BW Ache, JC Principe, R Baharloo, IM Park, and MA Reidenbach. Odor tracking in aquatic organisms: the importance of temporal and spatial intermittency of the turbulent plume. *Sci. Rep.*, 10:7961, 2020.

[MLJR91] S M Morrill, R G Lane, G Jacobson, and I I Rosen. Treatment planning optimization using constrained simulated annealing. *Physics in Medicine & Biology*, 36(10):1341, oct 1991.

[MLM$^+$22] Andrew M. M. Matheson, Aaron J. Lanz, Ashley M. Medina, Al M. Licata, Timothy A. Currier, Mubarak H. Syed, and Katherine I. Nagel. A neural circuit for wind-guided olfactory navigation. *Nature Communications*, 13:4613, 2022.

[MOBR19] Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2294–2340. PMLR, 25–28 Jun 2019.

[MPN$^+$23] Andrea Maracani, Vito Paolo Pastore, Lorenzo Natale, Lorenzo Rosasco, and Francesca Odone. In-domain versus out-of-domain transfer learning in plankton image classification. *Scientific Reports*, 13(1):10443, Jun 2023.

[MPT$^+$22] Matteo Migliorini, Jacopo Pazzini, Andrea Triossi, Marco Zanetti, and Alberto Zucchetta. Muon trigger with fast Neural Networks on FPGA, a demonstrator. *J. Phys. Conf. Ser.*, 2374(1):012099, 2022.

[MPvWT23] Tobia Marcucci, Mark Petersen, David von Wrangel, and Russ Tedrake. Motion planning around obstacles with convex optimization. *Science robotics*, 8(84):eadf7843, 2023.

[MRS$^+$21] Danylo Malyuta, Taylor P Reynolds, Michael Szmuk, Thomas Lew, Riccardo Bonalli, Marco Pavone, and Behcet Acikmese. Convex optimization for trajectory generation. *arXiv preprint arXiv:2106.09125*, 2021.

[Mun11] Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[Mun14] Rémi Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–129, 2014.

[MWHF12] Nimalan Mahendran, Ziyu Wang, Firas Hamze, and Nando De Freitas. Adaptive mcmc with bayesian optimization. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 751–760, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[MXZ06] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(95):2651–2667, 2006.

[Nac20] Benjamin Nachman. Anomaly Detection for Physics Analysis and Less than Supervised Learning. 10 2020. arXiv:2010.14554 [hep-ph].

[Nes13] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[Nes14] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.

[NP33] Jerzy Neyman and Egon Sharpe Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Phil. Trans. Roy. Soc. Lond. A*, 231(694-706):289–337, 1933.

[NS17a] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, apr 2017.

[NS17b] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[NS20] Benjamin Nachman and David Shih. Anomaly Detection with Density Estimation. *Phys. Rev. D*, 101:075042, 2020. arXiv:2001.04990 [hep-ph].

[NW99] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[OB17] Eric C Orenstein and Oscar Beijbom. Transfer learning and deep feature extraction for planktonic image data sets. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1082–1088. IEEE, 2017.

[ON78] J. O'Keefe and L. Nadel. *The Hippocampus as a Cognitive Map*. Oxford Univ. Press, 1978.

[Opi67] Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591 – 597, 1967.

[Osa20] Takayuki Osa. Multimodal trajectory optimization for motion planning. *The International Journal of Robotics Research*, 39(8):983–1001, 2020.

[Ost21] Bryan Ostdiek. Deep Set Auto Encoders for Anomaly Detection in Particle Physics. 9 2021. arXiv:2109.01695 [hep-ph].

[PABM22] Cindy Poo, Gautam Agarwal, Niccolo Bonacchi, and Zachary F. Mainen. Spatial maps in piriform cortex during olfactory navigation. *Nature*, 601:595–599, 2022.

[PCG$^+$19] Adrian Alan Pol, Gianluca Cerminara, Cecile Germain, Maurizio Pierini, and Agrima Seth. Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider. *Comput. Softw. Big Sci.*, 3(1):3, 2019.

[PCGP22] Adrian Alan Pol, Gianluca Cerminara, Cécile Germain, and Maurizio Pierini. Data Quality Monitoring Anomaly Detection. In *Artificial Intelligence for High Energy Physics*. World Scientific, March 2022.

[PG11] Rushen B Patel and Paul J Goulart. Trajectory generation for aircraft avoidance maneuvers using online optimization. *Journal of guidance, control, and dynamics*, 34(1):218–230, 2011.

[PGC$^+$17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[PGM$^+$19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[PMB22] Vito Paolo Pastore, Nimrod Megiddo, and Simone Bianco. An anomaly detection approach for plankton species discovery. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, pages 599–609, Cham, 2022. Springer International Publishing.

