Nile University

School of Communication and Information Technology (CIT)

Master of Science/Engineering MSCIT/ MCIT Program

Status Report 18

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version No.** | **Date** | **Description** | **Created By** | **e-mail** |
| 0.1 | 11-02-2019 | Status Report 18 | Ahmed Mohamed Abdel Rahman | [Robot209@gmail.com](mailto:Robot209@gmail.com) |
|  |  |  |  |  |

## Objective

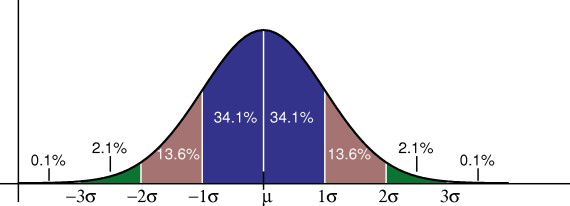
* Learn about black box optimizations mainly Bayesian Optimization using Gaussian Process

## Black Box Optimization

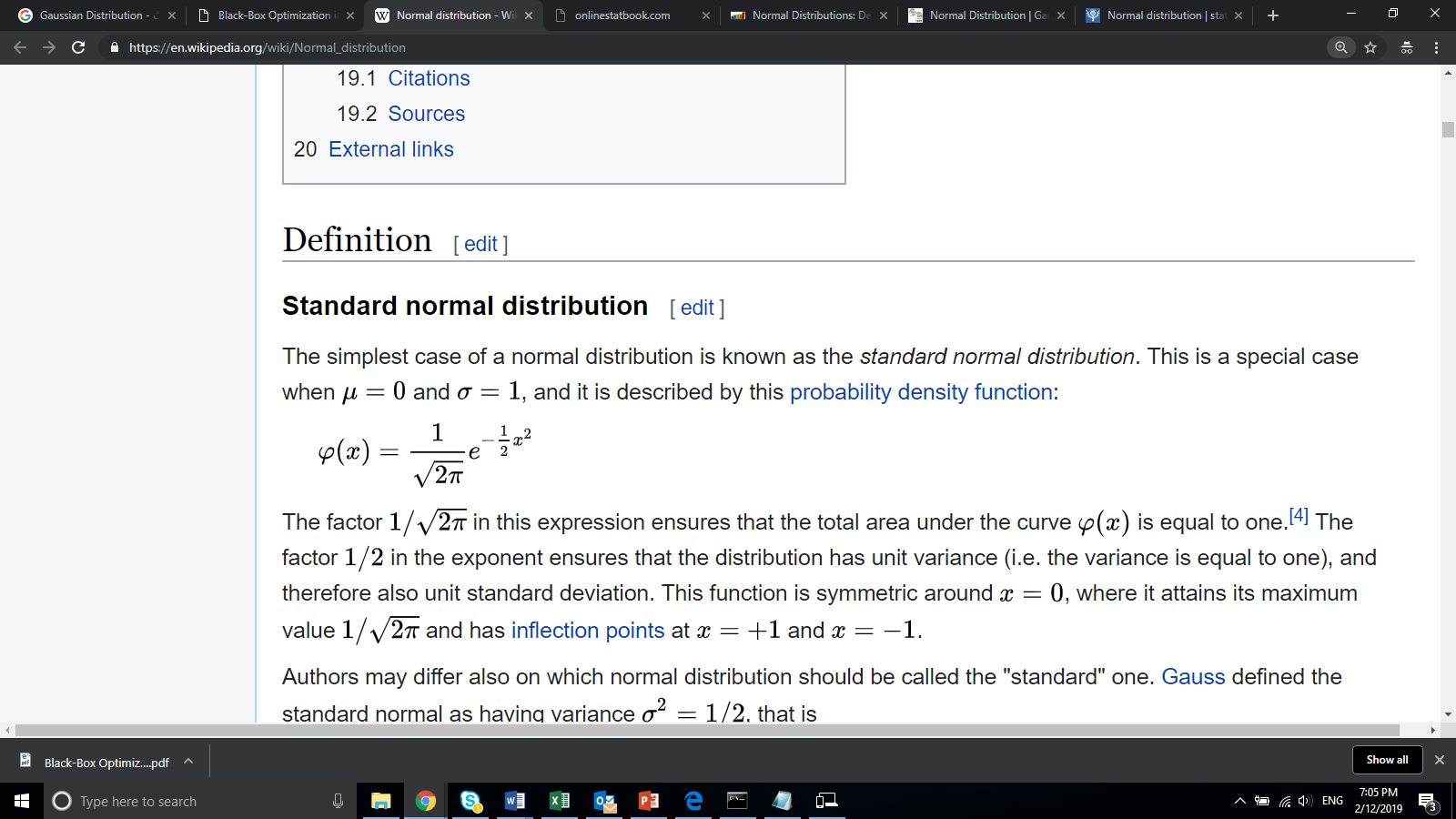
* "Black Box" optimization refers to a problem setup in which an optimization algorithm is supposed to optimize (e.g., minimize) an objective function through a so-called black-box interface
* the algorithm may query the value f(x) for a point x, but it does not obtain gradient information, and in particular it cannot make any assumptions on the analytic form of f (e.g., being linear or quadratic).
* So, Black Box Optimization methods do not rely on any explicit gradient computation, but assume that only function values can be computed, usually with noise.
* There are two relatively independent directions of research for black-box optimization
  + Bayesian Optimization (BO)
  + derivative free optimization (DFO)

## Gaussian Distribution

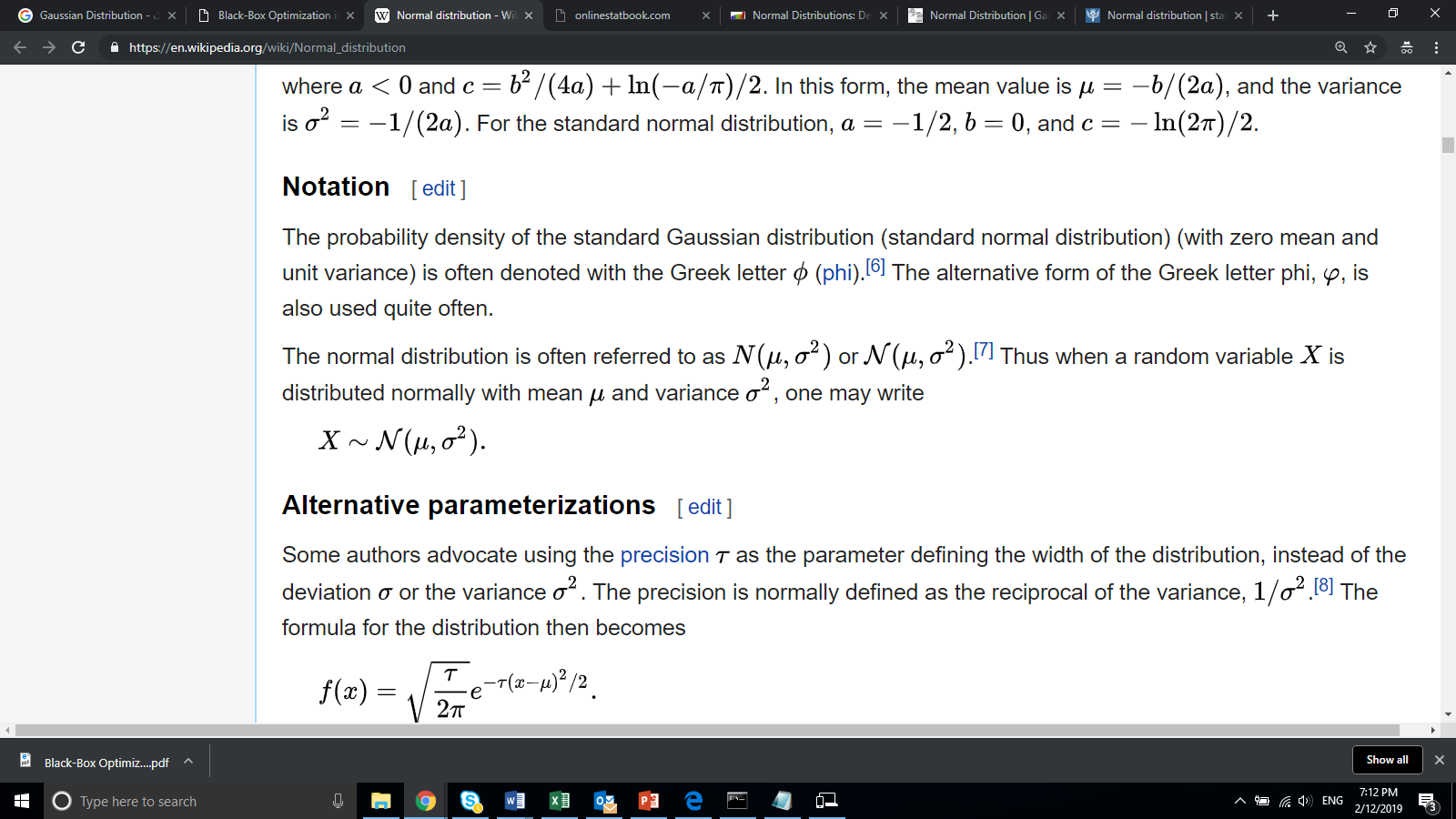
* **Normal distribution**, also called **Gaussian distribution**, the most common [distribution function](https://www.britannica.com/science/distribution-function) for independent, randomly generated variables
* The graph of the normal distribution is characterized by two parameters: the [mean](https://www.britannica.com/science/mean), or average, which is the [maximum](https://www.britannica.com/science/maximum) of the graph and about which the graph is always symmetric; and the [standard deviation](https://www.britannica.com/topic/standard-deviation), which determines the amount of dispersion away from the mean.



