applied sciences

MDPI

Article

Data-Driven Yield Estimation and Maximization Using Bayesian Optimization Under Uncertainty

Kei Sano [1,*], Daiki Kawahito [1], Yukiya Saito [1], Hironori Moki [1] and Dragan Djurdjanovic [2]

[1] Tokyo Electron Ltd., Tokyo 107-6325, Japan
[2] NSF Engineering Research Center, The University of Texas at Austin, Austin, TX 78712, USA
* Correspondence: kei.sano@tel.com

Featured Application

This paper applies yield estimation and optimization with minimal experiments and stringent specifications in the field of quality control.

Abstract

In this paper, we propose a novel method which utilizes samples of measured product quality characteristics to efficiently estimate the probabilities of those quality characteristics being within the desired specifications and, consequently, the process yield. Specifically, when dealing with 1D Gaussian distributions, we formally prove that the proposed yield estimator asymptotically gives a lower Mean Squared Error compared to the best unbiased estimator. In order to enable maximization of yield, this novel estimator is incorporated into the framework of Bayesian Optimization which iteratively seeks controllable tool parameters under which the outgoing product yield is maximized. The newly proposed yield maximization method is demonstrated in an application involving high-fidelity simulations of a reactive ion etch chamber, a tool component commonly used in semiconductor manufacturing. The aim of these simulations was to rapidly and reliably determine tool parameters that maximize the probability of delivering desired plasma density characteristics under stochastic variations in chamber conditions. The novel yield estimation and optimization methods show superiority when the number of experimental observations is limited and the distributions of outgoing product characteristics can be approximated well by a Gaussian distribution.

Keywords: Bayes procedures; optimization methods; probability; simulation; statistics

1. Introduction

In advanced manufacturing, product specifications have become increasingly stringent, and relevant tools must be meticulously regulated to ensure that the outgoing product quality meets these specifications. However, even when a tool operates under identical control parameter settings, the quality of fabricated products usually still varies. As a result, some portion of the outgoing products may potentially be defective. This variability arises because tools are influenced by numerous environmental factors that cannot be observed or controlled. To minimize defective quality and maximize yields, it is crucial to efficiently address the uncertainties arising from these uncontrollable and unobservable factors.

Problems in yield assessment vary across industry sectors, including additive manufacturing [1], semiconductor fabrication [2,3], and the pharmaceutical industry [4,5]. A good yield is qualitatively defined as having quality characteristics of outgoing products that

Check for updates

Received: 23 December 2025
Revised: 10 March 2026
Accepted: 11 March 2026
Published: 26 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

match the desired values with minimal variation. On a quantitative level, various formulations for defining yield have been considered, such as the sum of bias and variance [6,7] or the probability that the quality characteristics meet the specified constraints [8–12]. In cases where quality characteristics are multivariate, multi-criteria frameworks may be employed to define yield, as in [13,14]. In this manuscript, we define yield as the probability that the quality characteristic falls within the predefined specification limits (it should be noted that even when quality characteristics are multidimensional, yield defined as above is a scalar value), as discussed in [9,10].

A common approach is to treat yield as a Bernoulli parameter and estimate it via the maximum likelihood as in [9,11]. These studies regard the maximum likelihood estimation as a method using traditional Monte Carlo techniques. They propose the use of efficient sampling methods, such as importance sampling [11] or the cross-entropy method [9], each aiming for the highest possible accuracy using as few samples as possible. Nevertheless, improvement in sampling methods is only applicable when the probability density function of the quality characteristics is known. An alternative strategy for efficiently estimating yields is to assume a prior distribution for the parameter of a Bernoulli distribution and use samples to iteratively update that distribution. The study reported in [12] suggests employing a Beta distribution as the prior distribution of the Bernoulli parameter. While this approach can achieve high accuracy when provided with a precise initial yield estimation, it is often unreasonable to assume that such accurate yield estimations are available, especially without significant gathering of prior observations of the quality characteristics. Therefore, while leveraging prior knowledge can enhance accuracy with a limited number of samples, it is often preferable for estimation methods to operate without any assumptions about the true yield value, especially when one is developing a new process—i.e., when such knowledge is limited or does not exist.

To address this need, this article proposes estimating the yield by assuming and exploiting knowledge of the form of the underlying distribution of the product quality, with the standard assumption of independent and identically distributed (i.i.d.) samples. The main idea is that when one does have knowledge about the distribution form, one can first estimate its parameters and then subsequently calculate the corresponding probability of product quality falling within the desired specifications. Note that the estimator based on the Bernoulli distribution is the best unbiased estimator whose variance attains the Cramér–Rao bound [15]; however, it does not utilize any knowledge about the underlying distribution of the quality characteristic. Although the proposed estimator is generally biased, it can achieve a lower Mean Squared Error (MSE) due to the trade-off between bias and variance. In this article, we will formally prove that, under the assumptions that a scalar (univariate) quality output follows a Gaussian distribution and the desired quality specification area is an interval, the estimator based on the Gaussian distribution is asymptotically superior to the Bernoulli-based estimator in terms of the MSE metric.

Besides yield estimation from the product quality measurements, utilization of those measurements to adjust controllable parameters of the relevant manufacturing tool and thus maximize the outgoing product yield is a highly impacting, complementary problem. Such sampling-driven optimization of yield is a classical problem of optimization under uncertainty and has been examined in numerous manufacturing sectors, including pharmaceutical manufacturing [16,17], the petrochemical industry [18,19], circuit design [10,20], and material design [21], to name a few.

Regardless of the application area, the success metric for any sampling-driven yield optimization approach is the ability to determine controllable parameter settings that maximize the outgoing yield, while using as few samples of the outgoing product quality as possible. One highly promising approach for addressing this classical problem of

optimization under uncertainty is Bayesian Optimization (BO). In the field of BO, various forms of uncertainties and objective functions have been considered, as discussed in [22–27]. Similar problem formulations to the one considered in this article are proposed in [28,29], where the authors proposed a method based on a criterion they named *probability threshold robustness measure*, with the underlying assumption that the environmental factors are observable. However, in many situations, the environmental factors are not observable, which necessitates a different approach; this need will be addressed in this paper. More specifically, to maximize the probability that the quality characteristic falls within the desired specifications, we propose incorporating the novel yield estimator introduced in this paper into a BO framework. Unlike previous approaches modeling empirical probabilities directly such as [27], our method employs Gaussian Process Regression (GPR) to model the parameters of the quality distribution. To address noise effects caused by the unobservable and uncontrollable environmental factors, a repeated experimentation strategy will be employed, similar to how BO was performed in [25,27].

In summary, the contributions of this article are as follows:

- The introduction of a yield estimator based on the estimates of the mean and variance parameters of the distribution characterizing the outgoing product quality which are then used to estimate the probability that the outgoing quality falls within a desired specification interval.
- A mathematical proof of asymptotic superiority in terms of MSE of the aforementioned yield estimator compared to the traditionally utilized yield estimator based on estimating the Bernoulli distribution parameter under the assumption that the outgoing product quality variables follow an uncorrelated Gaussian distribution.
- The introduction of a BO algorithm which uses the aforementioned mean and variance parameter-based estimator of yield to tune controllable process parameters in a way that maximizes the probability that the random variable characterizing the outgoing product quality falls within the desired specifications.

The newly proposed approaches to yield estimation and optimization were evaluated in simulations of a plasma chamber reactor, which is a highly complex system commonly used in semiconductor manufacturing. The aim was to use the novel methods to iteratively drive the controllable tool parameters towards a setting at which key plasma parameters could be kept within the desired specifications with maximum probability, given the inherent uncertainties present in the plasma chamber reactor.

The remainder of this article is structured as follows: In Section 2.1, we introduce a yield estimator based on estimating the mean and variance of the distribution of the outgoing product quality, and we introduce several theorems characterizing that estimator and proving its superiority over the Bernoulli-based yield estimator when the outgoing product quality characteristics follow an uncorrelated Gaussian distribution. In Section 2.2, we formalize the optimization problem for sampling-driven maximization of yield and propose two BO algorithms to solve it—one using the newly proposed yield estimator, and the other employing the traditional yield estimator. In Section 3.1, we comparatively evaluate the performance of the yield estimators. First, we examine problems where quality characteristics follow Gaussian distributions and empirically validate the theorems presented in Section 2.1. Second, we estimate the probability of plasma parameters falling within the specifications in the presence of random condition variations in a plasma reactor, with plasma behaviors being simulated using a commercial high-fidelity simulator. Similar to Section 3.1, in Section 3.2, we assess the effectiveness of the proposed BO-based yield maximization algorithms for scenarios where the outgoing quality characteristics follow Gaussian distributions, as well as for a scenario in which that is not the case. Finally, Section 4 concludes this article and offers suggestions for possible future work.

2. Materials and Methods

2.1. Estimation of Yields

A. Problem formulation

Let $y \in \mathrm{R}$ be a random variable representing a quality characteristic and let $S = [a, b]$ denote the desired specification interval for y. The yield is defined as $p_* = Pr(y \in [a, b])$, i.e., as the probability that the quality characteristic falls within the desired specification limits. Suppose y follows a Gaussian distribution with unknown mean μ_* and variance σ_*^2. In addition, let us observe N independent samples $\{y_1, \ldots, y_N\}$ of the outgoing quality, where

$$y_n \overset{i.i.d}{\sim} \mathcal{N}(\mu_*, \sigma_*^2),\ n \in \{1, 2, \ldots, N\}. \tag{1}$$

The objective is to accurately estimate the probability p_* via some estimator $\hat{p}$ from the samples $\{y_1, \ldots, y_N\}$. The accuracy of the estimator will be assessed via its MSE, defined as

$$\mathbb{E}_{\{y_1,\ldots,y_N\}}\left[(p_* - \hat{p})^2\right]. \tag{2}$$

Obviously, estimators with a smaller MSE are generally regarded as better estimators. Note that our formulation of yield differs from outage error probability discussed in estimation theory as in [30]. Outage error probability is the probability that the error of the estimate exceeds a certain threshold. In our formulation, yield itself is the estimation target expressed as a probability.

B. Yield estimator based on estimating Bernoulli distribution parameter

By focusing on whether a sample y_n satisfies the specifications or not, yield estimation can be formulated as the problem of estimating the parameter of Bernoulli distribution, which describes the yield as in [31]. The maximum likelihood is given by

$$\hat{p}_1 = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_S(y_n), \tag{3}$$

where $\mathbf{1}_S$ is the indicator function defined as

$$\mathbf{1}_S(y) = \begin{cases} 1 & y \in S \\ 0 & y \notin S \end{cases}. \tag{4}$$

The MSE of $\hat{p}_1$ is easily derived as

$$\mathbb{E}\left[(\hat{p}_1 - p_*)^2\right] = \frac{1}{N} p_*(1 - p_*). \tag{5}$$

Note that $\hat{p}_1$ is the best unbiased estimator in the sense that it achieves the minimum Mean Squared Error.

C. Yield estimator based on estimates of mean and variance parameters

Under the assumption that y follows a Gaussian distribution, we can estimate the parameters of that distribution and subsequently estimate the required probability p_*. Let us use the unbiased estimators for the mean and variance of Gaussian distributions, which are, respectively,

$$\begin{aligned} \hat{\mu} &= \frac{1}{N}\sum_{n=1}^{N} y_n \\ \hat{\sigma}^2 &= \frac{1}{N-1}\sum_{n=1}^{N} (y_n - \hat{\mu})^2. \end{aligned} \tag{6}$$

Using those estimators, we can construct the estimator of the probability p_* as follows:

$$\begin{aligned} \hat{p}_2 &= Pr\left(y \in [a,b] | y \sim \mathcal{N}\left(\hat{\mu}, \hat{\sigma}^2\right)\right) \\ &= \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \int_a^b exp\left(-\frac{1}{2\hat{\sigma}^2}(z-\hat{\mu})^2\right) dz. \end{aligned} \tag{7}$$

Here, let us define a function $F(a,b,\mu,V)$ as

$$F(a,b,\mu,V) = \frac{1}{2}\left(erf\left(\frac{b-\mu}{\sqrt{2V}}\right) - erf\left(\frac{a-\mu}{\sqrt{2V}}\right)\right), \tag{8}$$

where $\mathrm{erf}(\cdot)$ is the error function defined by

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt. \tag{9}$$

Then, the estimator $\hat{p}_2$ can be expressed as

$$\hat{p}_2 = F\left(a,b,\hat{\mu},\hat{\sigma}^2\right). \tag{10}$$

In order to calculate the MSE of $\hat{p}_2$, we use the fact that $\hat{\mu}$ and $\hat{\sigma}^2$ independently follow $\mathcal{N}\left(\mu_*, \frac{\sigma_*^2}{N}\right)$ and $\frac{\sigma_*^2}{N-1}\chi^2_{N-1}$, respectively (e.g., see Chapter 6 of [32]). Then, the MSE of the estimator $\hat{p}_2$ is calculated as follows:

$$\mathbb{E}[(\hat{p}_2 - p_*)^2] = \mathbb{E}_{Y,Z}[F(a,b,Z,Y)^2] - 2p_*\mathbb{E}_{Y,Z}[F(a,b,Z,Y)] + p_*^2 \tag{11}$$

where distributions of Y and Z satisfy

$$\frac{(N-1)}{\sigma_*^2} Y \sim \chi^2_{N-1} \; ; \; Z \sim \mathcal{N}\left(\mu_*, \frac{\sigma_*^2}{N}\right). \tag{12}$$

Although the MSE of the estimates is mainly discussed in the main text, confidence intervals of the estimates are also important in applications. The derivation of the confidence intervals is explained in Appendix A.

D. Comparison of MSEs of two estimators

If $[a,b], \mu_*, \sigma_*^2$ and N are known, then the performance of the estimators becomes clear via calculation of (5) and (11). However, in reality, μ_* and σ_*^2 are never known. Here, we formulate and prove the theorem that shows the following: when the number of samples N is sufficiently large, metric (11) is smaller than (5) for any $[a,b]$, μ_* and σ_*^2.

First, let us apply the delta method (numerous statistics textbooks describe the delta method; e.g., one can refer to [33] or [34] for details of this method) to the MSE of the estimator $\hat{p}_2$ in order to derive its asymptotic behavior in the following Lemma.

Lemma 1 (Asymptotic MSE of $\hat{p}_2$)**.** *Suppose that we are given the specification interval $[a,b]$ and N independent samples $y_1, \ldots, y_N$, where $y_n \sim \mathcal{N}\left(\mu_*, \sigma_*^2\right)$, $n \in \{1,2,\ldots,N\}$. Let p_* denote the target probability that $y \in [a,b]$ if $y \sim \mathcal{N}\left(\mu_*, \sigma_*^2\right)$, i.e., let*

$$p_* = Pr\left(y \in [a,b] | y \sim \mathcal{N}\left(\mu_*, \sigma_*^2\right)\right). \tag{13}$$

Then, for the Gaussian parameter-based estimator $\hat{p}_2$ defined in (7), the corresponding MSE of $\hat{p}_2$ can be expressed as follows:

$$\begin{aligned}
&\mathbb{E}\left[(\hat{p}_2 - p_*)^2\right] \\
&= \frac{1}{4\pi\sigma_*^2 N} e^{-\frac{1}{\sigma_*^2}(a-\mu_*)^2}\left(2\sigma_*^2 + (a-\mu_*)^2\right) \\
&+ \frac{1}{4\pi\sigma_*^2 N} e^{-\frac{1}{\sigma_*^2}(b-\mu_*)^2}\left(2\sigma_*^2 + (b-\mu_*)^2\right) \\
&- \frac{1}{4\pi\sigma_*^2 N} e^{-\frac{1}{2\sigma_*^2}((a-\mu_*)^2+(b-\mu_*)^2)}\left(4\sigma_*^2 + 2(a-\mu_*)(b-\mu_*)\right) \\
&+ \mathcal{O}\left(\frac{1}{N^2}\right)
\end{aligned} \tag{14}$$

For details of the derivation, please refer to Appendix B.

Theorem 1 (comparison of asymptotic MSEs)**.** *Under the same premises as in Lemma 1, we have*

$$\lim_{N\to\infty} \frac{\mathbb{E}\left[(\hat{p}_1 - p_*)^2\right] - \mathbb{E}\left[(\hat{p}_2 - p_*)^2\right]}{\frac{1}{N}} \geq 0. \tag{15}$$

The strategy for the proof of this theorem is to observe the difference in the MSE portions associated with the factor $1/N$ as a function of parameters a and b and to compute extrema of this function. The details of the proof are enclosed in Appendix C.

The following corollary can be easily derived based on the inequality in (15) in Theorem 1.

Corollary of Theorem 1 (comparison of asymptotic MSEs)**.** *Asymptotically, the MSE of $\hat{p}_1$ is equal to or greater than that of $\hat{p}_2$.*

It should be noted that in the process of proving Theorem 1, the following proposition is derived.