[Pol87] B. T. Polyak. Introduction to optimization. *Optimization Software Inc., Publications Division, New York*, 1:32, 1987.

[PRR06] Christopher John Price, M Reale, and BL Robertson. A direct search method for smooth and nonsmooth unconstrained optimization. *ANZIAM Journal*, 48:C927–C948, 2006.

[PRU$^+$20] Sang Eon Park, Dylan Rankin, Silviu-Marian Udrescu, Mikaeel Yunus, and Philip Harris. Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge. *JHEP*, 21:030, 2020. arXiv:2011.03550 [hep-ph].

[PT04] Dobrivoje Popovic and Andrew R Teel. Direct search methods for nonsmooth optimization. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, volume 3, pages 3173–3178. IEEE, 2004.

[PVG$^+$11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PZBB19] Vito P Pastore, Thomas Zimmerman, Sujoy K Biswas, and Simone Bianco. Establishing the baseline for using plankton as biosensor. In *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVII*, volume 10881, page 108810H. International Society for Optics and Photonics, 2019.

[PZBB20] Vito P. Pastore, Thomas G. Zimmerman, Sujoy K. Biswas, and Simone Bianco. Annotation-free learning of plankton for classification and anomaly detection. *Scientific Reports*, 10(1):12142, Jul 2020.

[QCR05] Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[QKR17] Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[QLYS15] Hongwei Qin, Xiu Li, Zhixiong Yang, and Min Shang. When underwater imagery analysis meets deep learning: A solution at the age of big visual data. In *OCEANS 2015 - MTS/IEEE Washington*, pages 1–5, 2015.

[QnCR05] Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, December 2005.

[Ras03] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[RCR15] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015. arXiv:1507.04717 [stat.ML].

[RCR17] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[RCVR22] M. Rando, L. Carratino, S. Villa, and L. Rosasco. Ada-bkb: Scalable gaussian process optimization on continuous domains by adaptive discretization. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7320–7348. PMLR, 28–30 Mar 2022.

[RD18] Brad A. Radvansky and Daniel A. Dombeck. An olfactory virtual reality system for mice. *Nature Communications*, 9:839, 2018.

[RGGRP22] Álvaro Rubio-García, Juan José García-Ripoll, and Diego Porras. Portfolio optimization with discrete simulated annealing. *arXiv preprint arXiv:2210.00807*, 2022.

[RJI17] Dhruv Rathi, Sushant Jain, and S Indu. Underwater fish species classification using convolutional neural network and deep learning. In *2017 Ninth international conference on advances in pattern recognition (ICAPR)*, pages 1–6. IEEE, 2017.

[RMRS22] Nicola Rigolli, Nicodemo Magnoli, Lorenzo Rosasco, and Agnese Seminara. Learning to predict target location with turbulent odor plumes. *eLife*, 11:e72196, aug 2022.

[RMRV23] Marco Rando, Cesare Molinari, Lorenzo Rosasco, and Silvia Villa. An optimal structured zeroth-order algorithm for non-smooth optimization. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36738–36767. Curran Associates, Inc., 2023.

[RMV22] Gautam Reddy, Venkatesh N. Murthy, and Massimo Vergassola. Olfactory sensing and navigation in turbulent environments. *Annual Review of Condensed Matter Physics*, 13(1):191–213, 2022.

[RMVR22] M. Rando, C. Molinari, S. Villa, and L. Rosasco. Stochastic zeroth order descent with structured directions, 2022.

[Roc15] Ralph Tyrell Rockafellar. Convex analysis. 2015.

[Rov15] M Rovere. The Data Quality Monitoring Software for the CMS experiment at the LHC. *Journal of Physics: Conference Series*, 664(7):072039, dec 2015.

[Roz19] Leonel Rozo. Interactive trajectory adaptation through force-guided bayesian optimization. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7596–7603. IEEE, 2019.

[RR07] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[RR17] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017. arXiv:1602.04474 [stat.ML].

[RR22a] Lindon Roberts and Clément W. Royer. Direct search based on probabilistic descent in reduced spaces, 2022.

[RR22b] C. Rusu and L. Rosasco. Fast approximation of orthogonal matrices and application to pca. *Signal Processing*, 194:108451, 2022.

[RRSV22] Nicola Rigolli, Gautam Reddy, Agnese Seminara, and Massimo Vergassola. Alternation emerges as a multi-modal strategy for turbulent odor navigation. *eLife*, 11:e76989, aug 2022.

[RS71] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications**research supported by nih grant 5-r01-gm-16895-03 and onr grant n00014-67-a-0108-0018. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.