* The simplest case of a normal distribution is known as the standard normal distribution. This is a special case when mean = 0 and standard deviation = 1, and it is described by this probability density function (PDF):

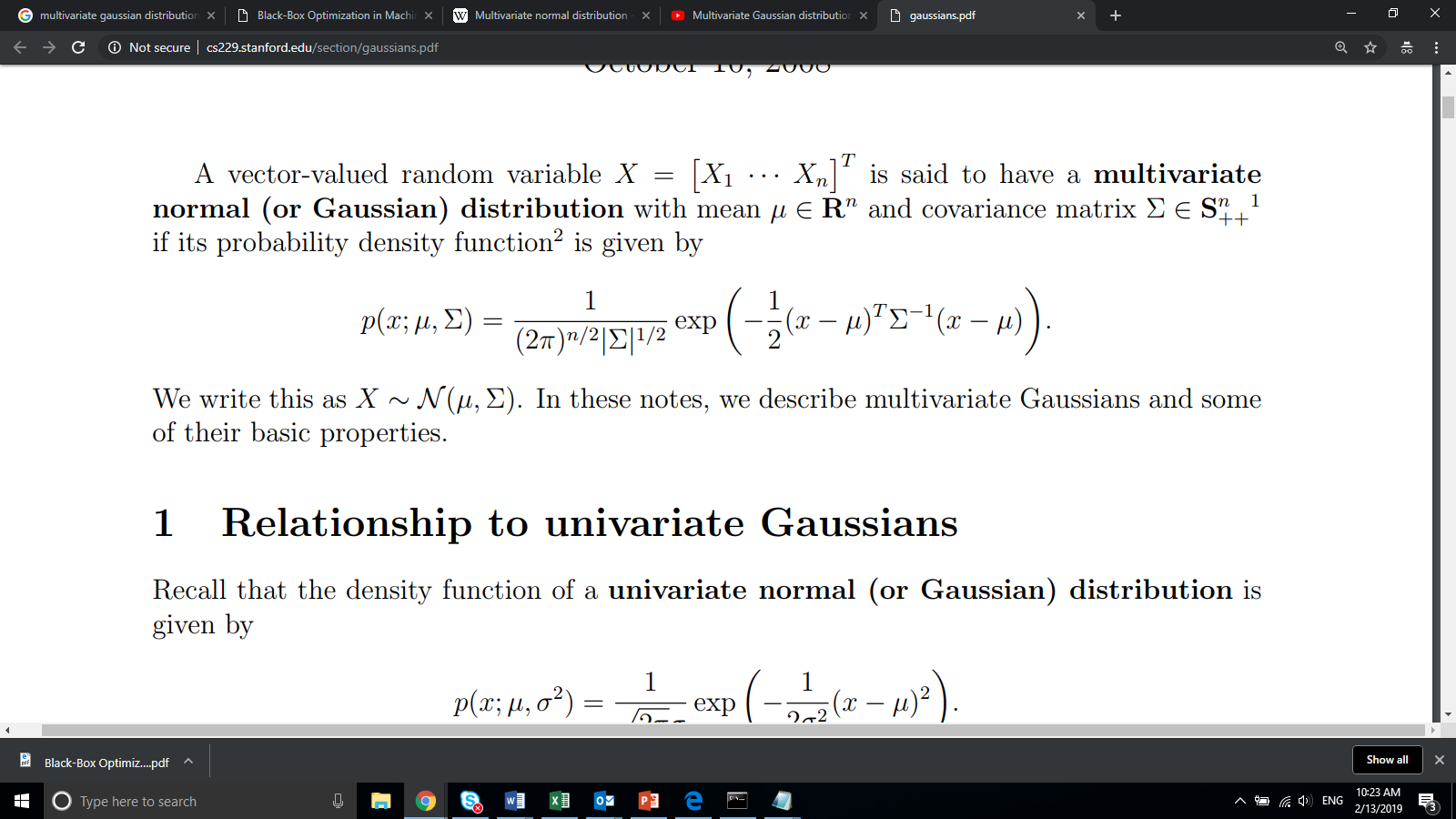


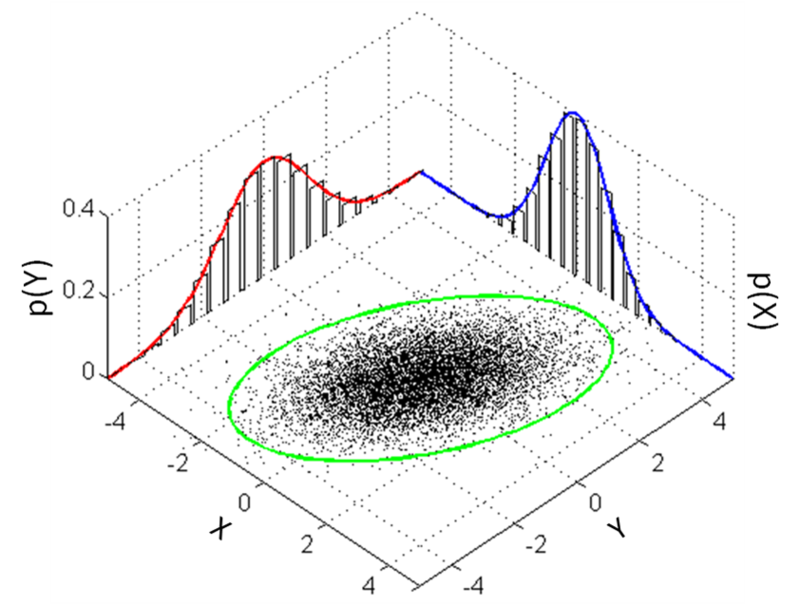
* Every normal distribution is a version of the standard normal distribution whose domain has been stretched by a factor (Standard Deviation) and translated by mean
* Thus when a random variable X is distributed normally with mean MU and variance Sigma, one may write