Proposition 1 (the maximum difference between MSEs of $\hat{p}_1$ and $\hat{p}_2$ for a centered manufacturing process)**.** *Consider the case where the true mean parameter corresponds to the center of the specification interval, i.e., $[a,b] = \left[\mu_* - w\sqrt{\sigma_*^2}, \mu_* + w\sqrt{\sigma_*^2}\right]$. Under the same premises as Lemma 1 and Theorem 1, the difference in asymptotic MSEs between yield estimators $\hat{p}_1$ and $\hat{p}_2$ reaches its maximum when $[a,b] = \left[\mu_* - w_*\sqrt{\sigma_*^2}, \mu_* + w_*\sqrt{\sigma_*^2}\right]$, where w_* is the solution of*

$$\sqrt{2\pi}\left(1 - 2erf\left(\frac{w_*}{\sqrt{2}}\right)\right) - \left(2w_* - 2w_*^3\right)e^{\frac{-w_*^2}{2}} = 0. \tag{16}$$

That solution can be determined to be $w_* \approx 0.48$ and, in that case, the maximum value for $\lim_{N\to\infty} \frac{\mathbb{E}\left[(\hat{p}_1 - p_*)^2\right] - \mathbb{E}\left[(\hat{p}_2 - p_*)^2\right]}{1/N}$ can be found to be

$$\lim_{N\to\infty} \frac{\mathbb{E}\left[(\hat{p}_1 - p_*)^2\right] - \mathbb{E}\left[(\hat{p}_2 - p_*)^2\right]}{1/N} \fallingdotseq 0.17.$$

E. Yield estimator based on parameters of uncorrelated multivariate Gaussian distribution

Theorem 1 can be easily extended to the case of multivariate quality characterization, with quality variable vectors following an uncorrelated multivariate Gaussian distribution. Suppose that a vector of quality characteristics $\boldsymbol{y} \in \mathrm{R}^D$ follows an uncorrelated multivariate Gaussian distribution as below,

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_*, diag\left(\sigma_{*,1}^2, \sigma_{*,2}^2, \sigma_{*,D}^2\right)\right). \tag{17}$$

The yield estimator based on the Bernoulli distribution parameter is the same as (**3**), though the specification area $\boldsymbol{S}$ is now defined as a set product over each dimension of the specification area, i.e.,

$$\boldsymbol{S} = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_D, b_D]. \tag{18}$$

On the other hand, the yield estimator based on the estimates of the mean and variance components of the quality variable vector $\boldsymbol{y}$ can be expressed as

$$\hat{p}_2 = Pr\left(\boldsymbol{y} \in \boldsymbol{S} \middle| \boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}, diag\left(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_D^2\right)\right)\right), \tag{19}$$

where

$$\begin{aligned} \hat{\mu}_d &= \frac{1}{N}\sum_{n=1}^{N} y_{n,d} \\ \hat{\sigma}_d^2 &= \frac{1}{N-1}\sum_{n=1}^{N} (y_{n,d} - \hat{\mu}_d)^2. \end{aligned} \tag{20}$$

Extension of Theorem 1 for this case is given below.

Theorem 2 (comparison of asymptotic MSEs for yield estimators based on samples of quality characteristics drawn from a multivariate uncorrelated Gaussian distribution)**.** *Let the desired specification area be* $\boldsymbol{S} = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_D, b_D]$ *and let us observe* N *independent vectorial samples* $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ *drawn from the multivariate uncorrelated Gaussian distribution* $\mathcal{N}\left(\boldsymbol{\mu}_*, diag\left(\sigma_{*,1}^2, \sigma_{*,2}^2, \sigma_{*,D}^2\right)\right)$*. Let* p_* *denote the yield we are trying to estimate, i.e.,*

$$p_* = Pr\left(\boldsymbol{y} \in \boldsymbol{S} \middle| \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_*, diag\left(\sigma_{*,1}^2, \sigma_{*,2}^2, \sigma_{*,D}^2\right)\right)\right) \ . \tag{21}$$

Then, the Bernoulli parameter-based yield estimator $\hat{p}_1$ *defined by* (3) *and the Gaussian parameter-based estimator* $\hat{p}_2$ *defined by* (19) *satisfy the following:*

$$\lim_{N\to\infty} \frac{\mathbb{E}\left[(\hat{p}_1 - p_*)^2\right] - \mathbb{E}\left[(\hat{p}_2 - p_*)^2\right]}{1/N} \geq 0$$

The approach to the proof is to independently consider D estimators for each dimension of the quality variable vector and to utilize the result of Theorem 1 for each of those dimensions.

F. Discussion

The proposed methodology for estimating yield based on estimating the parameters of the underlying distribution from the available samples is applicable to any distribution form, provided that a parametric distribution can be assumed and that the probability over the specification region can be computed. Nevertheless, formally proving the benefits of exploiting the distribution form and estimating its parameters may be challenging for correlated multivariate Gaussian distributions or for non-Gaussian distributions. In particular, when the distribution is heavy-tailed, estimating its parameters is difficult, and parameter estimation accuracy strongly affects yield estimation. A future challenge is to improve and provide guarantees for yield estimation by leveraging theoretical guarantees on parameter estimation (e.g., [35]).

In this article, we indeed formally only analyze the case when the underlying distribution of the outgoing quality is of a multivariate uncorrelated Gaussian form. However, there are application areas, such as semiconductor manufacturing processes, where the random

behavior of outgoing quality could significantly deviate from a Gaussian distribution form. In these cases, the superiority or inferiority of $\hat{p}_1$ and $\hat{p}_2$ depends on the sample size N, the actual underlying distribution form, and the specification interval.

Generally, $\hat{p}_2$ tends to be superior to $\hat{p}_1$ for small sample sizes N. This is because the available empirical clues for estimating yield probability from the Bernoulli distribution describing whether samples are falling into the specification area or not are insufficient when the sample size is small. In such cases, although not entirely accurate, assumption of Gaussianity helps $\hat{p}_2$ compensate for the scarcity of samples. On the other hand, when N is large, $\hat{p}_1$ will be superior to $\hat{p}_2$, in general, because $\hat{p}_1$ is a consistent estimator [36], while $\hat{p}_2$ is not. To select the preferred estimator, the sample size N for which the superiority of $\hat{p}_2$ to $\hat{p}_1$ flips is an important threshold. Unfortunately, derivation of that threshold is practically impossible without knowledge of the true distribution. Nevertheless, based on Theorem 1, it is reasonable to claim that the more closely the distribution of the outgoing quality resembles a Gaussian distribution, the more pronouncedly $\hat{p}_2$ is superior to $\hat{p}_1$ for larger N. Note that if Gaussianity is strongly violated and the true distribution is multimodal, the relative performance of $\hat{p}_2$ and $\hat{p}_1$ is expected to be reversed for very small sample sizes (e.g., $N = 2$).

From a practical viewpoint, selecting a subset of important dimensions to measure—rather than reducing the sample size N—can reduce measurement cost while maintaining estimation accuracy. Simple heuristics based on collinearity or formal subset selection frameworks [37] can be used. Combining these selection methods with the newly proposed yield estimator may improve the trade-off between measurement cost and estimation accuracy, particularly when the selected features are approximately regarded as uncorrelated.

2.2. Bayesian Optimization for Maximizing Yields

A classical problem in manufacturing is determining the settings of controllable parameters of the tool which maximize the outgoing yield. In this section, we propose a yield optimization method which incorporates the yield estimators described in the previous section into a Bayesian Optimization (BO) framework.

A. Problem formulation

We consider the optimization problem

$$\max_{x \in \mathcal{X}} Pr(\boldsymbol{y} \in \boldsymbol{S} | \boldsymbol{y} \sim \mathcal{D}(\boldsymbol{x})), \tag{22}$$

where $\boldsymbol{x}$ is a vector of control parameters, $\mathcal{D}(\boldsymbol{x})$ denotes the control parameter-dependent distribution of the vector of quality characteristics $\boldsymbol{y}$, $\mathcal{X}$ is the domain of control parameters and $\boldsymbol{S}$ is the desired multivariate specification area.

For a given candidate solution $\boldsymbol{x}_i$ of the vector of control parameters, let us assume that we can observe N independent samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_N$, of the vector of quality characteristics, where $\boldsymbol{y}_n \sim \mathcal{D}(\boldsymbol{x}_i)$. The objective is to iteratively move towards a solution $\hat{\boldsymbol{x}}$ such that $\Pr(\boldsymbol{y} \in \boldsymbol{S} | \boldsymbol{y} \sim \mathcal{D}(\hat{\boldsymbol{x}}))$ becomes as high as possible and to do that in as few iterations as possible (i.e., using as few experiments as possible). Algorithm 1 shown below summarizes the proposed optimization procedure. In each iteration, it encompasses two major functionalities: (i) suggestion of the next iteration's control parameter (step 3 in Algorithm 1), and (ii) estimation of the control parameter settings which maximize the outgoing yield based on observations available at that iteration (step 6 in Algorithm 1). Details of the implementation of this procedure will be discussed below.

Algorithm 1 Bayesian procedure for solving optimization problem

Input: An algorithm $\mathcal{A}$ initial observations $\mathcal{H}_{\text{init}}$,
the number of observations per iteration N,
and maximum iteration $I_{\max}$.

Output: Estimated best control parameters for each iteration $\hat{X} = (\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_{I_{\max}})$

1: $\mathcal{H} \leftarrow \mathcal{H}_{\text{init}}$
2: **for** $i = 1$ to $I_{\max}$ **do**
3: Algorithm $\mathcal{A}(\mathcal{H})$ suggests the next control parameter $\boldsymbol{x}_i$.
4: Observe the N outputs for $\boldsymbol{x}_i$, $\left(\boldsymbol{x}_i, \left(\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_i^{(2)}, \ldots, \boldsymbol{y}_i^{(N)}\right)\right)$.
5: Update the history of observations $\mathcal{H} \leftarrow \mathcal{H} \bigcup \left(\boldsymbol{x}_i, \left(\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_i^{(2)}, \ldots, \boldsymbol{y}_i^{(N)}\right)\right)$
6: Algorithm $\mathcal{A}(\mathcal{H})$ suggests the best control parameter, $\hat{\boldsymbol{x}}_i$.
7: $\hat{X} \leftarrow \hat{X} \bigcup \hat{\boldsymbol{x}}_i$.
8: **end for**

B. Bayesian Optimization using GPR modeling empirical yields

In this subsection, we describe the BO algorithm for solving (22) which relies on the GPR estimator of the yield, with yields being modeled from experiments using the empirical estimator of the Bernoulli distribution parameter, as described in Section 2.1-B. Let us define $f_*(\boldsymbol{x}) = Pr(\boldsymbol{y} \in \boldsymbol{S} | \boldsymbol{y} \sim \mathcal{D}(\boldsymbol{x}))$. Suppose that we have an available record of iterations of control parameter vectors $\boldsymbol{x}$ and the corresponding observations of the outgoing quality vectors $\boldsymbol{y}$ in the form

$$\begin{aligned} \mathcal{H} \quad = \Big\{ & \left(\boldsymbol{x}_1, \left(\boldsymbol{y}_1^{(1)}, \boldsymbol{y}_1^{(2)}, \ldots \boldsymbol{y}_1^{(N_1)}\right)\right) \ldots \\ & \left(\boldsymbol{x}_i, \left(\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_i^{(2)}, \ldots \boldsymbol{y}_i^{(N_i)}\right)\right) \ldots \\ & \left(\boldsymbol{x}_I, \left(\boldsymbol{y}_I^{(1)}, \boldsymbol{y}_I^{(2)}, \ldots \boldsymbol{y}_I^{(N_I)}\right)\right) \Big\}. \end{aligned} \tag{23}$$

For each setting of control parameters $\boldsymbol{x}_i$, we can use quality observations $\boldsymbol{y}_i^{(k)}, k = 1, 2, \ldots, N_i$ to empirically estimate the yield under those control settings using the simple formulation

$$\bar{f}(\boldsymbol{x}_i) = \frac{1}{N_i} \sum_{n=1}^{N_i} \mathbf{1}_{\boldsymbol{S}}\left(\boldsymbol{y}_i^{(n)}\right). \tag{24}$$

Then, one can model the yield as it depends on the control parameter vector $\boldsymbol{x}$ by training a GPR to model $f_*(\boldsymbol{x})$ using *control vector–yield estimate* pairs $(\boldsymbol{x}_1, \bar{f}(\boldsymbol{x}_1)), (\boldsymbol{x}_2, \bar{f}(\boldsymbol{x}_2)), \ldots, (\boldsymbol{x}_I, \bar{f}(\boldsymbol{x}_I))$ obtained from the available data $\mathcal{H}$. Let $\text{GPR}_f(\boldsymbol{x}|\boldsymbol{f})$ denote the GPR function modeling the expected value and variance of the yield $f(\boldsymbol{x})$, given observations $\boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots f(\boldsymbol{x}_I)]^\top$ of the yield at $\boldsymbol{x}_i, i = 1, 2, \ldots, I$. Let us denote $\bar{\boldsymbol{f}} = \left[\bar{f}(\boldsymbol{x}_1), \bar{f}(\boldsymbol{x}_2), \ldots, \bar{f}(\boldsymbol{x}_I)\right]^\top$. From the set of control parameters $\mathcal{X}_I = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_I\}$, let us select the vector $\hat{\boldsymbol{x}}$ which maximizes the posterior expectation of the yield, as modeled by $\text{GPR}_f\left(\boldsymbol{x}|\bar{\boldsymbol{f}}\right)$, i.e.,

$$\hat{\boldsymbol{x}} = \underset{\mathcal{X}_I}{argmax}\, \mathbb{E}_f\left[f \middle| f \sim \text{GPR}_f\left(\boldsymbol{x}|\bar{\boldsymbol{f}}\right)\right], \tag{25}$$

and let $\hat{f}_*$ be the expected yield, as modeled by $\text{GPR}_f\left(\hat{\boldsymbol{x}}|\bar{\boldsymbol{f}}\right)$, i.e.,

$$\hat{f}_* = \mathbb{E}_f\left[f \middle| f \sim \text{GPR}_f\left(\hat{\boldsymbol{x}}|\bar{\boldsymbol{f}}\right)\right]. \tag{26}$$

Now, as in any BO procedure, the vector of control parameters $\boldsymbol{x}$ in the next iteration needs to be pursued via a trade-off between exploration and exploitation enabled by that iteration [38]. Following the approach of [16,21], one can use Expected Improvement (EI) as the acquisition function and obtain the next candidate vector of controllable parameters $\boldsymbol{x}_{I+1}$ as

$$\boldsymbol{x}_{I+1} = \underset{\boldsymbol{x}\in\mathcal{X}}{argmax}\ \mathrm{EI}(\boldsymbol{x}) \quad , \tag{27}$$

where

$$\mathrm{EI}(\boldsymbol{x}) = \mathbb{E}_f\left[max\left(f-\hat{f}_*,0\right)\middle| f\sim \mathrm{GPR}_f\left(\boldsymbol{x}\middle|\bar{\boldsymbol{f}}\right)\right] \tag{28}$$

is the acquisition function which, for a given vector of control parameters $\boldsymbol{x}$, expresses the expected improvement in the yield relative to the current best expected yield $\hat{f}_*$.

Nevertheless, the EI acquisition function (28) does not account for the uncertainties of the observations $\bar{\boldsymbol{f}}$. In order to address this problem, following [39], in this paper we employ the so-called Noisy Expected Improvement (NEI) acquisition function which incorporates the uncertainty of observations $\bar{\boldsymbol{f}}$ into the selection of the next candidate solution $\boldsymbol{x}_{I+1}$, thus achieving general superiority over the EI acquisition function-based iterations described by (26). For completeness, NEI acquisition function-based selection of the next iteration for the candidate solution $\boldsymbol{x}_{I+1}$ is given below.