[RVV20] Lorenzo Rosasco, Silvia Villa, and Bang Công Vũ. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics & Optimization*, 82(3):891–917, Dec 2020.

[RZBS09] Nathan Ratliff, Matt Zucker, J Andrew Bagnell, and Siddhartha Srinivasa. Chomp: Gradient optimization techniques for efficient motion planning. In *2009 IEEE international conference on robotics and automation*, pages 489–494. IEEE, 2009.

[SAGF16] Moritz Sebastian Schmid, Cyril Aubry, Jordan Grigor, and Louis Fortier. The loki underwater imaging system and an automatic identification model for the detection of zooplankton taxa in the arctic ocean. *Methods in Oceanography*, 15-16:129–160, 2016. Computer Vision in Oceanography.

[SB98] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[SDH$^+$14] John Schulman, Yan Duan, Jonathan Ho, Alex Lee, Ibrahim Awwal, Henry Bradlow, Jia Pan, Sachin Patil, Ken Goldberg, and Pieter Abbeel. Motion planning with sequential convex optimization and convex collision checking. *The International Journal of Robotics Research*, 33(9):1251–1270, 2014.

[SGBK15] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, pages 997–1005. PMLR, 2015.

[SGT18] Yitong Sun, Anna Gilbert, and Ambuj Tewari. But how does it work in theory? linear svm with random features. 2018. arXiv:1809.04481.

[SGV98] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998.

[Sha17] O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

[SHC$^+$17] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning, 2017.

[SHCS17] Tim Salimans, Jonathan Ho, Xi Chen, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *ArXiv*, page arXiv:1703.03864, 2017.

[Sin12] Dr. Narinder Singh. Review of particle swarm optimization. *International Journal of Computational Intelligence and Information Security, April 2012*, Vol. 3:34–44, 01 2012.

[SJ18] Shubhanshu Shekhar and Tara Javidi. Gaussian process bandits with adaptive discretization. *Electronic Journal of Statistics*, 12(2):3829 – 3874, 2018.

[SJ20] Shubhanshu Shekhar and Tara Javidi. Multi-scale zero-order optimization of smooth functions in an rkhs, 2020.

[SKK20] Simon-Martin Schröder, Rainer Kiko, and Reinhard Koch. Morphocluster: efficient annotation of plankton images by clustering. *Sensors*, 20(11):3060, 2020.

[SKKS10] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022, 2010.

[SKKS12] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory - TIT*, 58:3250–3265, 05 2012.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[SNP00] E. Save, L. Nerad, and B. Poucet. Contribution of multiple sensory information to place field stability in hippocampal place cells. *Hippocampus*, 10:64, 2000.

[SO07] Heidi M Sosik and Robert J Olson. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*, 5(6):204–216, 2007.

[Spa03] James C. Spall. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc., USA, 1 edition, 2003.

[SPK13] G Shani, J Pineau, and R Kaplow. A survey of point-based pomdp solvers. *Autonomous Agents and MultiAgent Systems*, 27:1–51, 2013.

[SRB11] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization, 2011.

[SRB$^+$23] Subaselvi Sundarraj, R Vijaya Kumar Reddy, B Mahesh Babu, Gururaj Harinahalli Lokesh, Francesco Flammini, and Rajesh Natarajan. Route planning for an autonomous robotic vehicle employing a weight-controlled particle swarm-optimized dijkstra algorithm. *IEEE Access*, 2023.

[SRS$^+$15] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2171–2180, Lille, France, 07–09 Jul 2015. PMLR.

[SS00] B. Shraiman and E. Siggia. Scalar turbulence. *Nature*, 405:639, 2000.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[STDF13] Katja Schulze, Ulrich M. Tillich, Thomas Dandekar, and Marcus Frohme. Planktovision - an automated analysis system for the identification of phytoplankton. *BMC Bioinformatics*, 14(1):115, Mar 2013.

[SvBRB23] Satpreet H. Singh, Floris van Breugel, Rajesh P. N. Rao, and Bingni W. Brunton. Emergent behaviour and neural dynamics in artificial agents tracking odour plumes. *Nature Machine Intelligence*, 5:58–70, 2023.

[SVZ20] Sudeep Salgia, Sattar Vakili, and Qing Zhao. A computationally efficient approach to black-box optimization using gaussian process models. *arXiv preprint arXiv:2010.13997*, 2020.

[SVZ21] Sudeep Salgia, Sattar Vakili, and Qing Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. In *NeurIPS*, 2021.

[Tit09] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton

Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[Tot22] C. Totzeck. *Trends in Consensus-Based Optimization*, pages 201–226. Springer International Publishing, Cham, 2022.