{\displaystyle X\sim {\mathcal {N}}(\mu ,\sigma ^{2}).} 

## Multivariate normal distribution

* One definition is that a [random vector](https://en.wikipedia.org/wiki/Random_vector) is said to be k-variate normally distributed if every [linear combination](https://en.wikipedia.org/wiki/Linear_combination) of its k components has a univariate normal distribution.
* A vector-valued random variable X =[X1 · · · Xn] T is said to have a multivariate normal (or Gaussian) distribution with mean µ ∈ Rn and covariance matrix Σ ∈ S n ++ 1 if its probability density function is given by

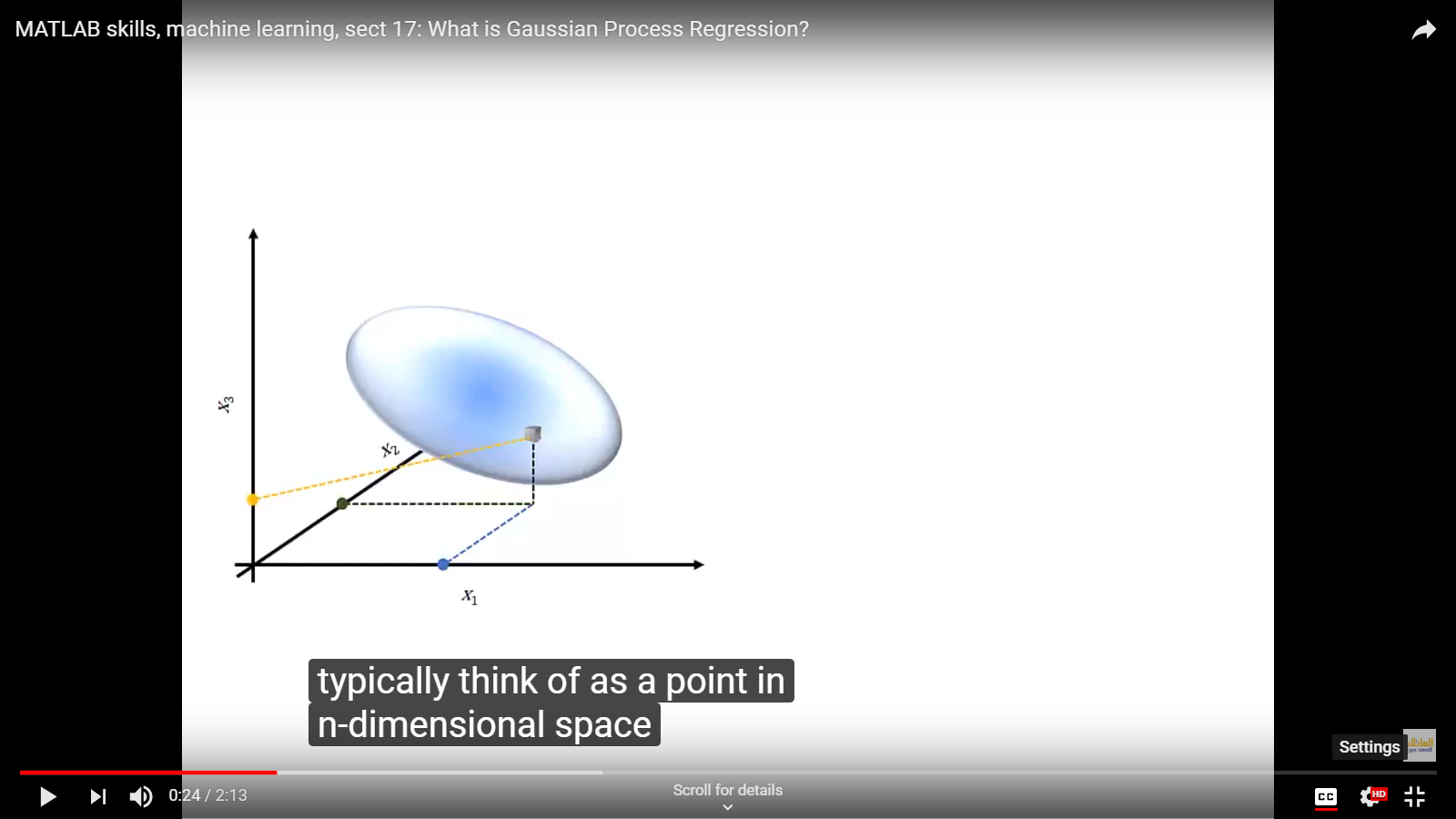




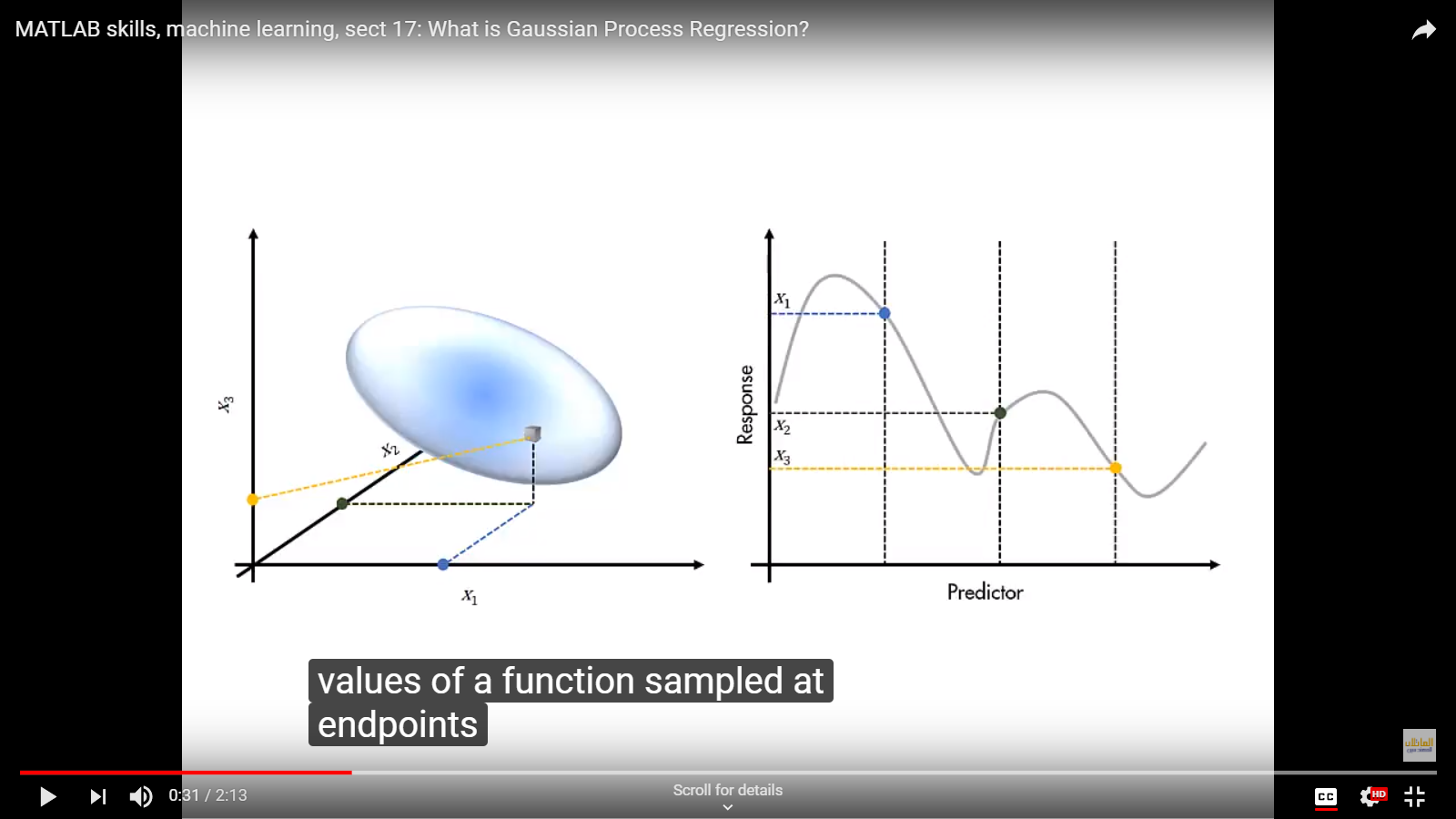
* In the definition of multivariate Gaussians, we required that the covariance matrix Σ be symmetric positive definite (i.e., Σ ∈ S n ++).
* The multivariate normal distribution is useful in analyzing the relationship between multiple normally distributed variables, and thus has heavy application to biology and economics where the relationship between approximately-normal variables is of great interest.