Let us introduce EI acquisition function $\alpha_{\mathrm{EI}}(\boldsymbol{x}|\boldsymbol{f})$ for any set of observations $\boldsymbol{f}$ of yield for control parameter vectors $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_I$, i.e.,

$$\alpha_{\mathrm{EI}}(\boldsymbol{x}|\boldsymbol{f}) = \mathbb{E}_f\left[max\left(f-\hat{f}_*,0\right)\middle| f\sim \mathrm{GPR}_f(\boldsymbol{x}|\boldsymbol{f})\right]. \tag{29}$$

Obviously, $\mathrm{EI}(\boldsymbol{x}) = \alpha_{\mathrm{EI}}\left(\boldsymbol{x}\middle|\bar{\boldsymbol{f}}\right)$. Let us note that as modeled by $\mathrm{GPR}_f\left(\boldsymbol{x}\middle|\bar{\boldsymbol{f}}\right)$, the vector of yields for control vector $\boldsymbol{x}_1,\boldsymbol{x}_2\ldots,\boldsymbol{x}_I$ can be modeled as a random vector $\boldsymbol{f}_{\mathbf{GPR}} = [f_{\mathrm{GPR}}(\boldsymbol{x}_1)f_{\mathrm{GPR}}(\boldsymbol{x}_2)\ \ldots\ f_{\mathrm{GPR}}(\boldsymbol{x}_I)]^T$ which follows a vectorial normal distribution:

$$\begin{aligned}
&\begin{bmatrix} f_{\mathrm{GPR}}(\boldsymbol{x}_1) \\ \vdots \\ f_{\mathrm{GPR}}(\boldsymbol{x}_I) \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{m}^{f_{\mathrm{GPR}}}, \boldsymbol{S}^{f_{\mathrm{GPR}}}\right) \\
&m_i^{f_{\mathrm{GPR}}} = \mathbb{E}\left[\mathrm{GPR}_f\left(\boldsymbol{x}_i\middle|\bar{\boldsymbol{f}}\right)\right],\ i=1,2,\ldots,I \\
&S_{ij}^{f_{\mathrm{GPR}}} = \mathbb{COV}\left[\mathrm{GPR}_f\left(\boldsymbol{x}_i\middle|\bar{\boldsymbol{f}}\right),\mathrm{GPR}_f\left(\boldsymbol{x}_j\middle|\bar{\boldsymbol{f}}\right)\right],\ i,j=1,2,\ldots,I
\end{aligned} \tag{30}$$

Then, the NEI acquisition function can be defined as

$$\mathrm{NEI}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{f}_{\mathbf{GPR}}}[\alpha_{\mathrm{EI}}(\boldsymbol{x}|\boldsymbol{f}_{\mathbf{GPR}})]. \tag{31}$$

Following [39], Monte Carlo sampling needs to be utilized in order to evaluate the expectation over $\boldsymbol{f}_{\mathbf{GPR}}$ in the $\mathrm{NEI}(\boldsymbol{x})$ acquisition function described by (31). Finally, selection of the next candidate vector of controllable parameters is accomplished by

$$\boldsymbol{x}_{I+1} = \underset{\boldsymbol{x}\in\mathcal{X}}{argmax}\ \mathrm{NEI}(\boldsymbol{x}). \tag{32}$$

C. Bayesian Optimization using GPR modeling parameters of the distribution of quality characteristics

Let us consider the situation where for a given setting of control parameters $\boldsymbol{x}$, the outgoing quality characteristics y follow a normal distribution (For simplicity, in this section, we consider the case in which quality output y is a scalar. Please note that following the discussions in Section 2.1-E, one can straightforwardly extend the approach described in this section to the multivariate case when the vector of quality characteristics $\boldsymbol{y}$ follows an

uncorrelated multivariate Gaussian distribution.) $\mathcal{N}\left(\mu_*(\boldsymbol{x}), \sigma_*^2(\boldsymbol{x})\right)$. In that case, a control parameter vector $\boldsymbol{x}$ determines the mean and variance of the distribution $\mathcal{D}(\boldsymbol{x})$, and the yield for that control parameter setting can be written as

$$Pr\left(y \in S \middle| y \sim \mathcal{N}\left(\mu_*(\boldsymbol{x}), \sigma_*^2(\boldsymbol{x})\right)\right). \tag{33}$$

Based on the results from Section 2.1, which show that the yield estimator relying on the estimation of Gaussian distribution parameters outperforms that relying on the empirical estimation of the Bernoulli distribution parameter, let us apply the GPR framework to model the mean and variance parameters of the distribution of the outgoing product quality with the available data $\mathcal{H}$, i.e.,

$$\begin{aligned} \bar{\mu}(\boldsymbol{x}_i) &= \frac{1}{N_i}\sum_{n=1}^{N_i} y_i^{(n)} \\ \bar{\sigma}^2(\boldsymbol{x}_i) &= \frac{1}{N_i - 1}\sum_{n=1}^{N_i}\left(y_i^{(n)} - \bar{\mu}(\boldsymbol{x}_i)\right)^2. \end{aligned} \tag{34}$$

Following [40], variances $\bar{\sigma}^2(\boldsymbol{x}_i)$ are transformed using the natural log function in order to ensure that the range for the GPR model estimating those transformed variances comprises the entire set of real numbers. Hence, GPRs modeling the mean and variance parameters will be trained on the pairs of control vectors and estimated distribution parameters as below,

$$\begin{aligned} &\left(\boldsymbol{x}_1, \left(\bar{\mu}(\boldsymbol{x}_1), log\left[\bar{\sigma}^2(\boldsymbol{x}_1)\right]\right)\right), \left(\boldsymbol{x}_2, \left(\bar{\mu}(\boldsymbol{x}_2), log\left[\bar{\sigma}^2(\boldsymbol{x}_2)\right]\right)\right), \ldots, \\ &\left(\boldsymbol{x}_I, \left(\bar{\mu}(\boldsymbol{x}_I), log\left[\bar{\sigma}^2(\boldsymbol{x}_I)\right]\right)\right). \end{aligned} \tag{35}$$

The variance of the estimator $\bar{\mu}(\boldsymbol{x}_i)$ can be expressed and approximated as

$$\begin{aligned} \mathbb{VAR}[\bar{\mu}(\boldsymbol{x}_i)] &= \mathbb{VAR}\left[\frac{1}{N_i}\sum_{n=1}^{N_i} y_i^{(n)}\right] \\ &= \frac{1}{N_i}\sigma_*^2(\boldsymbol{x}_i) \fallingdotseq \frac{1}{N_i}\bar{\sigma}^2(\boldsymbol{x}_i), \end{aligned} \tag{36}$$

while the variance of the estimator $log\left[\bar{\sigma}^2(\boldsymbol{x}_i)\right]$ can be approximated using the delta method as follows [34]:

$$\mathbb{VAR}\left[log\left[\bar{\sigma}^2(\boldsymbol{x}_i)\right]\right] \fallingdotseq \frac{\mathbb{VAR}\left[\bar{\sigma}^2(\boldsymbol{x}_i)\right]}{\left(\mathbb{E}[\bar{\sigma}^2(\boldsymbol{x}_i)]\right)^2} = \frac{2}{N_i - 1}. \tag{37}$$

Let $\mathrm{GPR}_{\mu}(\boldsymbol{x}|\boldsymbol{\mu})$ denote the GPR function modeling the mean parameter $\mu(\boldsymbol{x})$, given the observations of the estimated parameters $\boldsymbol{\mu} = [\mu(\boldsymbol{x}_1), \mu(\boldsymbol{x}_2), \ldots, \mu(\boldsymbol{x}_I)]$. Similarly, let $\mathrm{GPR}_{log[\sigma^2]}(\boldsymbol{x}|\boldsymbol{\theta})$ be the GPR function modeling the logarithm of the true variance parameter $\theta(\boldsymbol{x}) = log\left[\sigma^2(\boldsymbol{x})\right]$, given the observations of the estimated parameters $\boldsymbol{\theta} = \left[log\left[\sigma^2(\boldsymbol{x}_1)\right]\ log\left[\sigma^2(\boldsymbol{x}_2)\right]\ \ldots\ log\left[\sigma^2(\boldsymbol{x}_I)\right]\right]^\top$.

Then, let us denote with $\mathcal{F}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\theta})$ the distribution of the estimated yield, which can be expressed as

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\theta}) = Pr\Bigg(Y \in S \Bigg| Y \sim \mathcal{N}\left(\widetilde{\mu}, \widetilde{\sigma}^2\right)\Bigg) \\ \widetilde{\mu} \sim \mathrm{GPR}_{\mu}(\boldsymbol{x}|\boldsymbol{\mu}) \\ log\left[\widetilde{\sigma}^2\right] \sim \mathrm{GPR}_{log[\sigma^2]}(\boldsymbol{x}|\boldsymbol{\theta})\Bigg), \end{aligned} \tag{38}$$

and let $\bar{\boldsymbol{\mu}} = [\bar{\mu}(\boldsymbol{x}_1), \bar{\mu}(\boldsymbol{x}_2), \ldots, \bar{\mu}(\boldsymbol{x}_I)]^\top$ and $\bar{\boldsymbol{\theta}} = \left[log\left[\bar{\sigma}^2(\boldsymbol{x}_1)\right], log\left[\bar{\sigma}^2(\boldsymbol{x}_2)\right], \ldots, log\left[\bar{\sigma}^2(\boldsymbol{x}_I)\right]\right]^\top$. Now, from the set of control parameters $\mathcal{X}_I = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_I\}$, let us select vector $\hat{\boldsymbol{x}}$ which maximizes the posterior expectation of the yield modeled by $\mathcal{F}(\boldsymbol{x}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\theta}})$, i.e.,

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathcal{X}_I}{argmax} \mathbb{E}_f[f | f \sim \mathcal{F}(\boldsymbol{x}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\theta}})], \tag{39}$$

and let $\hat{f}_{**}$ be the expected value as per model $\mathcal{F}(\hat{\boldsymbol{x}}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\theta}})$, i.e.,

$$\hat{f}_{**} = \mathbb{E}_f[f | f \sim \mathcal{F}(\hat{\boldsymbol{x}}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\theta}})]. \tag{40}$$

The next iteration $\boldsymbol{x}_{I+1}$ of the candidate solution for the vector of control parameters is determined as

$$\boldsymbol{x}_{I+1} \quad = \underset{\boldsymbol{x} \in \mathcal{X}}{argmax} \mathrm{EI}_{\mathcal{F}}(\boldsymbol{x}), \tag{41}$$

where

$$\mathrm{EI}_{\mathcal{F}}(\boldsymbol{x}) \quad = \alpha_{\mathrm{EI}_{\mathcal{F}}}(\boldsymbol{x}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\theta}}) \tag{42}$$

is the acquisition function which, for a given vector of control parameters $\boldsymbol{x}$, expresses the expected improvement in the yield relative to the inferred best expected yield $\hat{f}_{**}$.

A major challenge associated with the calculation of (39) and (41) is that $\mathcal{F}(\boldsymbol{x}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\theta}})$ is not an analytically tractable Gaussian distribution. Therefore, following [40], Monte Carlo sampling from the distribution $\mathcal{F}(\boldsymbol{x}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\theta}})$ is proposed for evaluating the expected value of the yield in (39) and the acquisition function $\mathrm{EI}_{\mathcal{F}}(\boldsymbol{x})$ for any vector $\boldsymbol{x}$ of the control parameters.

Let us now introduce $\alpha_{\mathrm{EI}_{\mathcal{F}}}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\theta})$ for any set of observed estimated parameters $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ for control parameter vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_I$, i.e.,

$$\alpha_{\mathrm{EI}_{\mathcal{F}}}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\theta}) \quad = \mathbb{E}_f\left[max\left(f - \hat{f}_{**}, 0\right) \middle| f \sim \mathcal{F}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\theta})\right] \ . \tag{43}$$

Obviously, $\mathrm{EI}_{\mathcal{F}}(\boldsymbol{x}) = \alpha_{\mathrm{EI}_{\mathcal{F}}}(\boldsymbol{x}|\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\theta}})$. It can be noted that the mean parameters for control vector $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_I$ can be modeled as random vector $[\mu_{\mathrm{GPR}}(\boldsymbol{x}_1), \mu_{\mathrm{GPR}}(\boldsymbol{x}_2), \ldots, \mu_{\mathrm{GPR}}(\boldsymbol{x}_I)]^T$ as below:

$$\begin{bmatrix} \mu_{\mathrm{GPR}}(\boldsymbol{x}_1) \\ \vdots \\ \mu_{\mathrm{GPR}}(\boldsymbol{x}_I) \end{bmatrix} \sim \mathcal{N}(\boldsymbol{m}^{\mu_{\mathrm{GPR}}}, \boldsymbol{S}^{\mu_{\mathrm{GPR}}})$$
$$m_i^{\mu_{\mathrm{GPR}}} = \mathbb{E}\left[\mathrm{GPR}_{\mu}(\boldsymbol{x}_i|\bar{\boldsymbol{\mu}})\right], i = 1, 2, \ldots, I$$
$$S_{ij}^{\mu_{\mathrm{GPR}}} = \mathbb{COV}\left[\mathrm{GPR}_{\mu}(\boldsymbol{x}_i|\bar{\boldsymbol{\mu}}), \mathrm{GPR}_{\mu}(\boldsymbol{x}_j|\bar{\boldsymbol{\mu}})\right]$$
$$i, j = 1, 2, \ldots, I \tag{44}$$

Similarly, the logarithm of variance parameters for control vector $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_I$ can be modeled as random vector $[\theta_{\mathrm{GPR}}(\boldsymbol{x}_1), \theta_{\mathrm{GPR}}(\boldsymbol{x}_2), \ldots \theta_{\mathrm{GPR}}(\boldsymbol{x}_I)]^T$ following the distribution

$$\begin{bmatrix} \theta_{\mathrm{GPR}}(\boldsymbol{x}_1) \\ \vdots \\ \theta_{\mathrm{GPR}}(\boldsymbol{x}_I) \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{m}^{\theta_{GPR}}, \boldsymbol{S}^{\theta_{GPR}}\right)$$
$$m_i^{\theta_{GPR}} = \mathbb{E}\left[\mathrm{GPR}_{log[\sigma^2]}(\boldsymbol{x}_i|\bar{\boldsymbol{\theta}})\right], i = 1, 2, \ldots, I$$
$$S_{ij}^{\theta_{GPR}} = \mathbb{COV}\left[\mathrm{GPR}_{log[\sigma^2]}(\boldsymbol{x}_i|\bar{\boldsymbol{\theta}}), \mathrm{GPR}_{log[\sigma^2]}\left(\boldsymbol{x}_j|\bar{\boldsymbol{\theta}}\right)\right],$$
$$i, j = 1, 2, \ldots, I. \tag{45}$$

If we denote $\boldsymbol{\mu}_{\mathrm{GPR}} = [\mu_{\mathrm{GPR}}(\boldsymbol{x}_1), \mu_{\mathrm{GPR}}(\boldsymbol{x}_2), \ldots, \mu_{\mathrm{GPR}}(\boldsymbol{x}_I)]^\top$ and $\boldsymbol{\theta}_{\mathrm{GPR}} = [\theta_{\mathrm{GPR}}(\boldsymbol{x}_1), \theta_{\mathrm{GPR}}(\boldsymbol{x}_2), \ldots, \theta_{\mathrm{GPR}}(\boldsymbol{x}_I)]^\top$, then the NEI acquisition function is defined as

$$\mathrm{NEI}_{\mathcal{F}}(\boldsymbol{x}) \quad = \mathbb{E}_{\boldsymbol{\mu}_{\mathrm{GPR}}, \boldsymbol{\theta}_{\mathrm{GPR}}}\left[\alpha_{\mathrm{EI}_{\mathcal{F}}}(\boldsymbol{x}|\boldsymbol{\mu}_{\mathrm{GPR}}, \boldsymbol{\theta}_{\mathrm{GPR}})\right] \ . \tag{46}$$

Finally, the selection of $\boldsymbol{x}_{I+1}$ is accomplished as

$$\boldsymbol{x}_{I+1} \quad = \underset{\boldsymbol{x}\in\mathcal{X}}{argmax}\mathrm{NEI}_{\mathcal{F}}(\boldsymbol{x}). \tag{47}$$

In order to calculate $\mathrm{NEI}_{\mathcal{F}}(\boldsymbol{x})$, Monte Carlo sampling is required for evaluating the expectation of $\alpha_{\mathrm{EI}_{\mathcal{F}}}(\boldsymbol{x}|\boldsymbol{\mu}_{\mathrm{GPR}}, \boldsymbol{\theta}_{\mathrm{GPR}})$ over $\boldsymbol{\mu}_{\mathrm{GPR}}$ and $\boldsymbol{\theta}_{\mathrm{GPR}}$, which means that the calculation cost for NEI becomes higher than that for EI. Nevertheless, the calculation of Monte Carlo sampling is apparently parallelizable in nature. Therefore, multiple cores are able to compensate for the burden of increase in calculations.

D. Discussion

In this section, we discuss the suitability of the BO algorithms described in Section 2.2-B,C. For problems where control parameters determine means and variances of Gaussian distributions, we deduce from Theorem 1 that applying GPRs to the parameters of Gaussian distributions is better than applying the GPR concept to the empirical, Bernoulli distribution-based estimators of yield. It is difficult to discuss the suitability and performance of the proposed approaches for the cases where the distributions of the outgoing quality are not Gaussian. Nevertheless, drawing an analogy with the discussion in Section 2.1-F, when the sample size N and the number of iterations is large, applying GPR to empirical probabilities is preferable due to the consistency of the corresponding estimator. On the other hand, when the sample size N and the number of iterations are small, applying GPRs to the parameters of the Gaussian distributions may be preferable, as the assumption of the Gaussian form of the outgoing quality characteristics compensates for the scarcity of samples.