[Tro19] T. Trogdon. On spectral and numerical properties of random butterfly matrices. *Applied Mathematics Letters*, 95:48–58, 2019.

[TTH$^+$21] Akinori Tanaka, Akio Tomiya, Koji Hashimoto, Akinori Tanaka, Akio Tomiya, and Koji Hashimoto. Basics of neural networks. *Deep Learning and Physics*, pages 35–55, 2021.

[VCM13] Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pages 19–27. PMLR, 2013.

[VKM$^+$13] Michal Valko, Nathaniel Korda, Remi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, page 654?666, 2013.

[VM21] A. Helen Victoria and G. Maragatham. Automatic tuning of hyperparameters using bayesian optimization. *Evolving Systems*, 12(1):217–223, Mar 2021.

[VMV20] Francesco Viola, Valentina Meschini, and Roberto Verzicco. Fluid–structure-electrophysiology interaction (fsei) in the left-heart: a multi-way coupled computational model. *European Journal of Mechanics-B/Fluids*, 79:212–232, 2020.

[VPC23] K. V. B. Verano, E Panizon, and A Celani. Olfactory search with finite-state controllers. *Proc Nat Acad Sci*, 120(34):e2304230120, 2023.

[VVS07] M. Vergassola, E. Villermaux, and B.I. Shraiman. 'Infotaxis' as a strategy for searching without gradients. *Nature*, 445:406, 2007.

[Wal43] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

[WCZ$^+$19] Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, and Si-Hao Deng. Hyperparameter optimization for machine learning models based on bayesian optimizationb. *Journal of Electronic Science and Technology*, 17(1):26–40, 2019.

[WF23] Tianyu Wang and Yasong Feng. Convergence rates of stochastic zeroth-order gradient descent for łojasiewicz functions, 2023.

[Whi94] Darrell Whitley. A genetic algorithm tutorial. *Statistics and computing*, 4(2):65–85, 1994.

[WHZ$^+$16] Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando De Feitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

[Wil38] S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals Math. Statist.*, 9(1):60–62, 1938.

[WKS21] Veit Wild, Motonobu Kanagawa, and Dino Sejdinovic. Connections and Equivalences between the Nyström Method and Sparse Variational Gaussian Processes. *arXiv e-prints*, page arXiv:2106.01121, jun 2021.

[WSJF14] Ziyu Wang, Babak Shakibi, Lin Jin, and Nando Freitas. Bayesian Multi-Scale Optimistic Optimization. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 1005–1014, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[WSsf$^+$19] Josh Warner, Jason Sexauer, scikit fuzzy, twmeggs, alexsavio, Aishwarya Unnikrishnan, Guilherme Castelão, Felipe Arruda Pontes, Tobias Uelwer, pd2f, laurazh, Fernando Batista, alexbuy, Wouter Van den Broeck, William Song, The Gitter Badger, Roberto Abdelkader Martínez Pérez, James F. Power, Himanshu Mishra, Guillem Orellana Trullols, Axel Hörteborn, and 99991. Jdwarner/scikit-fuzzy: Scikit-fuzzy version 0.4.2, November 2019.

[WW16] Constantin Weisser and Mike Williams. Machine learning and multivariate goodness of fit. 12 2016. arXiv:1612.07186 [physics.data-an].

[YNS12] F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

[ZBCM20] Hui Zhang, Thomas J Best, Anton Chivu, and David O Meltzer. Simulation-based optimization to improve hospital patient assignment to physicians and clinical units. *Health care management science*, 23:117–141, 2020.

[ZD15] S Zhang and Manahan-Vaughan D. Spatial olfactory learning contributes to place field formation in the hippocampus. *Cereb Cortex*, 25:423–32, 2015.

[ZLS10] Feng Zhao, Feng Lin, and Hock Soon Seah. Binary sipper plankton image classification using random subspace. *Neurocomputing*, 73(10):1853–1860, 2010. Subspace Learning / Selected papers from the European Symposium on Time Series Prediction.

[ZWY$^+$17] Haiyong Zheng, Ruchen Wang, Zhibin Yu, Nan Wang, Zhaorui Gu, and Bing Zheng. Automatic plankton image classification combining multiple view features via multiple kernel learning. *BMC Bioinformatics*, 18(16):570, Dec 2017.

[ZYF$^+$22] Meng-Yue Zhang, Shi-Chun Yang, Xin-Jie Feng, Yu-Yi Chen, Jia-Yi Lu, and Yao-Guang Cao. Route planning for autonomous driving based on traffic information via multi-objective optimization. *Applied Sciences*, 12(22):11817, 2022.