## Gaussian Process

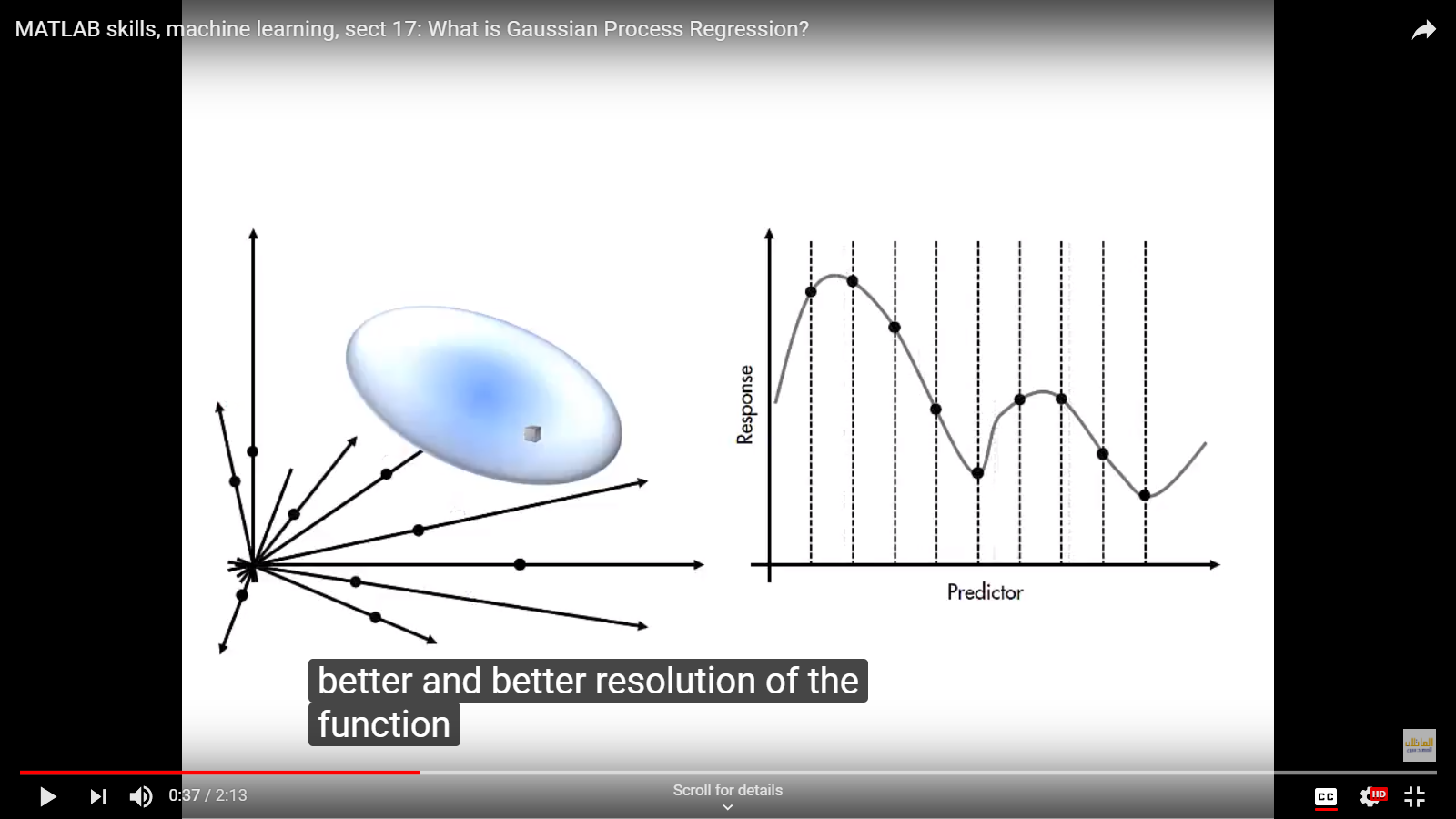
* A Gaussian process is a generalization of the Gaussian probability distribution
* It is extending to the idea of a probability distribution of numbers to probability distribution of functions.
* When we sample an n-dimensional normal distribution we get n numbers (point in n-dimension space)



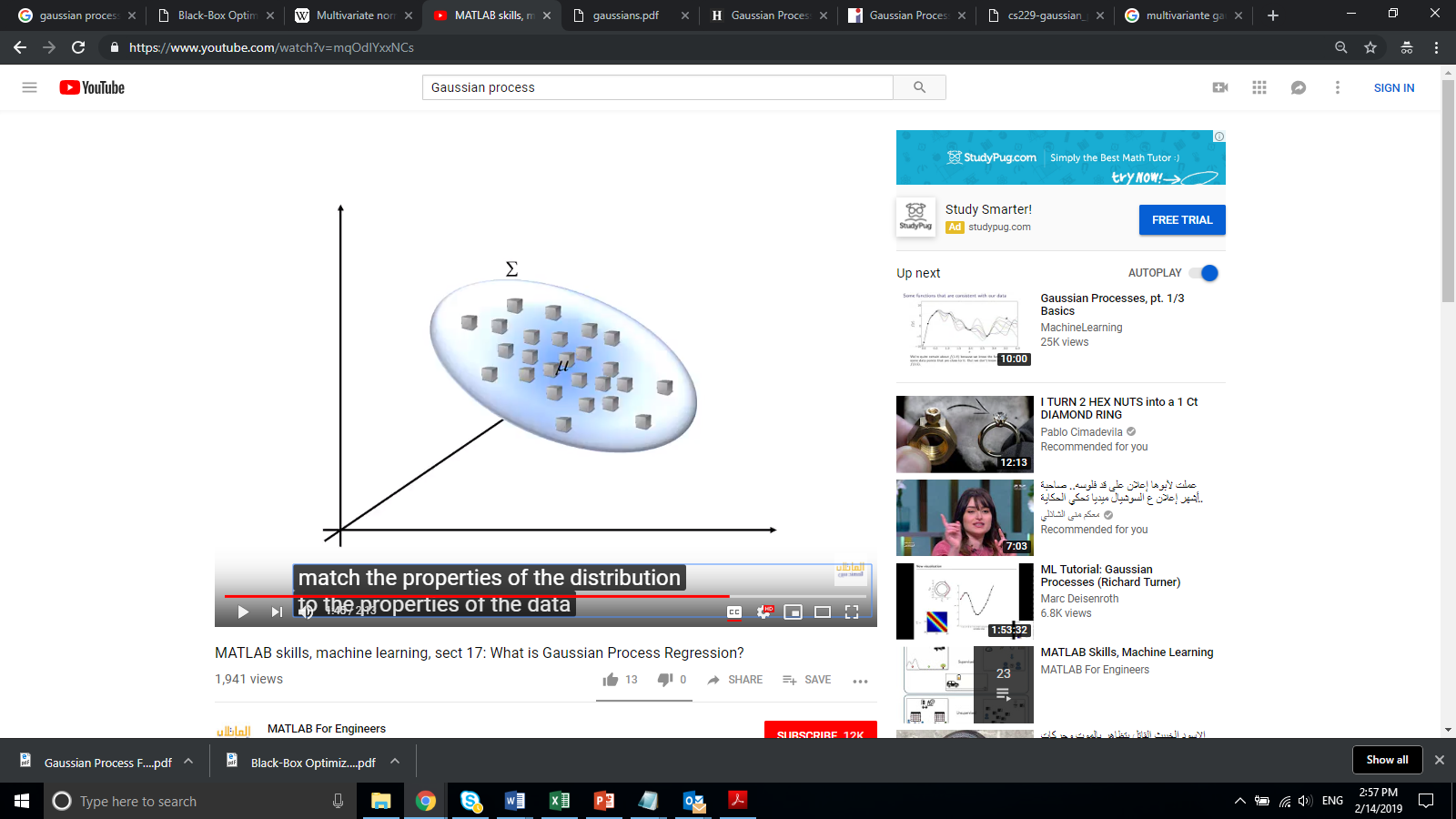
* But we can also think of them as the values of a function sampled