Furthermore, when the specification area is narrow, applying GPRs to parameters of Gaussian distributions is expected to outperform applying the GPR concept to Bernoulli distribution-based estimators of yield, especially when the number of iterations is small. Essentially, if quality characteristics corresponding to a certain set of control parameters fall near the boundaries of the specification areas, these control parameters should be regarded as somewhat promising, even if the empirical probability that the quality characteristics satisfy the specification is zero. Applying GPRs to distribution parameters incorporates this intuitive insight. On the other hand, when GPR is applied to empirical probabilities, the algorithm cannot exploit the proximity of the quality characteristics to the specification area.

Finally, we comment on computational cost. Applying GPRs to the parameters of Gaussian distributions requires additional computation because Monte Carlo sampling is needed (see Section 2.2-C). The cost of these Monte Carlo calculations grows roughly with the number of samples and with the dimensionality of the characteristic space; so, high-dimensional problems or large sample sizes can become computationally demanding. When a single workstation is insufficient, the embarrassingly parallel nature of Monte Carlo sampling allows for straightforward distribution of independent samples (or independent control parameter evaluations) across multiple workstations.

3. Results

3.1. Experiments with Various Methods for Estimating Yields

In this section, we evaluate the performance of the yield estimators discussed in Section 2.1. In Section 3.1-A, we analyze the performance of yield estimators when the distribution of the outgoing quality is exactly Gaussian, while in Section 3.1-B, we consider the distribution of the outgoing quality not exactly following the Gaussian distribution form.

A. Estimation of yield when quality characteristic follows Gaussian distribution

Let us assume that a quality characteristic follows the normal distribution $\mathcal{N}(0,1)$ and that the desired interval for that characteristic is $\left[-w^*, +w^*\right]$, where w^* is the solution of

$$\sqrt{2\pi}\left(1 - 2erf\left(\frac{w_*}{\sqrt{2}}\right)\right) - \left(2w_* - 2w_*^3\right)e^{\frac{-w_*^2}{2}} = 0, \tag{48}$$

as described in Proposition 1. This situation is illustrated in Figure 1. MSEs for the classical yield estimator based on Bernoulli distribution $\hat{p}_1$ and the newly proposed yield estimator $\hat{p}_2$ for different sample sizes N are shown in Figure 2 and Table 1. We see that the theoretical expressions for relevant MSEs, namely (5) and (11), coincide with the empirical results.

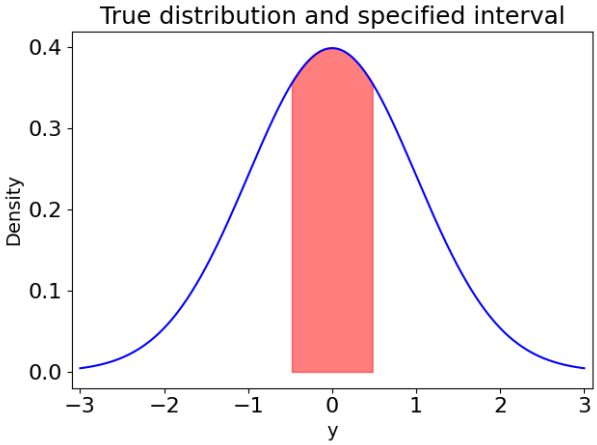


Figure 1. The probability density function of the standard normal distribution and the specification interval. This distribution and interval correspond to the case described in Proposition 1.

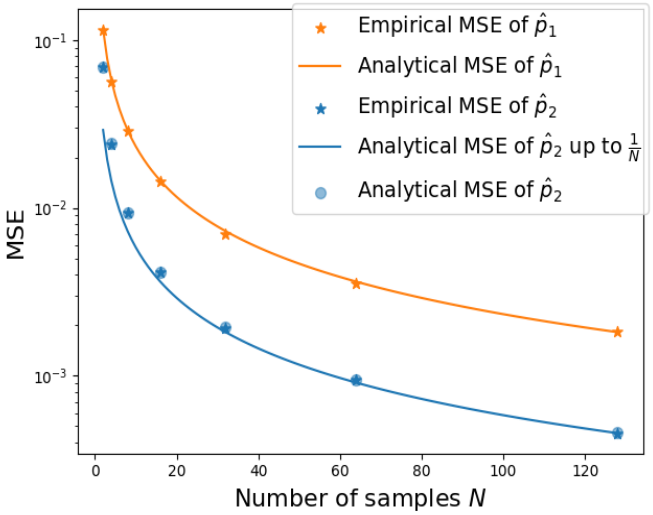


Figure 2. MSEs of the classical estimator $\hat{p}_1$ and the proposed estimator $\hat{p}_2$ for estimating the probability shown in Figure 1 for sample sizes $N = \{2, 4, 8, 16, 32, 64, 128\}$). For empirical MSEs, we repeated 10^4 experiments and took the average.

Table 1. Empirical squared errors of estimators (Mean ± STD).

N	Empirical Squared Errors of $\hat{p}_1$	Empirical Squared Errors of $\hat{p}_2$
2	$1.15 \times 10^{-1} \pm 1.25 \times 10^{-1}$	$6.93 \times 10^{-2} \pm 1.01 \times 10^{-1}$
4	$5.68 \times 10^{-2} \pm 7.27 \times 10^{-2}$	$2.39 \times 10^{-2} \pm 4.87 \times 10^{-2}$
8	$2.89 \times 10^{-2} \pm 3.90 \times 10^{-2}$	$9.30 \times 10^{-3} \pm 1.90 \times 10^{-2}$
16	$1.43 \times 10^{-2} \pm 1.99 \times 10^{-2}$	$4.16 \times 10^{-3} \pm 7.91 \times 10^{-3}$
32	$7.01 \times 10^{-3} \pm 9.86 \times 10^{-3}$	$1.92 \times 10^{-3} \pm 3.23 \times 10^{-3}$
64	$3.57 \times 10^{-3} \pm 5.05 \times 10^{-3}$	$9.44 \times 10^{-4} \pm 1.48 \times 10^{-3}$
128	$1.83 \times 10^{-3} \pm 2.57 \times 10^{-3}$	$4.54 \times 10^{-4} \pm 6.64 \times 10^{-4}$

The advantage of the newly proposed yield estimator is statistically supported by one-sided hypothesis testing using the bootstrap method, as shown in Table 2, for both $N = 2$ and $N = 128$. In both cases, sufficiently small p-values are obtained. In addition, the portion of MSE related to the $1/N$ factor, as characterized by (14), predicts the MSE of $\hat{p}_2$ when N is large, and the MSE of $\hat{p}_2$ is smaller than that of $\hat{p}_1$, as stipulated by Theorem 1.

Table 2. One-sided hypothesis testing using the bootstrap method (number of resamples = 10^4).

N	One-Sided Alternative Hypothesis	p-Value	Effect Size (Difference in Average)	95% CI of Difference in Squared Error (Percentiles from 2.5 to 97.5)
2	Squared error of $\hat{p}_2$ is less than that of $\hat{p}_1$.	$<10^{-16}$	-0.0460	$[-0.04912, -0.04284]$
128	Squared error of $\hat{p}_2$ is less than that of $\hat{p}_1$.	$<10^{-16}$	-0.00138	$[-0.00143, -0.001327]$

B. Estimation of Yield Expressed as Probabilities of Electron Densities and Temperatures Falling Within Specifications

In this subsection, we evaluate the performance of the classical yield estimator $\hat{p}_1$ and the newly proposed yield estimator $\hat{p}_2$ when the distribution of the outgoing process quality is not exactly Gaussian. This study will be performed through examination of the problem of ensuring plasma stability in an Inductively Coupled Plasma (ICP) chamber, which is a system widely utilized in semiconductor manufacturing.

ICP chambers use coils to generate a strong electric field within a vacuum chamber and thus generate plasmas. In the simulations considered in this paper, Ar and O_2 gases are ionized to create plasmas, where Ar ions mainly provide mechanical energy for sputtering, while O_2 ions enhance chemical reactions. Maintaining key plasma parameters within desired specification limits is crucial for ascertaining the desired quality of outgoing wafers, as those parameters have significant effects on the surface modifications of wafers. Specifically, following [41], this article focuses on the electron density and temperature among plasma parameters.

(1) Configuration of simulations:

COMSOL software package (version: 6.2) is utilized to perform simulations, with the model details available in [42]. A schematic sketch of the chamber considered in this paper is shown in Figure 3. It is assumed that one is able to set powers of the four coils to distinct values and thus generate an electric field within the chamber. From the gas inlet, Ar and O_2 gases flow into the chamber and leave the chamber via the outlet. Random

variations in the Ar/O_2 concentration ratios occur due to the effects of residual gases inside the chamber, while random fluctuations of pressures attributes arise from uncertainties in the control actions of pressure controllers. Given inputs, COMSOL simulations provide the stationary spatial fields of electron densities and temperatures across the chamber, based on which measurements of electron densities and temperatures are assumed to be obtained via a plasma absorption probe installed at a specific location in the chamber. A plasma absorption probe is a device capable of measuring local absolute electron density [43] and can be used to estimate electron temperature [44]. The radius of the dielectric tube used in such probes is typically a few millimeters [43,45]. Therefore, for this simulation, probe measurements were modeled as the spatial average of the plasma parameters over an area of 1 cm $\times$ 1 cm, with the aim of keeping these measurements within some a priori specified intervals. The objective of the numerical experiments in this section is to accurately estimate the probability that the measurements of electron densities and temperatures will fall within the desired specification intervals.

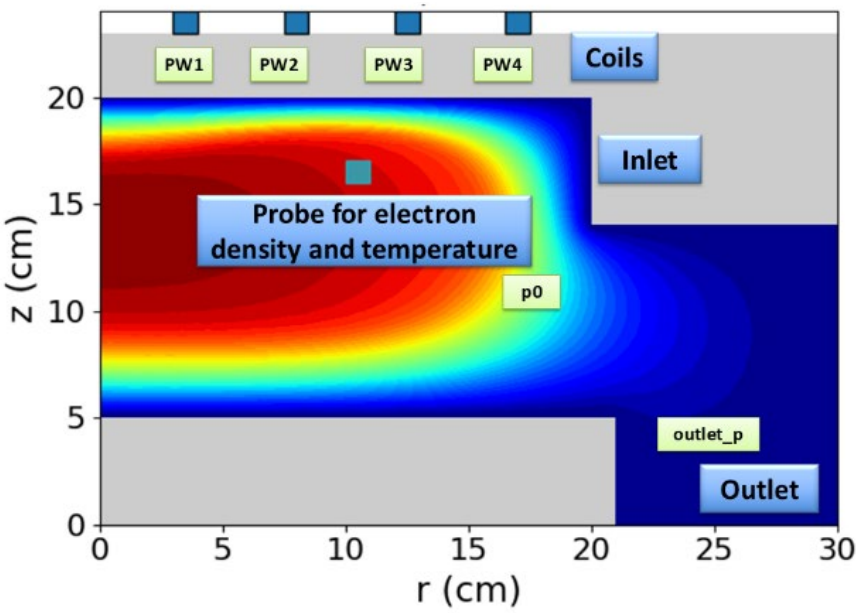


Figure 3. A sketch of the ICP chamber used in the numerical experiments considered in this paper. The settings of simulation parameters are based on [42]. The model of simulations is revised so that powers of four coils can be set to distinct values and are considered control parameters. In addition, pressures at the outlet are also parametrized to emulate the variations in pressure controllers.

(2) Probability of electron density falling within specification limits:

In this subsection, we compare the performance of the classical yield estimator $\hat{p}_1$ and the newly proposed yield estimator $\hat{p}_2$ in terms of their performance in estimating the probability that the sensed electron density falls within the desired specification interval. Stochastic variations in chamber conditions are simulated by randomly disturbing inputs of the simulator, and an empirical distribution of electron densities is acquired by running simulation replications with these disturbed inputs. The settings used in the simulations are summarized in Table 3. In order to ensure that the parameters stay within the valid range, we clip the values of xO2 to fall within $[0, 1]$ and ensure that the values of p0 are positive after sampling from the Gaussian distribution. A total of 1000 simulation replications were run and, from the valid simulation runs, the spatial average of electron density was calculated for the area assumed to be occupied by the probe.

Table 3. Input parameters of simulations for constructing true distribution of electron density.

Item	Type	Value/Distribution
Power (Pw1-4)	Fixed	100 [W]
Ratio of O2 at inlet (xO2)	Varied	$\mathcal{N}\left(0.6, (0.1)^2\right)$
Pressure (p0)	Varied	$\mathcal{N}\left(0.04, (0.01)^2\right)$ [torr]
Pressure at outlet (outlet_p)	Fixed	0 [torr]

Figure 4 shows the histogram of electron densities obtained using the above-described simulations, with the desired specification interval superimposed over the true distribution. While the distribution illustrated in Figure 4 appears to be similar to a Gaussian distribution in terms of its mono-modality, the Shapiro–Wilk test [46] implies that it does not follow a Gaussian form ($p = 2.64 \times 10^{-12}$). This is consistent with the fact that, in semiconductor manufacturing, nonlinear effects and substantial noise often lead to non-Gaussian distributions of quality characteristics. Nevertheless, we will still attempt to estimate the probability that electron density falls within the desired specification interval by estimating Gaussian distribution parameters from N independent, identically distributed samples, as in (7).

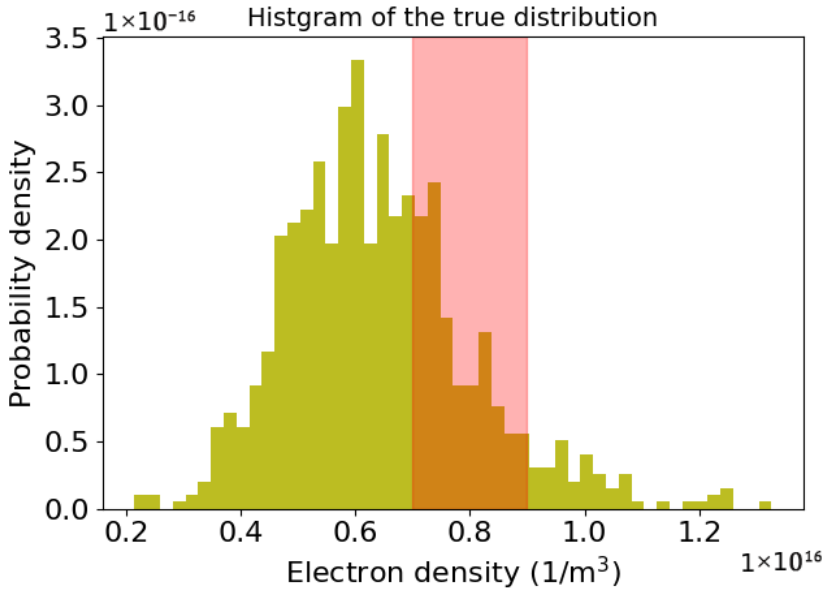


Figure 4. The true distribution of electron density constructed from multiple simulations and the specified interval. Though this distribution is mono-modal and appropriate to be approximated with a Gaussian distribution in this regard, the p-value of the Shapiro–Wilk test is 2.64×10^{-12}.

In Figure 5 and Table 4, the MSEs of yield estimators $\hat{p}_1$ and $\hat{p}_2$ for the above-mentioned probability are shown. As discussed in Section 2.1-F, the MSE of the proposed estimator $\hat{p}_2$ is smaller than that of the classical estimator $\hat{p}_1$ until a certain threshold value of the sample size N is reached, beyond which the superiority flips and the MSE of $\hat{p}_1$ becomes smaller. This finding is supported by one-sided hypothesis tests using the bootstrap method for $N = 2$ and $N = 128$, as in Table 5. For $N = 2$, the test indicates that the empirical squared errors of $\hat{p}_2$ are significantly smaller than those of $\hat{p}_1$. Conversely, for $N = 128$, the test indicates that the empirical square errors of $\hat{p}_1$ are significantly smaller than those of $\hat{p}_2$.

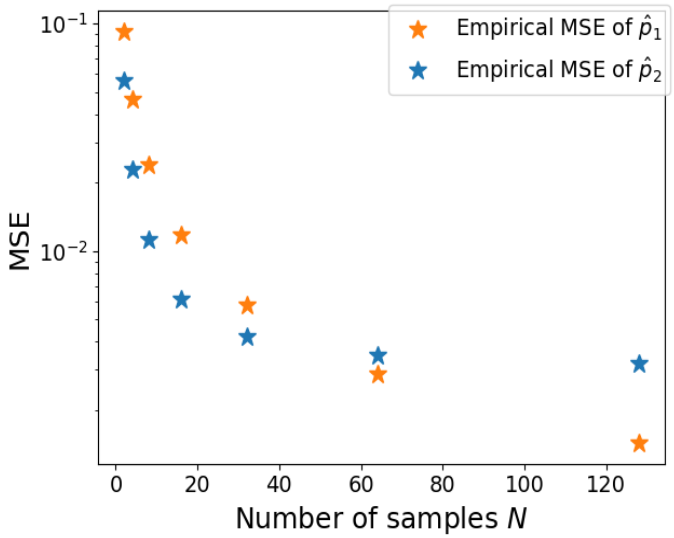


Figure 5. The MSEs of yield estimators $\hat{p}_1$ and $\hat{p}_2$ for estimating the probability shown in Figure 4 for sample sizes ($N = \{2, 4, 8, 16, 32, 64, 128\}$). For empirical MSEs, we repeated 10^4 experiments and took the average.

Table 4. Empirical squared errors of estimators (Mean ± STD).

N	Empirical Squared Errors of $\hat{p}_1$	Empirical Squared Errors of $\hat{p}_2$
2	$9.26 \times 10^{-2} \pm 1.20 \times 10^{-1}$	$5.58 \times 10^{-2} \pm 1.06 \times 10^{-1}$
4	$4.66 \times 10^{-2} \pm 6.21 \times 10^{-2}$	$2.27 \times 10^{-2} \pm 4.19 \times 10^{-2}$
8	$2.40 \times 10^{-2} \pm 3.34 \times 10^{-2}$	$1.12 \times 10^{-2} \pm 1.72 \times 10^{-2}$
16	$1.18 \times 10^{-2} \pm 1.67 \times 10^{-2}$	$6.16 \times 10^{-3} \pm 8.32 \times 10^{-3}$
32	$5.81 \times 10^{-3} \pm 8.36 \times 10^{-3}$	$4.18 \times 10^{-3} \pm 4.73 \times 10^{-3}$
64	$2.86 \times 10^{-3} \pm 3.99 \times 10^{-3}$	$3.48 \times 10^{-3} \pm 3.24 \times 10^{-3}$
128	$1.43 \times 10^{-3} \pm 2.00 \times 10^{-3}$	$3.19 \times 10^{-3} \pm 2.32 \times 10^{-3}$

Table 5. One-sided hypothesis testing using the bootstrap method (number of resamples = 10^4).

N	One-Sided Alternative Hypothesis	*p*-Value	Effect Size (Difference in Average)	95% CI of Difference in Squared Error (Percentiles from 2.5 to 97.5)
2	Squared error of $\hat{p}_2$ is less than that of $\hat{p}_1$.	$< 10^{-16}$	-0.0369	$[-0.0400, -0.0337]$
128	Squared error of $\hat{p}_1$ is less than that of $\hat{p}_2$.	$< 10^{-16}$	-0.00176	$[-0.00183, -0.00170]$

(3) Probability of electron densities and temperatures falling within specifications:

This subsection evaluates the performance of the classical estimator $\hat{p}_1$ and the newly proposed estimator $\hat{p}_2$ in the simulations when the outgoing quality characteristics are multivariate.

Suppose that the desired specification intervals for electron density and temperature are given. As mentioned before, both electron density and temperature are measured as averages over the areas covered by the probes schematically shown in Figure 3. As in the previous subsection, we simulate the variability in chamber conditions by randomly disturbing inputs of the simulator, which yields the true distribution of output variables.

Figure 6 depicts the true joint distribution of electron densities and temperatures, along with the desired specification area. Simple visual inspection indicates that, though monomodal, the true distribution does not follow a Gaussian form and is also correlated, which does not match the assumptions from Theorem 2 in Section 2.1-E.

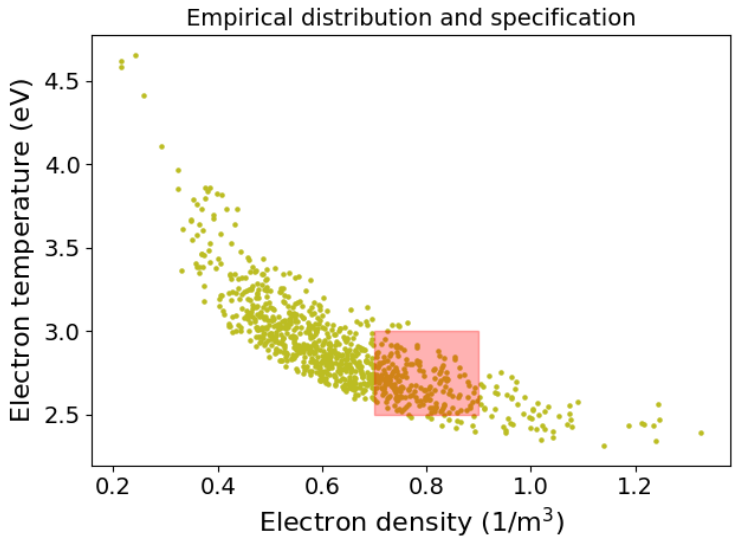


Figure 6. The true distribution of electron density and temperature constructed from multiple runs of simulations and the specified intervals. We can see that the true distribution is not a Gaussian distribution and correlated. The desired specification area is represented with the pink square.

A comparison of yield estimators $\hat{p}_1$ and $\hat{p}_2$ is shown in Figure 7 and Table 6. Similar to what we observed in the scalar case, the MSE of the proposed yield estimator $\hat{p}_2$ is smaller than that of the classical yield estimator $\hat{p}_1$ until a certain threshold sample size N is reached, after which the performance of $\hat{p}_1$ and $\hat{p}_2$ flips, as discussed in Section 2.1-F. As in the scalar case, the hypothesis test using the bootstrap method indicates that the proposed estimator $\hat{p}_2$ is significantly preferable for $N = 2$, whereas the classical yield estimator $\hat{p}_1$ is significantly preferable for $N = 128$, as shown in Table 7.

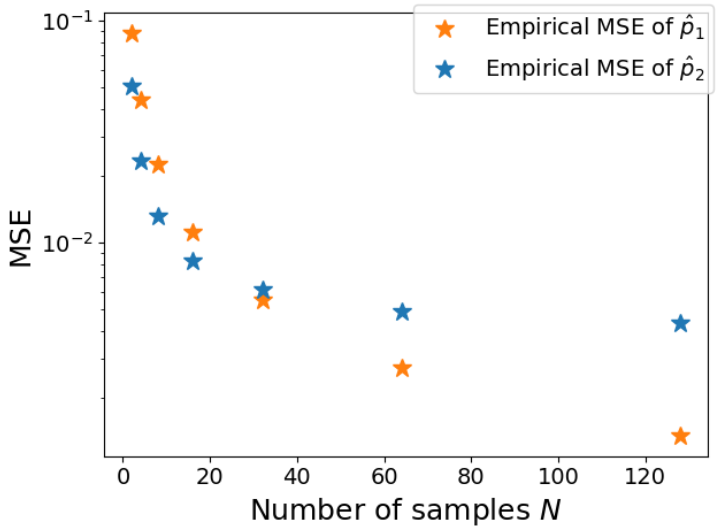


Figure 7. The MSEs of yield estimators $\hat{p}_1$ and $\hat{p}_2$ for estimating the probability shown in Figure 6 for various sample sizes ($N = \{2, 4, 8, 16, 32, 64, 128\}$). For empirical MSEs, we repeated the experiment 10^4 times and took the average.

Table 6. Empirical squared errors of estimators (Mean ± STD).

N	Empirical Squared Errors of $\hat{p}_1$	Empirical Squared Errors of $\hat{p}_2$
2	$8.81 \times 10^{-2} \pm 1.19 \times 10^{-1}$	$5.07 \times 10^{-2} \pm 8.96 \times 10^{-2}$
4	$4.41 \times 10^{-2} \pm 6.06 \times 10^{-2}$	$2.34 \times 10^{-2} \pm 3.53 \times 10^{-2}$
8	$2.26 \times 10^{-2} \pm 3.14 \times 10^{-2}$	$1.31 \times 10^{-2} \pm 1.50 \times 10^{-2}$
16	$1.12 \times 10^{-2} \pm 1.57 \times 10^{-2}$	$8.29 \times 10^{-3} \pm 8.82 \times 10^{-3}$
32	$5.53 \times 10^{-3} \pm 8.01 \times 10^{-3}$	$6.10 \times 10^{-3} \pm 5.97 \times 10^{-3}$
64	$2.71 \times 10^{-3} \pm 3.81 \times 10^{-3}$	$4.90 \times 10^{-3} \pm 4.20 \times 10^{-3}$
128	$1.34 \times 10^{-3} \pm 1.86 \times 10^{-3}$	$4.35 \times 10^{-3} \pm 3.05 \times 10^{-3}$

Table 7. One-sided hypothesis testing using the bootstrap method (number of resamples = 10^4).

N	One-Sided Alternative Hypothesis	p-Value	Effect Size (Difference in Average)	95% CI of Difference in Squared Error (Percentiles from 2.5 to 97.5)
2	Squared error of $\hat{p}_2$ is less than that of $\hat{p}_1$.	$<10^{-16}$	-0.0373	$[-0.0402, -0.0344]$
128	Squared error of $\hat{p}_1$ is less than that of $\hat{p}_2$.	$<10^{-16}$	-0.00301	$[-0.00309, -0.00294]$

3.2. Numerical Experiments for Maximizing Yield

In this section, we compare the performance of BO algorithms for maximizing the yield of the outgoing product quality. Specifically, we focus on the difference between applying the GPR paradigm to directly model the empirical yields, as described in Section 2.2-B, versus applying GPRs to model parameters of the Gaussian distributions of the outgoing quality, as described in Section 2.2-C.

In Section 3.2-A, we compare the performance of the yield estimation and optimization algorithms for a synthetic problem in which the outgoing quality follows a Gaussian distribution. In Section 3.2-B, we consider yield estimation and optimization in ICP chamber simulations. Finally, in Section 3.2-C, we use ICP chamber simulations to perform a set of sensitivity studies regarding the preferability of different estimation and optimization algorithms for that specific problem.

A. Optimization of control parameters which determine parameters of Gaussian distribution

In this subsection, we consider a synthetic problem of determining the control input for a process which will result in a maximal outgoing yield under the assumption that the outgoing quality follows a Gaussian distribution. More precisely, the problem is defined as follows. The domain of the control input parameter is $\mathcal{X} = [-3, +3]$ and is discretized into 1000 points. The distribution of the outgoing quality y is assumed to be distributed as $\mathcal{N}(\mu(x), \sigma^2(x))$, where

$$\mu(x) = 2x^2 \tag{49}$$

and

$$\sigma^2(x) = \begin{cases} -2x^2 + 10 - 4x & (0.1 \leq x) \\ 0.1 & (x < 0.1) \end{cases}. \tag{50}$$

The desired specification interval for the outgoing quality variable y is assumed to be $S = [-2, +2]$. Figure 8 illustrates the dependency of the Gaussian distribution parameters and the corresponding yield on the control parameter x. It can be noted that though the

mean $\mu(x)$ of the outgoing quality takes the value of 0 when $x = 0$, which corresponds to the center of the specification interval S, the variance $\sigma^2(x)$ of the outgoing quality is relatively large when $x = 0$. Therefore, the maximum yield is realized when $x \fallingdotseq 0.72$, for which, even though the expected value of the outgoing quality is not centered in the middle of the specification interval, the variance $\sigma^2(x)$ is relatively small. This synthetic problem is an example that demonstrates that it is inadequate to only focus on the mean parameter for optimization of the yield.

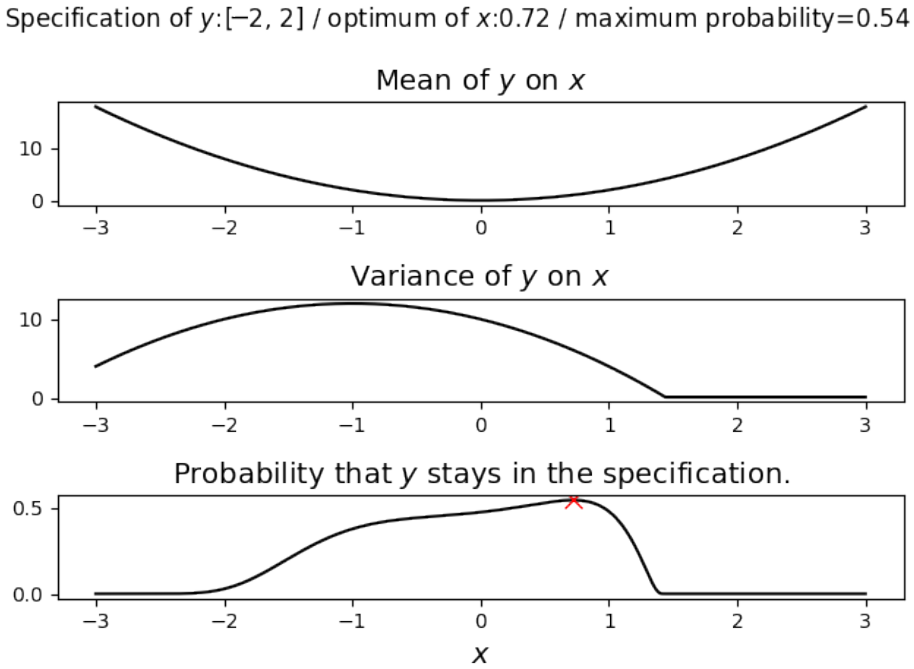


Figure 8. The relationship between control parameter x and Gaussian distribution parameters. Both mean parameters and variance parameters are determined by given x. The red X corresponds the maximum yield in this problem.

Two versions of the BO algorithm are used for solving the problem described above. In one version, labeled in the remainder of the paper as *BO PROB*, GPR is employed to model the empirical yields obtained via estimation of the relevant Bernoulli parameter, as described in Section 2.2-B. In the other BO version, labeled as *BO DIST* in the remainder of this paper, GPRs are employed to model the mean and variance parameters of the outgoing product quality, as described in Section 2.2-C. Both *BO PROB* and *BO DIST* exploit the NEI acquisition functions, as discussed in Section 2.2. In our experiments, we use RBF (squared-exponential) kernels for *BO PROB* and *BO DIST*. As new observations are collected, the kernel length-scale parameters are optimized by maximizing the marginal log-likelihood. We impose a log-normal prior on each length-scale with $\ln\ell \sim \mathcal{N}(2 + 0.5d, \sigma^2 = 3)$, where d denotes the number of control parameters. Detailed configurations are available in the Supplementary Codebase.

A set of four initial guesses for the control parameter x were randomly generated following a Sobol Sequence [47], and $N = 8$ samples of the outgoing quality y were generated for each of those control parameters. Starting from that, *BO PROB* and *BO DIST* iterative procedures were run following Algorithm 1. In order to evaluate the stochastic behavior of the resulting BO methods, we performed 16 trials of the optimization procedures, with the progression of resulting quantiles of the yield suggested by the algorithms depicted in Figure 9 and Table 8.

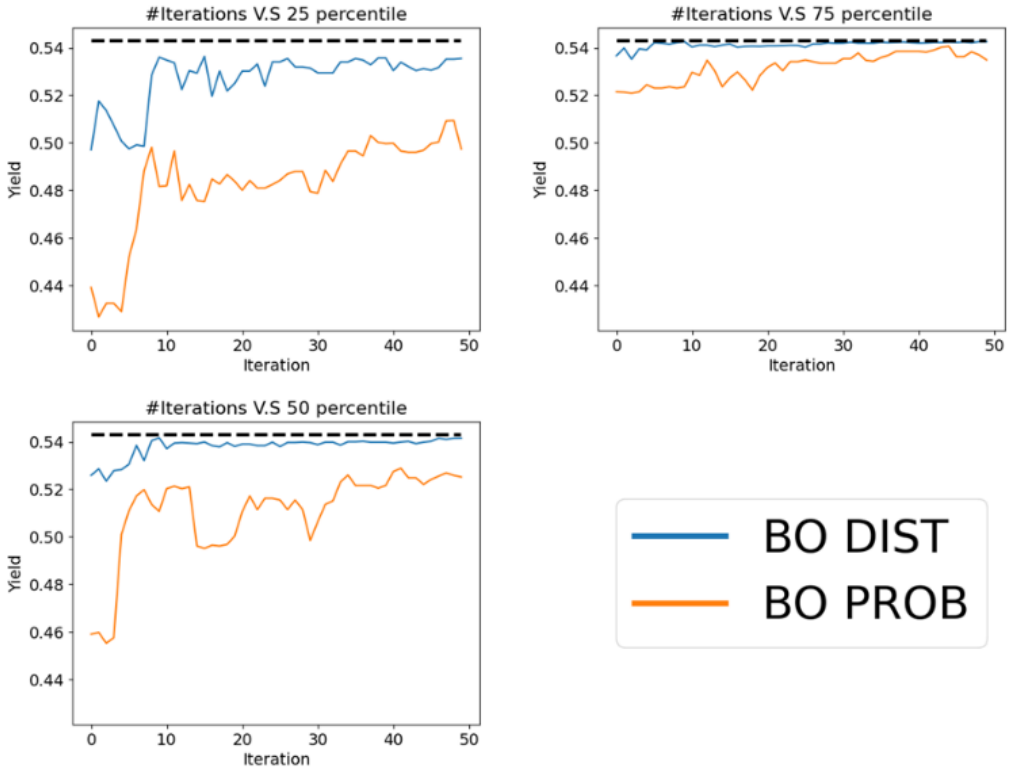


Figure 9. The progression of the 25th, 50th and 75th percentiles of yields obtained for the control parameters suggested over 16 random replications of the *BO PROB* and *BO DIST* algorithms. The black dashed line corresponds to the highest possible yield in this synthetic problem.