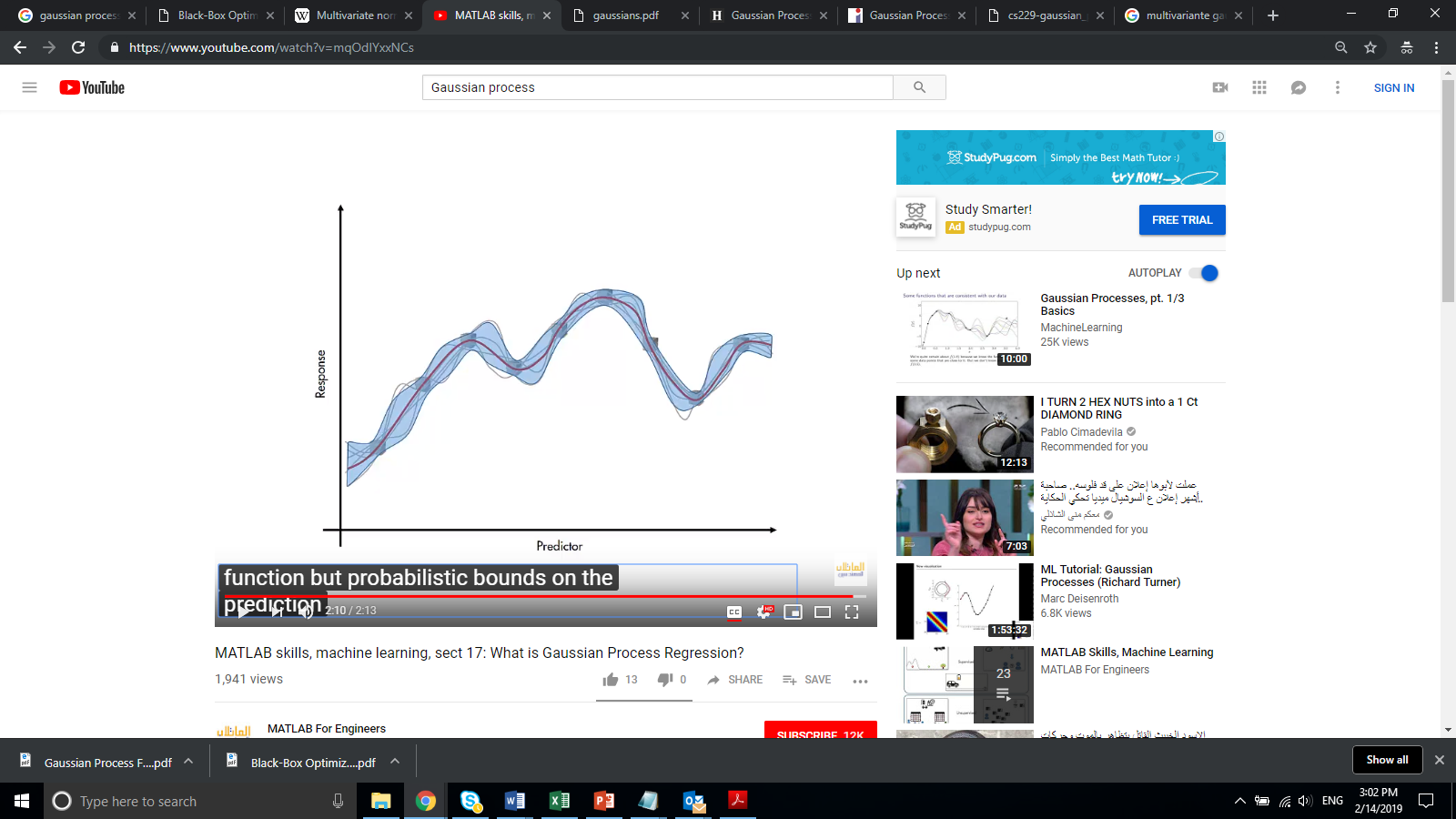
* If we let n get larger and larger we will get better resolution of the function, theoretically a functional is fully represented by a point in an infinite dimensional space (but a large value of n will be sufficient in practical purpose)



* Points in a finite dimensional space can be sampled from a probability distribution determined by a mean vector and a covariance matrix, similarly we can have a probability distribution of functions determined by a mean function and a covariance function.
* For Gaussian process regression, the covariance function is determined by a chosen kernel function that describe how much influence one point has on another (this effectively determine the smoothness of the functions in the distribution)
* Given a set of data points we can fit a probability distribution to them by choosing the distribution parameters (to match the properties of the distribution to the properties of the data)



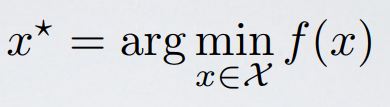
* Similarly given a set of function values we can fit a probability distribution of functions that closely match the given function values.
* Considering the hole fitted distribution of functions we can determine the mean as well as a confidence interval



* Whereas a probability distribution describes random variables which are scalars or vectors (for multivariate distributions),
* one can loosely think of a function as a very long vector, each entry in the vector specifying the function value f(x) at a particular input x.
* Indeed, many models that are commonly employed in both machine learning and statistics are in fact special cases of, or restricted kinds of Gaussian processes
* Gaussian processes are the extension of multivariate Gaussians to infinite-sized collections of real valued variables. In particular, this extension will allow us to think of Gaussian processes as distributions not just over random vectors but in fact distributions over random functions.
* probability distributions over functions with finite domains can be represented using a finite-dimensional multivariate Gaussian distribution over function outputs f(x1), . . ., f(xm) at a finite number of input points x1, . . ., xm.
* can we specify probability distributions over functions when the domain size may be infinite? For this, we turn to a fancier type of probability distribution known as a Gaussian process.
* one can think of a function f(·) drawn from a Gaussian process prior as an extremely high-dimensional vector drawn from an extremely high-dimensional multivariate Gaussian. Here, each dimension of the Gaussian corresponds to an element x from the index set X , and the corresponding component of the random vector represents the value of f(x).

## Hyperparameter Optimization

* The aim of hyperparameter optimization in machine learning is to find the hyperparameters of a given machine learning algorithm that return the best performance as measured on a validation set. (Hyperparameters, in contrast to model parameters, are set by the machine learning engineer before training.
* The number of trees in a random forest is a hyperparameter while the weights in a neural network are model parameters learned during training. I like to think of hyperparameters as the model settings to be tuned.)