Table 8. Yields suggested by Bo methods as iterations progressed.

Percentile	**25**		**50**		**75**	
Method	**BO DIST**	**BO PROB**	**BO DIST**	**BO PROB**	**BO DIST**	**BO PROB**
Iteration						
4	0.501	0.429	0.528	0.501	0.539	0.524
9	0.536	0.482	0.542	0.511	0.543	0.524
14	0.529	0.476	0.539	0.496	0.541	0.524
19	0.525	0.484	0.538	0.500	0.541	0.528
24	0.534	0.482	0.540	0.516	0.541	0.534
29	0.531	0.479	0.540	0.498	0.542	0.534
34	0.534	0.497	0.540	0.526	0.542	0.534
39	0.536	0.500	0.540	0.522	0.542	0.539
44	0.531	0.497	0.540	0.522	0.542	0.541
49	0.535	0.497	0.542	0.525	0.543	0.535

As a tendency, one can see that as the iterations progress, the yields obtained using the suggested control parameters become higher. However, this tendency is not monotone due to the random nature of the problem and the BO procedure, as well as the limited number of trials. Inspection of the results corresponding to the 75th percentile shows little difference between *BO DIST* and *BO PROB* because both methods realize the yield close to the highest possible value. However, progressions of the 25th percentiles show the superiority of *BO DIST* in the sense that it realizes a relatively high yield very early in the

iteration process. As per the discussions in Section 2.2-D, the superiority of the *BO DIST* method is reasonable because the outgoing quality y in this synthetic problem actually follows a Gaussian distribution.

We present additional experiments in Appendix D that examine the effects of kernel selection and the differences between the EI and NEI acquisition functions.

B. Optimization of control parameters to maximize probability that electron densities and temperatures are within specifications

In this subsection, we study an optimization problem in which the objective is to maximize the yield defined as the probability that electron densities and temperatures in an ICP chamber fall within the desired specification limits. Similarly to what was performed in the previous subsection, we compare the performance of the *BO PROB* and *BO DIST* algorithms.

For this purpose, we use the simulation environment described in Section 3.1-B, with the basic simulation settings depicted in Figure 3 and discussed in Section 3.1-B-1. In order to derive the empirical distributions of electron densities and temperatures for each setting of control parameters, one needs to conduct a prohibitively large number of simulations. To address this challenge, we used the simulation environment to construct a Machine Learning (ML)-based surrogate model capable of predicting measurements of electron densities and temperatures based on a limited number of simulation results. Specifically, a Random Forest regression with z-score normalization [48] was employed because of its well-documented ability to achieve high accuracy even with a limited number of training samples [49]. This model was trained using the results of 2000 COMSOL simulation runs conducted for input parameters sampled using the Sobol Sequence method from the domain described in Table 9, thus pursuing an informative training set via a limited number of simulations.

Table 9. Domain of input parameters used to generate training data for Random Forest-based surrogate model of ICP chamber.

Item	Range
Coil Power (PW1)	10 [W]
Coil Powers (PW2, PW3)	(0, 300) [W]
Coil Powers (PW4)	250 [W]
Ratio of O2 at inlet (xO2)	(0.7, 0.9)
Pressure (p0)	(0.01, 0.02) [torr]
Pressure at outlet (outlet_p)	(0, 0.02) [torr]

Table 10 summarizes the input settings for the problem. The parameters labeled as *Control* are iteratively selected through the optimization algorithms, while the parameters labeled as *Randomly Varied* correspond to the uncontrollable inputs whose variations lead to variability in chamber conditions and, consequently, randomness in the measured electron densities and temperatures. As for the parameters labeled as *Fixed*, their values are not varied as inputs into the ML surrogate model. The domain of each control parameter, i.e., the domains of values for each of the two controllable power coils, is discretized into 20 points. The goal of the BO procedures is to determine control parameters that would maximize the probability of electron densities and temperatures measured in the probe areas illustrated Figure 3 being within desired specification limits. Specification limits for the electron densities and temperatures are summarized in Table 11.

Table 10. Input parameters for simulations in order to construct true distribution of electron densities and temperatures.

Item	Type	Domain/Value/Distribution
Power1 (Pw1)	Fixed	10 [W]
Power2 (Pw2)	Control	10–250 [W]
Power3 (Pw3)	Control	10–250 [W]
Power4 (Pw4)	Fixed	250 [W]
Ratio of O2 at inlet (xO2)	Randomly Varied	$N\left(0.8, (0.01)^2\right)$
Pressure (p0)	Randomly Varied	$\mathcal{N}\left(0.015, (0.0005)^2\right)$ [torr]
Pressure at the outlet (outlet_p)	Randomly Varied	$\mathcal{N}\left(0.01, (0.0005)^2\right)$ [torr]

Table 11. Specification of electron densities and temperatures.

Item	Range
Electron density $\left(1/\mathrm{m}^3\right)$	$\left(1.8 \times 10^{16}, 2.0 \times 10^{16}\right)$
Electron temperature (eV)	$(3.8, 4.0)$

To clarify the relationship between surrogate-based evaluations and real experiments, we summarize their correspondences in Table 12. While the surrogate model and simulations are designed to mimic the measurements obtained from the physical system and therefore produce data that can be treated with the same evaluation metrics, real experiments require repeated measurements at fixed settings to directly observe variability arising from environmental factors and measurement noise. A typical real experiment loop consists of performing repeated measurements at a chosen setting, training a Gaussian Process (GP) on the acquired data, computing the acquisition function, and selecting the next setting sequentially. Because real experiments incur substantial time, material, and operational costs, a pragmatic workflow is to first screen and evaluate promising methods using surrogate models and then validate selected candidates on the actual equipment.

Table 12. Correspondence between surrogate model experiments and real experiments.

Aspect	Surrogate Model	Real Experiments
Control variables/operating windows	Coils (see Table 3 for details)	Same conditions
Measurement cost /response lag	Negligible per eval/fast surrogate; (0.1~1 [s] per evaluation on workstation)	Measurement of sensors for processing wafers /repetition of wafer processing (tens of minutes ~ hours)
Yield metrics	Probability both density/temp within specification limits	Same metrics

Figure 10 illustrates the relationship between control parameters and the probabilities that electron densities and temperatures fall within the specification intervals, as determined by the surrogate ML model with empirical samples from the domain described in Table 10.

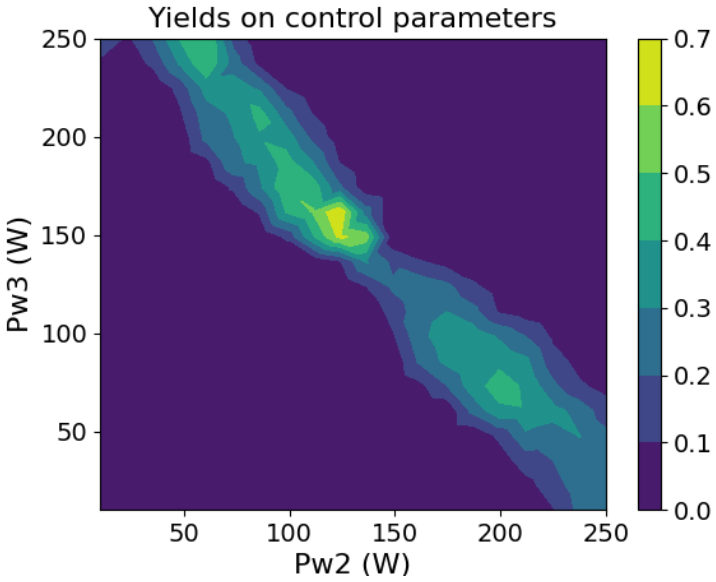


Figure 10. The relationship between control parameters and yield in the problem, obtained using the surrogate model. Pw2 and Pw3 correspond to the powers of coils described in Figure 3.

A set of four initial guesses for the pair of coil powers were randomly selected following a Sobol Sequence, and $N = 8$ samples of electron densities and temperatures were produced by the surrogate ML model for each of those pairs. Starting from that, *BO PROB* and *BO DIST* iterative procedures were run following Algorithm 1 in order to optimize the pair of controllable coil powers. In order to evaluate the stochastic behavior of the resulting BO methods, we ran 16 trials of the optimization procedures, with the progression of the 25th, 50th and 75th percentiles of the yields obtained by the BO algorithms depicted in Figure 11 and Table 13. For one trial, approximately 10 s and 80 s were necessary for *BO PROB* and *BO DIST*, respectively, in this evaluation setting with the standard workstation.

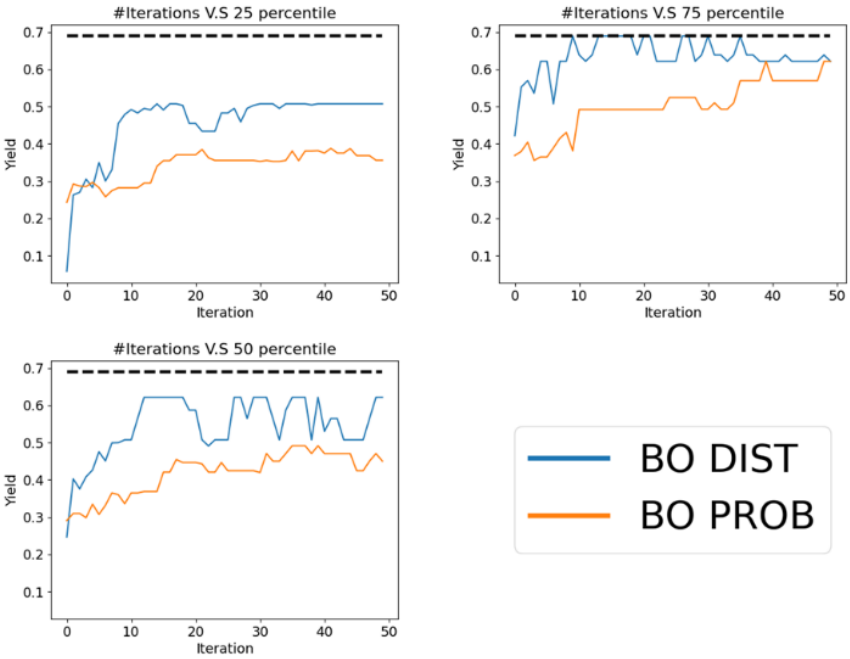


Figure 11. The progression of the 25th, 50th and 75th percentiles of yields obtained for the control parameters suggested over 16 random replications of the BO algorithms considered in this paper. The black dashed line corresponds to the maximal possible yield, obtained via analysis of yield dependency on the ICP control parameters illustrated in Figure 10.

Table 13. Yields suggested by Bo methods for each iteration.

Percentile	25		50		75	
Method	**BO DIST**	**BO PROB**	**BO DIST**	**BO PROB**	**BO DIST**	**BO PROB**
Iteration						
4	0.283	0.297	0.426	0.335	0.621	0.365
9	0.479	0.282	0.507	0.336	0.689	0.381
14	0.507	0.340	0.621	0.369	0.689	0.492
19	0.455	0.371	0.587	0.447	0.638	0.492
24	0.483	0.356	0.507	0.447	0.621	0.524
29	0.503	0.356	0.621	0.424	0.638	0.493
34	0.507	0.356	0.587	0.471	0.638	0.510
39	0.507	0.382	0.621	0.492	0.621	0.621
44	0.507	0.387	0.507	0.471	0.621	0.570
49	0.507	0.356	0.621	0.450	0.621	0.621

One can see that when the number of iterations is small, the *BO DIST* approach which relies on estimating means and variances of the underlying distributions achieves higher yields than the *BO PROB* method which relies on directly modeling empirical yields. As discussed in Section 2.2-D, this may be attributed to the scarcity of experiments compensated for by the assumption of a Gaussian form. If one looks at the progression of the 75th percentiles, *BO PROB* improves its performance as iterations progress, while the *BO DIST* method deteriorates in the later iterations. This may be attributed to the incorrect Gaussianity assumption, leading to the *BO DIST* method being informed by incorrect yield estimations, as discussed in Section 2.1-F.

C. Sensitivity studies for BO performance in ICP chamber simulations

As discussed in Section 2.2-D, the preferability of the newly proposed *BO DIST* algorithm, which relies on GPRs modeling the Gaussian distribution parameters, depends on the properties of the underlying problem. Here, we perform two sensitivity studies in order to illustrate the following:

- When the desired specification area is narrower, the newly proposed algorithm works more favorably.
- When the number of experiments is smaller, the newly proposed algorithm works more favorably.

Table 14 shows the specification intervals used in numerical experiments investigating the influence of the strictness of the desired specification on BO performance. In the case labeled as *STRICT*, the specified intervals of electron density and temperature are narrow, and the maximal possible yield is approximately 0.52. In the case labeled as *LENIENT*, the specification intervals are wider, with the maximal possible yield being approximately 0.71.

Table 14. The specification of electron densities and temperatures for a sensitive study for the strictness of specification.

Item	STRICT	LENIENT
Electron density $(1/\mathrm{m}^3)$	$(1.8 \times 10^{16}, 1.9 \times 10^{16})$	$(1.8 \times 10^{16}, 2.1 \times 10^{16})$
Electron temperature (eV)	$(3.8, 3.9)$	$(3.8, 4.1)$

The number of experiments per iteration N was set to be 8 and the initial settings of control parameters were randomly selected, after which the *BO PROB* and *BO DIST* iterative procedures were run following Algorithm 1 in order to maximize the yield via optimization of the pair of controllable coil powers. The performance of each algorithm was investigated using the ratio of the yield determined by that algorithm vs. the actual maximal possible yield for the problems. For each BO algorithm, we performed 32 trials of the optimization procedure and calculated the medians of the aforementioned yield ratios, which are depicted in Figure 12.

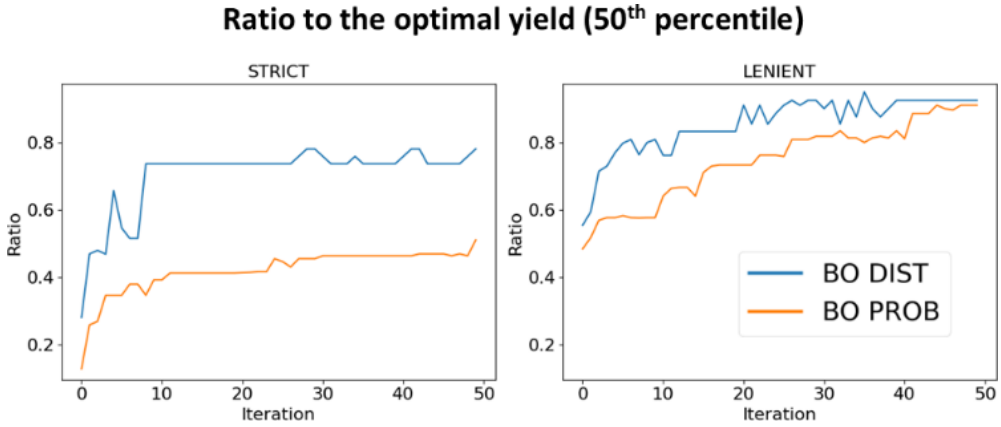


Figure 12. Progressions of the ratios to the maximized yield produced by the algorithms versus the maximal possible yield for the *STRICT* and *LENIENT* problems defined in Table 14.

It can be seen that when the specifications are strict, the *BO DIST* algorithm relying on the newly proposed concept of GPRs modeling the mean and variance parameters of the outgoing quality outperforms the classical approach relying on the GPR modeling empirical yields. On the other hand, when the specifications are lenient, the classical approach achieves high yields for a relatively smaller number of iterations. As discussed in Section 2.2-D, these observations may be attributed to the fact that the newly proposed algorithm takes into consideration the proximity of the observed characteristics to the desired specifications, thus facilitating better yield estimates from a smaller number of samples.

Under the same input parameter settings and specification limits as those considered in the previous subsection (input parameter settings listed in Table 10 and specification limits summarized in Table 11), we evaluated the performance of the optimization algorithms for different sample sizes N available at each iteration in the BO algorithms. A small sample size $N = 4$ and a large sample size of $N = 256$ were considered, with 32 random trials of the BO procedures conducted. Figure 13 reports the progression of medians of the resulting ratios of the yields produced by the *BO PROB* and *BO DIST* methods versus the maximal possible yield.