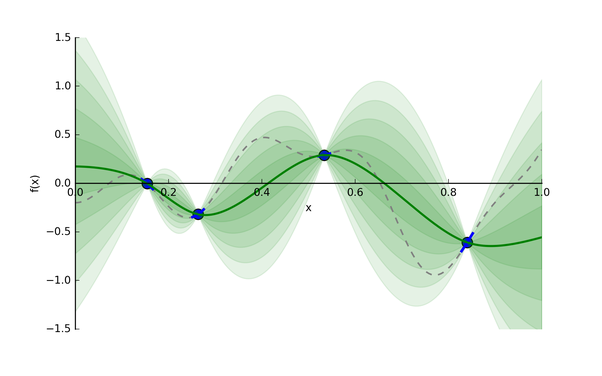
* Here f(x) represents an objective score to minimize— such as RMSE or error rate— evaluated on the validation set; x\* is the set of hyperparameters that yields the lowest value of the score, and x can take on any value in the domain X.
* In simple terms, we want to **find the model hyperparameters that yield the best score on the validation set metric**.
* The problem with hyperparameter optimization is that evaluating the objective function to find the score is extremely expensive. Each time we try different hyperparameters, we have to train a model on the training data, make predictions on the validation data, and then calculate the validation metric. With a large number of hyperparameters and complex models such as ensembles or deep neural networks that can take days to train, this process quickly becomes intractable to do by hand!
* Grid search and [random search are slightly better](http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf) than manual tuning because we set up a grid of model hyperparameters and run the train-predict -evaluate cycle automatically in a loop while we do more productive things (like[feature engineering](https://www.featuretools.com/)).
* However, even these methods are relatively inefficient because they do not choose the next hyperparameters to evaluate based on previous results. **Grid and random search are completely** *uninformed***by past evaluations,**and as a result, often spend a significant amount of time evaluating “bad” hyperparameters.

## Bayesian Optimization (using GP)

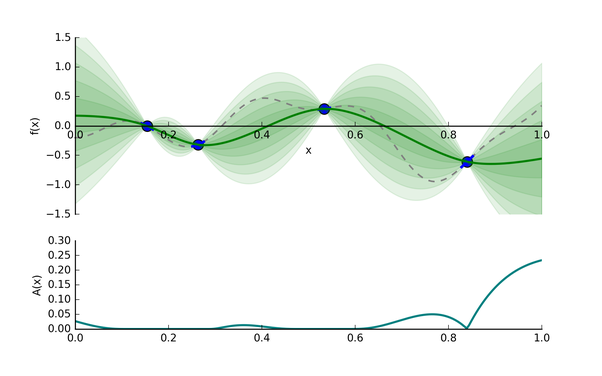
* [Bayesian hyperparameter optimization](https://sigopt.com/static/pdf/SigOpt_Bayesian_Optimization_Primer.pdf) is: build a probability model of the objective function and use it to select the most promising hyperparameters to evaluate in the true objective function.
* [Bayesian approaches](https://www.iro.umontreal.ca/~bengioy/cifar/NCAP2014-summerschool/slides/Ryan_adams_140814_bayesopt_ncap.pdf), in contrast to random or grid search, keep track of past evaluation results which they use to form a probabilistic model mapping hyperparameters to a probability of a score on the objective function:

https://cdn-images-1.medium.com/max/800/1*u00KlxHhd1fz6-Jaeou6PA.png

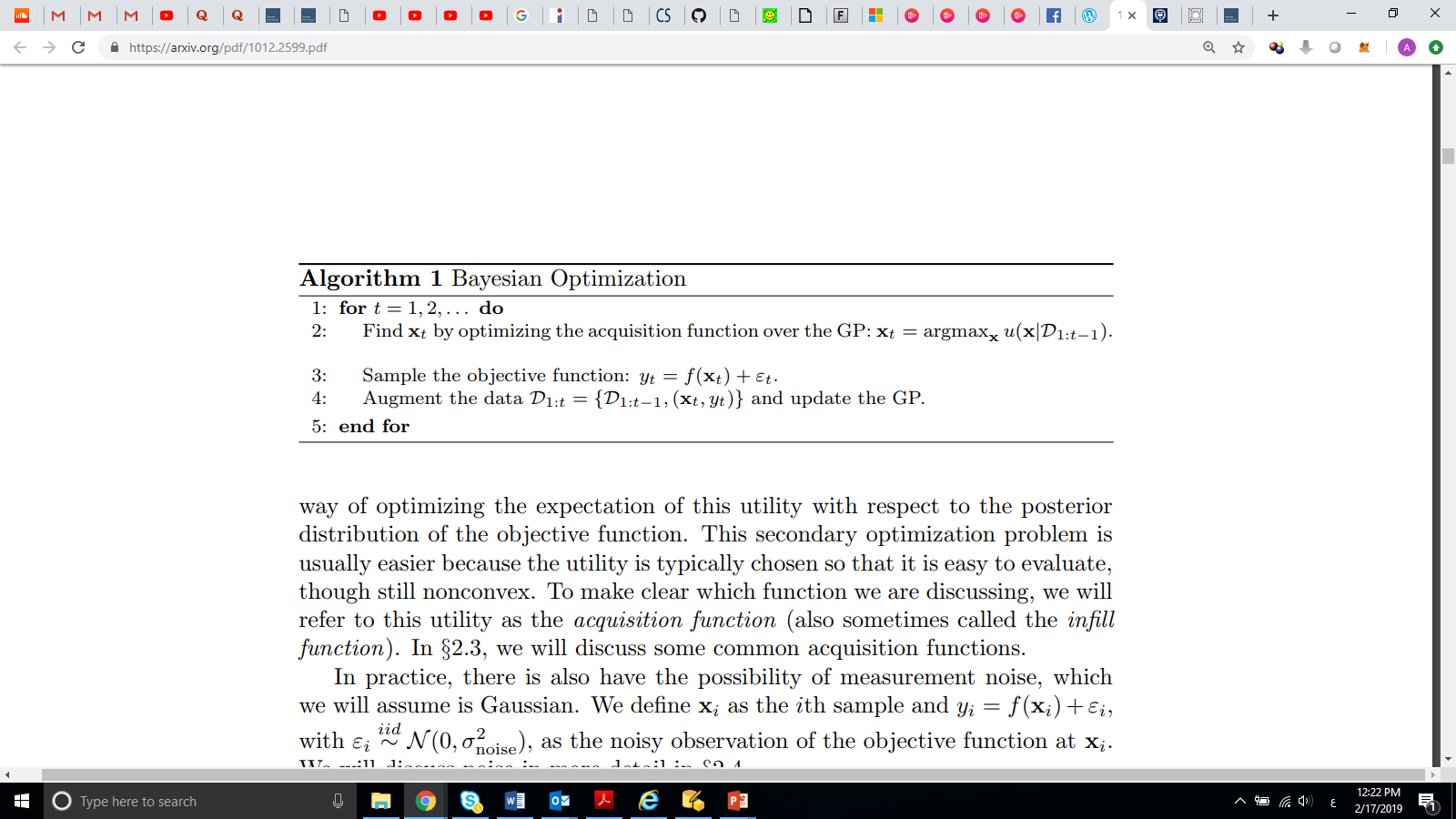
* this model is called a “surrogate” for the objective function and is represented as p(y | x). The surrogate is much easier to optimize than the objective function and Bayesian methods work by finding the next set of hyperparameters to evaluate on the actual objective function by selecting hyperparameters that perform best on the surrogate function.
* In other words:
  1. Build a surrogate probability model of the objective function
  2. Find the hyperparameters that perform best on the surrogate
  3. Apply these hyperparameters to the true objective function
  4. Update the surrogate model incorporating the new results
  5. Repeat steps 2–4 until max iterations or time is reached
* he basic idea is: **spend a little more time selecting the next hyperparameters in order to make fewer calls to the objective function.**
* In practice, the time spent selecting the next hyperparameters is inconsequential compared to the time spent in the objective function.
* Bayesian optimization is a powerful strategy for finding the extrema of objective functions that are expensive to evaluate.
* Sample some input-outputs (ex: less than 10) and use them to guess the true function with 'Gaussian Process' (or ‘GP’ here).
* Then use that guessed function to determine where to evaluate next.
* Evaluate that point, add it to our set of input-outputs and infer the guessed function once again.
* Repeat this until you've exhausted your budget of evaluations (or some other stopping criteria). If the GP is any good at guessing the true function, we’ll do better than random sampling.



* The solid green line is our guess of the true function. Each additional band of green is another half standard deviation on the output distribution.
* We have to develop “*acquisition function*” that balance between exploration-exploitation (That is, we should evaluate on points where the solid green line is low and We should check areas we know less about).
* This function is cheap, so we can optimize it and use that x as our next point to search next.

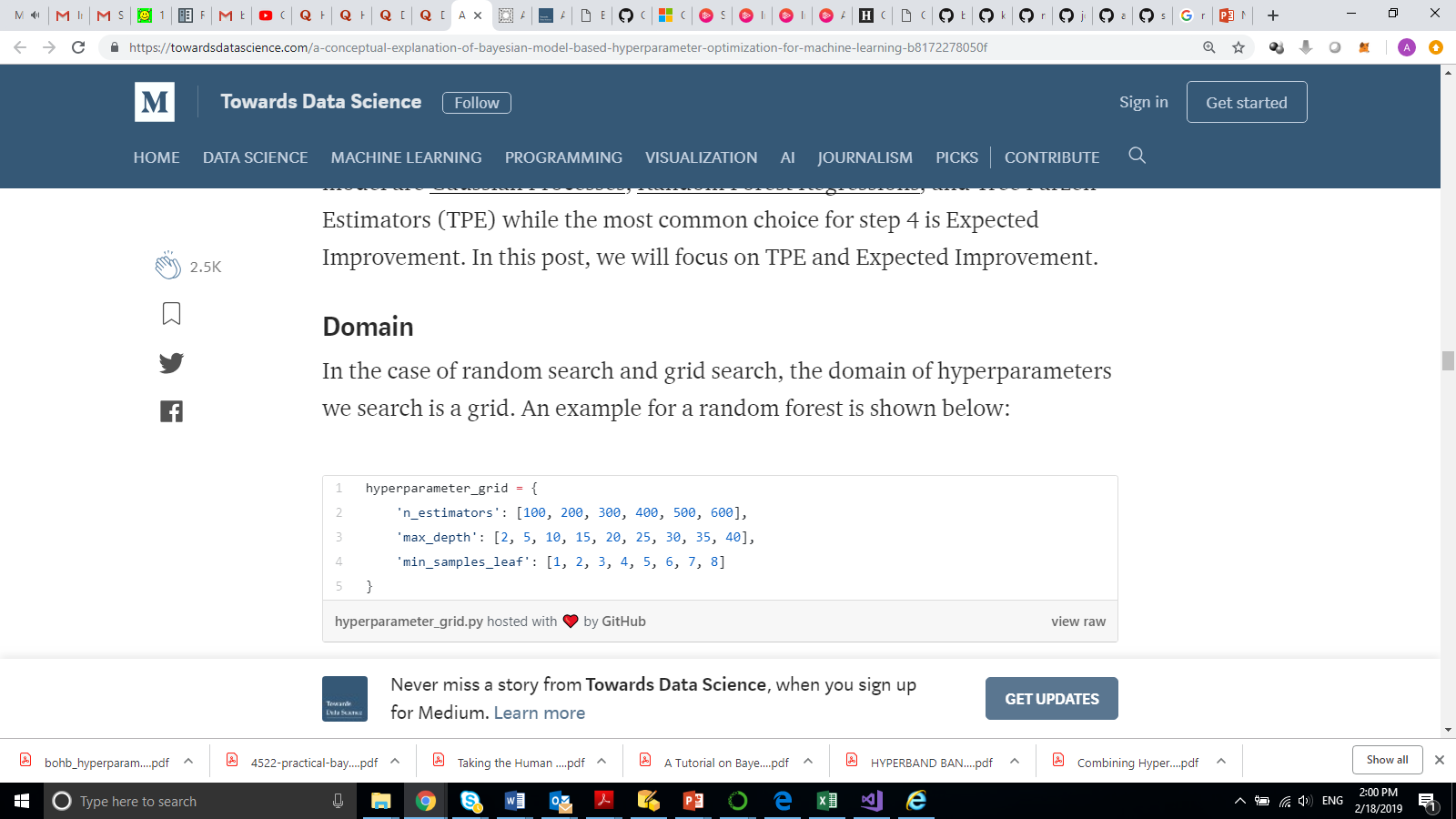


* Generally, the optimization is over more than just one dimension. GPs scale to more than one dimension naturally,



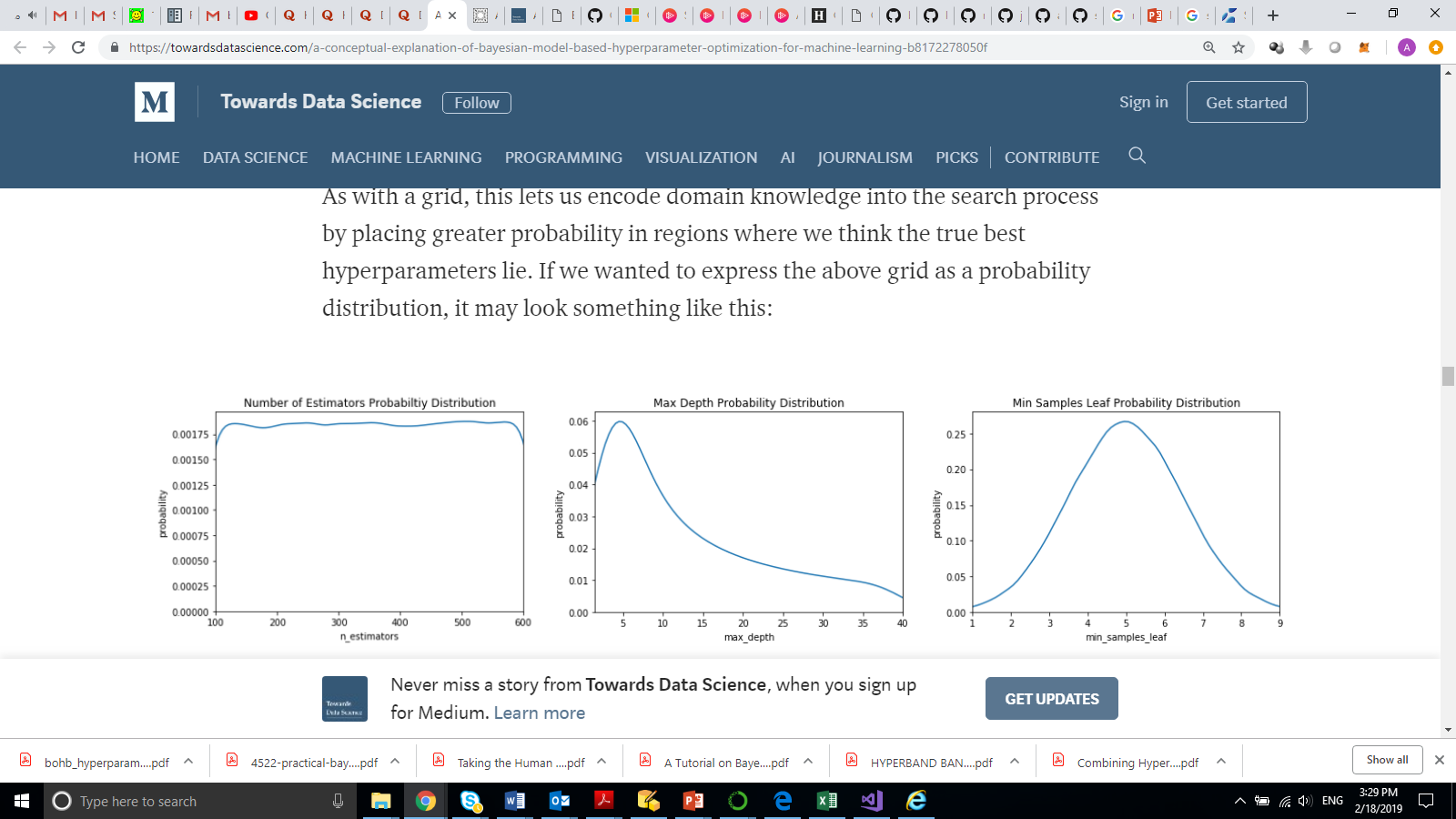
## Sequential Model-Based Optimization (SMBO)