One can see that when the sample size N is small, the *BO DIST* approach relying on the newly proposed concept of GPRs modeling the parameters of the Gaussian distribution performed better. On the other hand, when the sample size N is large, the classical approach relying on the GPR modeling empirical yield becomes superior, except at the beginning of the iterations. As discussed in Section 2.2-D, the effectiveness of the proposed algorithm for small sample sizes N can be attributed to the ability of the Gaussian distribution assumption to compensate for a limited number of experiments, even when the Gaussianity assumption does not hold exactly true. In contrast, the classical *BO PROB* algorithm shows its strength with the large sample size N due to the consistency of the Bernoulli distribution-based yield estimator—a property that the yield estimator proposed in this paper does not possess.

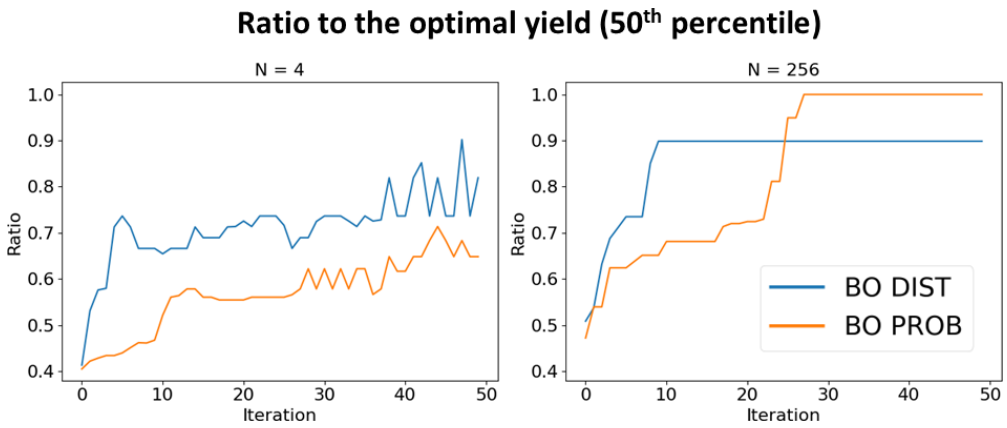


Figure 13. Progressions of the ratios to the maximized yield produced by the *BO PROB* and *BO DIST* algorithms versus the maximal possible yield as determined via analysis of the function illustrated in Figure 10.

4. Conclusions

In this paper, we introduced novel methodologies for estimating and maximizing yield in a manufacturing process, with yield being defined as the probability that quality characteristics fall within the desired specifications. The newly proposed yield estimator relies on estimating the expected value and variance of the outgoing product quality, based on which a Gaussian distribution form is assumed to calculate the yield. It was mathematically proven that if the outgoing product quality characteristics follow an uncorrelated multivariate Gaussian distribution, the newly proposed yield estimator achieves asymptotically lower MSE values compared to the classical yield estimator, which directly estimates the relevant Bernoulli distribution parameters. Additionally, we integrated this yield estimator into a BO framework in order to effectively optimize yield in the presence of uncertainties in the relevant manufacturing processes.

Theoretical insights regarding the properties of the novel yield estimator were verified in a synthetic example of yield estimation and optimization. In addition, the BO-based yield optimization procedures using the novel and traditional yield estimators were evaluated through high-fidelity simulations of a reactive ion etch chamber. The objective of these simulations was to determine tool parameters that maximize the probability of delivering desired electron density and temperature characteristics in plasma. BO using the newly proposed yield estimation algorithm showed advantages when the number of experiments was limited and when the desired specifications were strict, with advantages being more prevalent when the underlying distribution of the outgoing quality characteristics can be approximated well with an uncorrelated multivariate Gaussian distribution.

Looking ahead, future research should emphasize applications of the proposed method to experiments performed with physical equipment. Furthermore, the proposed method can be applied to estimate and maximize yields of various outgoing characteristics in different fields of manufacturing. From the perspective of theoretical developments, extending our methodologies to account for distributions beyond uncorrelated multivariate Gaussian distributions is essential. It will enable one to handle the critical question of when and to what degree the incorporation of prior knowledge about the underlying distribution form leads to a reduction in MSE when the distribution form differs from the uncorrelated multivariate Gaussian distributions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16073213/s1.

Author Contributions: Conceptualization, K.S. and D.K.; methodology, K.S. and D.K.; software, K.S.; validation, K.S., D.K., and D.D.; formal analysis, K.S. and D.K.; resources, D.K., Y.S. and H.M.; writing—original draft preparation, K.S.; writing—review and editing, K.S., D.K. and D.D.; project administration, Y.S., H.M. and D.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets generated during the current study are available from the corresponding author on reasonable request. Please note that, as this research was conducted within a corporate setting, access to the data and scripts may require a Non-Disclosure Agreement (NDA) or other contractual arrangements. The script generated during the current study is provided as the Supplemental Materials.

Acknowledgments: We thank Takahito Matsuzawa and Masato Kazui for their assistance and support. We also gratefully acknowledge the helpful comments and feedback on the manuscript from members of Dragan Djurdjanovic's laboratory. Finally, we thank our colleagues for their support during this work. During the preparation of this manuscript/study, the authors used GPT-4.1, GPT-4.1 mini, GPT-4o, and GPT-5 mini for the purposes of text editing and proof reading.

Conflicts of Interest: Authors Kei Sano, Daiki Kawahito, Yukiya Saito, Hironori Moki were employed by the Tokyo Electron Ltd. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Notations and Abbreviations

Symbol	**Description**
x	Scalar value
$\boldsymbol{x}$	Vector (column vector)
$\boldsymbol{X}$	Matrix
x_i	i-th element of vector $\boldsymbol{x}$
X_{ij}	(i,j)-th element of matrix $\boldsymbol{X}$
Abbreviation/Symbol	**Full Term/Meaning**
BO	Bayesian Optimization
MSE	Mean Squared Error
GPR	Gaussian Process Regression
EI	Expected Improvement (an acquisition function)
NEI	Noisy Expected Improvement (an acquisition function)
ICP	Inductively Coupled Plasma
ML	Machine learning
y	Quality characteristics
$S = [a, b]$	Specification interval
p_*	True yield
μ_*	True mean
σ_*^2	True variance
N	Number of samples
$\hat{p}_1$	Estimator based on Bernoulli distribution
$\hat{p}_2$	Proposed estimator
$\hat{\mu}$	Unbiased estimator of mean
$\hat{\sigma}^2$	Unbiased estimator of variance
erf	Error function
$\mathcal{N}(\cdot, \cdot)$	Gaussian distribution
χ^2_{N-1}	Chi-square distribution with $N-1$ degrees of freedom
$\Pr(\cdot)$	Probability Operator
$\mathbb{E}[\cdot]$	Expectation Operator

$\mathbb{VAR}[\cdot]$	Variance Operator
$\boldsymbol{x}$	Control parameter vector
$\bar{f}(\boldsymbol{x})$	Empirical estimate of yield at $\boldsymbol{x}$
$\alpha_{\mathrm{EI}}(x \mid \bar{\boldsymbol{f}})$	Acquisition function based on EI given observations $\bar{\boldsymbol{f}}$
$\mathrm{GPR}_f(x \mid \boldsymbol{f})$	Distribution of f modeled by Gaussian Process Regression given observations $\boldsymbol{f}$

Appendix A. Derivation of the Confidence Interval for the Newly Proposed Yield Estimator

While the main text of our paper discusses the accuracy of the newly proposed yield estimator, its confidence interval is also important from a practical application perspective. In this appendix, we present the derivation of the confidence interval for the estimated yield using the newly proposed method. This derivation is supported with numerical experiments as well.

Appendix A.1. Confidence Interval for Newly Proposed Yield Estimator

Let V denote a random variable describing the proposed estimator under the assumptions outlined in Section 2.1-C, where the quality characteristic y follows a Gaussian distribution. Then, V is expressed by (10), i.e.,

$$V = F(a, b, Y, Z), \tag{A1}$$

where

$$\begin{aligned} Z &\sim N\left(\mu_*, \tfrac{\sigma_*^2}{N}\right) \\ \tfrac{Y}{\sigma_*^2(N-1)} &\sim \chi^2_{N-1}. \end{aligned} \tag{A2}$$

By substituting μ_* and σ_*^2 with the corresponding estimates $\hat{\mu}$ and $\hat{\sigma}^2$, respectively, we can approximate the random variable V with V' as

$$V' = F(a, b, Y', Z') \ , \tag{A3}$$

where

$$\begin{aligned} Z' &\sim N\left(\hat{\mu}, \tfrac{\hat{\sigma}^2}{N}\right) \\ \tfrac{Y'}{\hat{\sigma}^2(N-1)} &\sim \chi^2_{N-1}. \end{aligned} \tag{A4}$$

Let us denote $\hat{p}'(y)$ as the probability density function of V. Note that Monte Carlo sampling is necessary to evaluate $\hat{p}'(y)$, as its analytical form is intractable. Let α denote the significance level. Then, it is possible to define the confidence interval as

$$[\hat{p}_2 - h_\alpha, \hat{p}_2 + h_\alpha] \ , \tag{A5}$$

where

$$h_\alpha = \min_{h>0} \left\{ h \mid \textstyle\int_{\hat{p}_2-h}^{\hat{p}_2+h} \hat{p}_2'(y)dy \geq 1-\alpha \right\} \ . \tag{A6}$$

Appendix A.2. Numerical Evaluation of the Confidence Interval

We now examine the confidence intervals of the yield estimates. The experimental settings are the same as those described in Section 3.1-A,B-2 of the main text. For the confidence interval of $\hat{p}_1$, the Wald interval [50] is used, while for the confidence interval of $\hat{p}_2$, the method described above is employed, with the number of samples for Monte Carlo sampling for $\hat{p}'(y)$ set to 10^3. The significance level is set to be $\alpha = 0.05$. The confidence intervals in Section 3.1-A,B-2 are shown in Figure A1 and Figure A2 respectively.

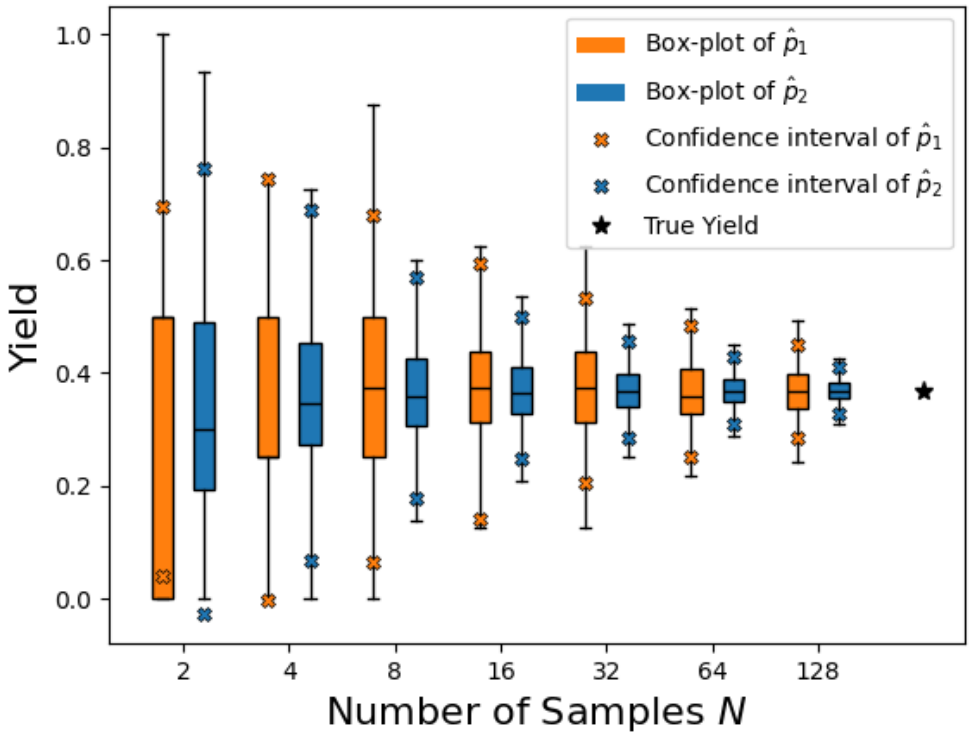


Figure A1. The confidence intervals of the classical estimator $\hat{p}_1$ and the proposed estimator $\hat{p}_2$ for the evaluation setting in Section 3.1-A of the main text. Sample sizes [2, 4, 8, 16, 32, 64, 128] and the significance level $\alpha = 0.05$. We repeated the 10^4 experiments, created the box plot and took the average for the upper and lower bounds of the confidence intervals.

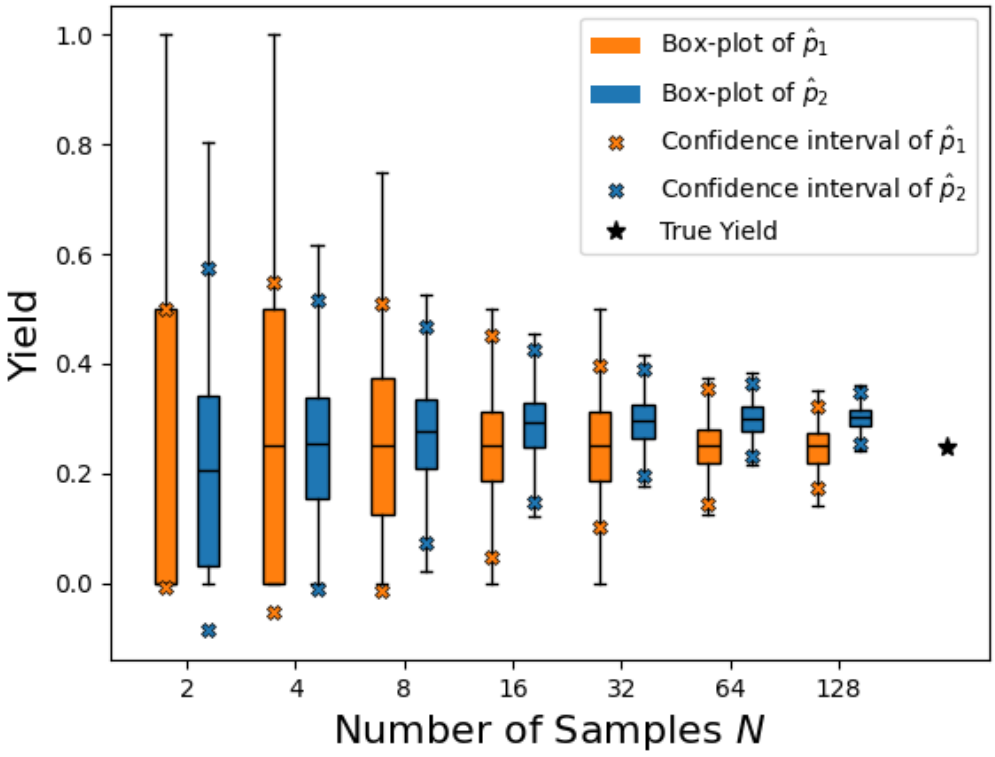


Figure A2. The confidence intervals of the classical estimator $\hat{p}_1$ and the proposed estimator $\hat{p}_2$ for the evaluation setting in Section 3.1-B-1 of the main text. Sample sizes [2, 4, 8, 16, 32, 64, 128] and the significance level $\alpha = 0.05$. We repeated the 10^4 experiments, created the box plot and took the average for the upper and lower bounds of the confidence intervals.

As shown in Figure A1, the quality characteristics follow a Gaussian distribution. In this case, the superiority of the proposed estimator $\hat{p}_2$ is clearly visible. When the sample size N is small, the estimates are less accurate and the confidence intervals are wider.

In particular, the confidence interval of the classical estimator is unstable when $N = 2$ and $N = 4$ since the Wald interval is based on the central limit theorem, which is not valid for small N. As the sample size N increases, we observe that both the confidence intervals and estimation variability of the proposed estimator $\hat{p}_2$ are smaller than those of the classical estimator $\hat{p}_1$.

As shown in Figure A2, the quality characteristics do not follow a Gaussian distribution. Although, when the sample size N is small, the estimates of the proposed estimator $\hat{p}_2$ are closer to the true yield than those of the classical estimator $\hat{p}_1$, the confidence intervals of the classical estimator $\hat{p}_1$ are clearly superior to those of the newly proposed estimator $\hat{p}_2$, especially when the sample size N is sufficiently large. As the sample size N increases, we observe that the confidence intervals and estimation variability of both estimators become smaller. Nevertheless, the confidence interval and the estimates of the proposed estimator $\hat{p}_2$ diverge from the true yield since the assumption of Gaussianity does not hold. On the other hand, the confidence interval and estimates of the classical estimator $\hat{p}_1$ converge to the true yield as the sample size N increases.