* [Sequential model-based optimization (SMBO) methods](https://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf)  are a formalization of Bayesian optimization
* The sequential refers to running trials one after another, each time trying better hyperparameters by applying Bayesian reasoning and updating a probability model (surrogate).
* There are five aspects of model-based hyperparameter optimization:
  1. A domain of hyperparameters over which to search
  2. An objective function which takes in hyperparameters and outputs a score that we want to minimize (or maximize)
  3. The surrogate model of the objective function
  4. A Criteria, called a selection function, for evaluating which hyperparameters to choose next from the surrogate model
  5. A history consisting of (score, hyperparameter) pairs used by the algorithm to update the surrogate model
* There are several variants of [SMBO methods that differ](https://sigopt.com/static/pdf/SigOpt_Bayesian_Optimization_Primer.pdf) in steps 3–4, namely, how they build a surrogate of the objective function and the criteria used to select the next hyperparameters.
* Several common choices for the surrogate model are
  + [Gaussian Processes](https://en.wikipedia.org/wiki/Gaussian_process)
  + [Random Forest Regressions](http://aad.informatik.uni-freiburg.de/papers/13-GECCO-BBOB_SMAC.pdf)
  + Tree Parzen Estimators (TPE)
* **Domain:**
  + In the case of random search and grid search, the domain of hyperparameters we search is a grid. For example

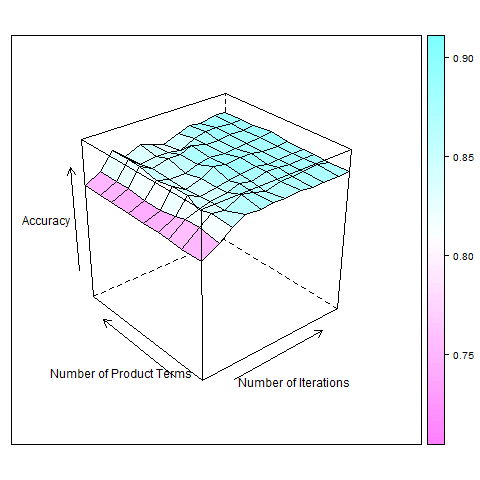


* + For a model-based approach, the domain consists of *probability distributions*.

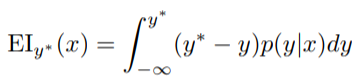
If we wanted to express the above grid as a probability distribution, it may look something like this:



* + Here we have a uniform, log-normal, and normal distribution.
* **Objective Function**
  + The objective function takes in hyperparameters and outputs a single real-valued score that we want to minimize (or maximize).
  + The entire concept of Bayesian model-based optimization is to reduce the number of times the objective function needs to be run by choosing only the most promising set of hyperparameters to evaluate based on previous calls to the evaluation function.
* **Surrogate Function (Probability Model)**
  + The surrogate function, also called the response surface, is the probability representation of the objective function built using previous evaluations.
  + This is called sometimes called a response surface because it is a high-dimensional mapping of hyperparameters to the probability of a score on the objective function.



* + There are several different forms of the surrogate function:
    - [Gaussian Processes](https://en.wikipedia.org/wiki/Gaussian_process)
    - [Random Forest Regressions](http://aad.informatik.uni-freiburg.de/papers/13-GECCO-BBOB_SMAC.pdf)
    - Tree Parzen Estimators (TPE)
* **Selection Function**
  + The selection function is the criteria by which the next set of hyperparameters are chosen from the surrogate function.
  + The most common choice of criteria is Expected Improvement:



## Hyperband Algorithm and Successive Halving algorithm

* **Successive Halving**
  + uniformly allocate a budget to a set of hyperparameter configurations, evaluate the performance of all configurations, throw out the worst half, and repeat until one configuration remains.
  + The algorithm allocates exponentially more resources to more promising configurations.
* Hyperband
  + It considers several possible values of n (hyperparameters) for a fixed Budget B (e.g., an hour of training time to choose a hyperparameter configuration) [it is like performing a grid search over feasible value of n]
  + Associated with each value of n is a minimum resource r that is allocated to all configurations before some are discarded; a larger value of n corresponds to a smaller r and hence more aggressive early-stopping
  + There are two components to Hyperband;
    1. the inner loop invokes Successive Halving for fixed values of n and r and
    2. the outer loop iterates over different values of n and r

## Libraries and Framework for optimizing hyperparams

* SMAC
  + SMAC: Sequential Model-based Algorithm Configuration
  + <http://www.cs.ubc.ca/labs/beta/Projects/SMAC/>
  + (Java)
* [spearmint](https://github.com/JasperSnoek/spearmint)
  + Spearmint is a package to perform Bayesian optimization
  + <https://github.com/JasperSnoek/spearmint>
  + (Python)
* HyperOpt
  + <http://hyperopt.github.io/hyperopt/>
  + (Python)
* REMBO
  + <https://github.com/ziyuw/rembo>
  + (Matlab)
* HPOLib
  + It provides a common interface to three state of the art hyperparameter optimization packages: SMAC, spearmint and hyperopt
  + <https://github.com/automl/HPOlib>
* Google Vizier
  + A Service for Black-Box Optimization
  + <https://ai.google/research/pubs/pub46180>
  + <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/46180.pdf>
* BOHB
  + ROBUST AND EFFICIENT HYPERPARAMETER OPTIMIZATION AT SCALE
  + <https://www.automl.org/blog_bohb/>
* auto-sklearn
  + *auto-sklearn* is an automated machine learning toolkit and a drop-in replacement for a scikit-learn estimator:
  + <https://automl.github.io/auto-sklearn/master/>

## References

* Gaussian Multivariate Distribution

<http://cs229.stanford.edu/section/gaussians.pdf>

* Gaussian Process
* <http://cs229.stanford.edu/section/cs229-gaussian_processes.pdf>
* http://www.stat.columbia.edu/~rdavis/papers/VAG002.pdf
* <https://www.youtube.com/watch?v=mqOdIYxxNCs>

* Bayesian Optimization
* <https://arxiv.org/pdf/1012.2599.pdf>
* <https://www.cs.ox.ac.uk/people/nando.defreitas/publications/BayesOptLoop.pdf>
* https://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf
* <https://www.youtube.com/watch?v=-Ekte3my510>
* <https://www.youtube.com/watch?v=vz3D36VXefI>
* <https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f>
* Black Box Optimization in machine Learning
* <https://engineering.lehigh.edu/sites/engineering.lehigh.edu/files/_DEPARTMENTS/ise/pdf/tech-papers/17/17T_005.pdf>
* Bayesian Optimization Implementation Example
* <http://krasserm.github.io/2018/03/19/gaussian-processes/>
* <https://github.com/joefarned/SparkGaussianProcess/blob/master/GaussianProcess.scala>
* Hyperband
* <https://arxiv.org/pdf/1603.06560.pdf>
* Distributed Hyperparameter optimization
* <http://www.cs.ubc.ca/labs/beta/Projects/SMAC/papers/parallel_smac_pre_conference_CRC.pdf>
* Sequential Model-Based Optimization for General Algorithm Configuration
* <https://www.cs.ubc.ca/~hutter/papers/10-TR-SMAC.pdf>
* <http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf>
* Introduction
  + <https://arimo.com/data-science/2016/bayesian-optimization-hyperparameter-tuning/>