Appendix B. Proof of Lemma 1

In this appendix, we derive the following lemma needed for proving Lemma 1.

Lemma A1 (asymptotic expectation and variance of $\hat{p}_2$)**.** *Under the same premises as in Lemma 1, the mean and variance of* $\hat{p}_2$ *are represented as follows:*

$$\begin{aligned}\mathbb{E}[\hat{p}_2] &= p_* + \frac{\sigma_*^{-1}}{4N\sqrt{2\pi}} e^{-\frac{1}{2\sigma_*^2}(a-\mu_*)^2}\left(\sigma_*^{-2}(a-\mu_*)^3 - (a-\mu_*)\right) \\ &\quad - \frac{\sigma_*^{-1}}{4N\sqrt{2\pi}} e^{-\frac{1}{2\sigma_*^2}(b-\mu_*)^2}\left(\sigma_*^{-2}(b-\mu_*)^3 - (b-\mu_*)\right) + \mathcal{O}\left(\frac{1}{N^2}\right)\end{aligned} \tag{A7}$$

$$\begin{aligned}\mathbb{VAR}[\hat{p}_2] &= \frac{1}{4\pi\sigma_*^2 N} e^{-\frac{1}{\sigma_*^2}(a-\mu_*)^2}\left(2\sigma_*^2 + (a-\mu_*)^2\right) + \frac{1}{4\pi\sigma_*^2 N} e^{-\frac{1}{\sigma_*^2}(b-\mu_*)^2}\left(2\sigma_*^2 + (b-\mu_*)^2\right) \\ &\quad - \frac{1}{4\pi\sigma_*^2 N} e^{-\frac{1}{2\sigma_*^2}\left((a-\mu_*)^2+(b-\mu_*)^2\right)}\left(4\sigma_*^2 + 2(a-\mu_*)(b-\mu_*)\right) + \mathcal{O}\left(\frac{1}{N^2}\right)\end{aligned} \tag{A8}$$

Proof of Lemma A1. Let $F_{a,b}(x,y) : \mathrm{R} \times \mathrm{R}^+ \to [0,1]$ be a function whose value is the probability that a random variable $Y \sim \mathcal{N}(x,y)$ falls within an interval $[a,b]$. The function $F_{a,b}(x,y)$ is expressed as

$$F_{a,b}(x,y) = \frac{1}{2}\left(\mathrm{erf}\left(\frac{b-x}{\sqrt{2y}}\right) - \mathrm{erf}\left(\frac{a-x}{\sqrt{2y}}\right)\right).$$

Note that $\hat{p}_2 = F_{a,b}\left(\hat{\mu}, \hat{\sigma}^2\right)$. Let us now apply Taylor expansion to $\hat{p}_2$ around $\left(\mu_*, \sigma_*^2\right)$, noting that since $\mathrm{erf}(\cdot)$ is analytical, as is $y^{-\frac{1}{2}}$ for $y \in \mathrm{R}^+$, the Taylor series expansion of $F_{a,b}\left(\hat{\mu}, \hat{\sigma}^2\right)$ around $\left(\mu_*, \sigma_*^2\right)$ converges pointwise to $F_{a,b}$ for any $(x,y) \in \mathrm{R} \times \mathrm{R}^+$.

The partial derivatives of F become as follows:

$$\frac{\partial F_{a,b}(x,y)}{\partial x} = \frac{y^{-\frac{1}{2}}}{\sqrt{2\pi}}\left(\mathrm{e}^{-\frac{1}{2y}(a-x)^2} - \mathrm{e}^{-\frac{1}{2y}(b-x)^2}\right) \tag{A9}$$

$$\frac{\partial F_{a,b}(x,y)}{\partial y} = \frac{y^{-\frac{3}{2}}}{2\sqrt{2\pi}}\left((a-x)\mathrm{e}^{-\frac{1}{2y}(a-x)^2} - (b-x)\mathrm{e}^{-\frac{1}{2y}(b-x)^2}\right) \tag{A10}$$

$$\frac{\partial^2 F_{a,b}(x,y)}{\partial x^2} = \frac{y^{-\frac{3}{2}}}{\sqrt{2\pi}}\left((a-x)\mathrm{e}^{-\frac{1}{2y}(a-x)^2} - (b-x)\mathrm{e}^{-\frac{1}{2y}(b-x)^2}\right) \tag{A11}$$

$$\frac{\partial^2 F_{a,b}(x,y)}{\partial y^2} = \frac{y^{-\frac{7}{2}}}{4\sqrt{2\pi}}\left(\left((a-x)^3 - 3(a-x)y\right)\mathrm{e}^{-\frac{1}{2y}(a-x)^2} - \left((b-x)^3 - 3(b-x)y\right)\mathrm{e}^{-\frac{1}{2y}(b-x)^2}\right) \tag{A12}$$

Since $\hat{\mu} \sim \mathcal{N}\left(\mu_*, \frac{\sigma_*^2}{N}\right)$, the moments of $\hat{\mu}$ are as follows:

$$\mathbb{E}[\hat{\mu}] = \mu_*;\ \mathbb{VAR}[\hat{\mu}] = \frac{\sigma_*^2}{N}; \mathbb{E}\left[(\hat{\mu}-\mu_*)^{2m+1}\right] = 0;\ \mathbb{E}\left[(\hat{\mu}-\mu_*)^{2m}\right] = \mathcal{O}\left(\frac{1}{N^m}\right) \tag{A13}$$

Furthermore, since $\frac{(N-1)\hat{\sigma}^2}{\sigma_*^2} \sim \chi^2_{N-1}$, the central moment of $\hat{\sigma}^2$ is as follows:

$$\mathbb{E}\left[\left(\hat{\sigma}^2 - \sigma_*^2\right)^m\right] = \left(\frac{\sigma_*^2}{(N-1)}\right)^m \mathbb{E}\left[(V - \mathbb{E}[V])^m\right] \tag{A14}$$

where V is a random variable which follows a χ^2_{N-1} distribution. From the recurrence relation of the central moments for χ^2_{N-1} [51], the m-th central moment is

$$\mathbb{E}\left[(V-\mathbb{E}[V])^2\right] = 2N; \quad \mathbb{E}\left[(V-\mathbb{E}[V])^{2m}\right] = \mathbb{E}\left[(V-\mathbb{E}[V])^{2m+1}\right] = \mathcal{O}(N^m).$$

Therefore,

$$\mathbb{E}\left[\left(\hat{\sigma}^2 - \sigma_*^2\right)\right] = 0;\ \mathbb{E}\left[\left(\hat{\sigma}^2 - \sigma_*^2\right)^2\right] = \frac{2\sigma_*^4 N}{(N-1)^2}\quad , \tag{A15}$$

and for $s \geq 1$,

$$\mathbb{E}\left[\left(\hat{\sigma}^2 - \sigma_*^2\right)^{2s}\right] = \mathcal{O}\left(\frac{1}{N^s}\right); \mathbb{E}\left[\left(\hat{\sigma}^2 - \sigma_*^2\right)^{2s+1}\right] = \mathcal{O}\left(\frac{1}{N^s}\right)\ . \tag{A16}$$

Since $\hat{\mu}$ and $\hat{\sigma}^2$ are independent, for any $s, t \geq 1$, we also have

$$\mathbb{E}\left[(\hat{\mu}-\mu_*)^{2s-1}\left(\hat{\sigma}^2 - \sigma_*^2\right)^t\right] = 0 \tag{A17}$$

$$\mathbb{E}\left[(\hat{\mu}-\mu_*)^{2s}\left(\hat{\sigma}^2 - \sigma_*^2\right)^{2t}\right] = \mathcal{O}\left(\frac{1}{N^{s+t}}\right) \tag{A18}$$

$$\mathbb{E}\left[(\hat{\mu}-\mu_*)^{2s}\left(\hat{\sigma}^2 - \sigma_*^2\right)^{2t+1}\right] = \mathcal{O}\left(\frac{1}{N^{s+t}}\right)\ . \tag{A19}$$

Focusing on the coefficients related to the factor of $\frac{1}{N}$, (63)–(69) lead to the following expression for the Taylor expansion of $\hat{p}_2 = \mathbb{E}\left[F\left(\hat{\mu}, \hat{\sigma}^2\right)\right]$ around $\left(\mu_*, \sigma_*^2\right)$:

$$\begin{aligned}\mathbb{E}[\hat{p}_2] &= F_{a,b}\left(\mu_*, \sigma_*^2\right) + \tfrac{1}{2}\mathbb{E}\left[(\hat{\mu}-\mu_*)^2\right]\left[\tfrac{\partial^2 F_{a,b}(x,y)}{\partial x^2}\right]_{x=\mu_*, y=\sigma_*^2} + \tfrac{1}{2}\mathbb{E}\left[\left(\hat{\sigma}^2 - \sigma_*^2\right)^2\right]\left[\tfrac{\partial^2 F_{a,b}(x,y)}{\partial y^2}\right]_{x=\mu_*, y=\sigma_*^2} + \mathcal{O}\left(\tfrac{1}{N^2}\right) \\ &= F\left(\mu_*, \sigma_*^2\right) + \tfrac{1}{2}\tfrac{\sigma_*^2}{N}\left[\tfrac{\partial^2 F_{a,b}(x,y)}{\partial x^2}\right]_{x=\mu_*, y=\sigma_*^2} + \tfrac{1}{2}\tfrac{2\sigma_*^4}{N}\left[\tfrac{\partial^2 F_{a,b}(x,y)}{\partial y^2}\right]_{x=\mu_*, y=\sigma_*^2} + \mathcal{O}\left(\tfrac{1}{N^2}\right)\end{aligned} \tag{A20}$$

Similarly,

$$\begin{aligned}\mathbb{VAR}[\hat{p}_2] &= \mathbb{VAR}[\hat{\mu}]\left(\left[\tfrac{\partial F_{a,b}(x,y)}{\partial x}\right]_{x=\mu_*, y=\sigma_*^2}\right)^2 + \mathbb{VAR}\left[\hat{\sigma}^2\right]\left(\left[\tfrac{\partial F_{a,b}(x,y)}{\partial y}\right]_{x=\mu_*, y=\sigma_*^2}\right)^2 + \mathcal{O}\left(\tfrac{1}{N^2}\right) \\ &= \tfrac{\sigma_*^2}{N}\left(\left[\tfrac{\partial F_{a,b}(x,y)}{\partial x}\right]_{x=\mu_*, y=\sigma_*^2}\right)^2 + \tfrac{2\sigma_*^4}{N}\left(\left[\tfrac{\partial F_{a,b}(x,y)}{\partial y}\right]_{x=\mu_*, y=\sigma_*^2}\right)^2 + \mathcal{O}\left(\tfrac{1}{N^2}\right).\end{aligned} \tag{A21}$$

By substituting (A9)–(A12) into (A20) and (A21), we obtain Equations (A7) and (A8), thus deriving Lemma A1. □

Note regarding Lemma A1. Since $\mathbb{E}\left[(\hat{p}_2 - p_*)^2\right] = (\mathbb{E}[\hat{p}_2] - p_*)^2 + \mathbb{VAR}[\hat{p}_2]$, Lemma 1 is easily derived from Lemma A1.

Appendix C. Proof of Theorem 1 and Proposition 1

Due to the possibility of transforming any Gaussian random variable into a standard normal random variable, without loss of generality, it is sufficient to consider Theorem 1 only for the case in which y follows the standard normal distribution. Let us denote $S_1 = \left\{(a,b) \in \mathrm{R}^2 \middle| 0 \le a \le b\right\}$ and $S_2 = \left\{(a,b) \in \mathrm{R}^2 \middle| a \le 0, -a \le b\right\}$. Let ∂S and S^{o}, respectively, be the boundary and interior of $S_1 \cup S_2$, i.e., $\partial S = \left\{(a,b) \in \mathrm{R}^2 \middle| a \le 0, -a = b\right\} \cup \left\{(a,b) \in \mathrm{R}^2 \middle| 0 \le a, a = b\right\}$ and $S^{\mathrm{o}} = \{(a,b) \in \mathrm{R}^2 | 0 \le a < b\}$ $\cup \left\{(a,b) \in \mathrm{R}^2 | a \le 0, -a < b\right\}$.

Let us define the function $G(a,b) : S_1 \cup S_2 \to \mathbb{R}$ as follows:

$$G(a,b) = \lim_{N\to\infty} \frac{\mathbb{E}\left[(\hat{p}_1 - p_*)^2\right] - \mathbb{E}\left[(\hat{p}_2 - p_*)^2\right]}{1/N}$$

The function $G(a,b)$ is expressed as follows:

$$G(a,b) = q(a,b)(1-q(a,b)) - \frac{1}{4\pi}\left((2+a^2)\mathrm{e}^{-a^2} + (2+b^2)\mathrm{e}^{-b^2} - (4+2ab)\mathrm{e}^{\frac{-(a^2+b^2)}{2}}\right)$$

where:

$$q(a,b) = \frac{1}{2}\left(\mathrm{erf}\left(\frac{b}{\sqrt{2}}\right) - \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right)$$

We aim to show that $G(a,b) \ge 0\ \forall (a,b) \in S_1 \cup S_2$. In order to do that, let us analyze the behavior of $G(a,b)$ via analysis of its partial derivatives. The partial derivatives of $G(a,b)$ are as follows:

$$\frac{\partial G(a,b)}{\partial a} = \frac{\mathrm{e}^{-\frac{a^2}{2}}}{2\pi}\left(-\sqrt{2\pi}(1-2q(a,b)) - \left(2a + a^2b - b\right)\mathrm{e}^{-\frac{b^2}{2}} + \left(a + a^3\right)\mathrm{e}^{-\frac{a^2}{2}}\right)$$

$$\frac{\partial G(a,b)}{\partial b} = \frac{\mathrm{e}^{-\frac{b^2}{2}}}{2\pi}\left(\sqrt{2\pi}(1-2q(a,b)) - \left(2b + ab^2 - a\right)\mathrm{e}^{-\frac{a^2}{2}} + \left(b + b^3\right)\mathrm{e}^{-\frac{b^2}{2}}\right)$$

Let us introduce $g(a,b)$ and $h(a,b)$ as simpler functions whose signs are equivalent to $\frac{\partial G(a,b)}{\partial a}$ and $\frac{\partial G(a,b)}{\partial b}$, respectively. Namely, we define $g(a,b)$ and $h(a,b)$ as

$$g(a,b) = -\sqrt{2\pi}(1-2q(a,b)) - \left(2a + a^2b - b\right)\mathrm{e}^{-\frac{b^2}{2}} + \left(a + a^3\right)\mathrm{e}^{-\frac{a^2}{2}},$$

$$h(a,b) = \sqrt{2\pi}(1-2q(a,b)) - \left(2b + ab^2 - a\right)\mathrm{e}^{-\frac{a^2}{2}} + \left(b + b^3\right)\mathrm{e}^{-\frac{b^2}{2}}.$$

The partial derivatives of $g(a,b)$ and $h(a,b)$ are as follows:

$$\frac{\partial g(a,b)}{\partial a} = -\left(\left(a^4 - 2a^2 + 1\right)\mathrm{e}^{-\frac{a^2}{2}} + 2(1+ab)\mathrm{e}^{-\frac{b^2}{2}}\right)$$

$$\frac{\partial g(a,b)}{\partial b} = \left(a^2b^2 - (b-a)^2 + 3\right)\mathrm{e}^{-\frac{b^2}{2}}$$

$$\frac{\partial h(a,b)}{\partial a} = \left(a^2b^2 - (b-a)^2 + 3\right)\mathrm{e}^{-\frac{a^2}{2}}$$

$$\frac{\partial h(a,b)}{\partial b} = -\left(\left(b^4 - 2b^2 + 1\right)\mathrm{e}^{-\frac{b^2}{2}} + 2(1+ab)\mathrm{e}^{-\frac{a^2}{2}}\right)$$

Let us use $g_{b=\infty}(a)$ to denote $\lim_{b\to\infty} g(a,b)$. The function $g_{b=\infty}(a)$ and its derivative are expressed as below.

$$g_{b=\infty}(a) = \lim_{b\to\infty} g(a,b) = -\sqrt{2\pi}\mathrm{erf}\left(\frac{a}{\sqrt{2}}\right) + \left(a + a^3\right)\mathrm{e}^{-\frac{a^2}{2}}$$

$$\frac{d}{da} g_{b=\infty}(a) = -\left(a^4 - 2a^2 + 1\right)\mathrm{e}^{-\frac{a^2}{2}}$$

Obviously, $\frac{d}{da} g_{b=\infty}(a) = -\left(a^4 - 2a^2 + 1\right)\mathrm{e}^{-\frac{a^2}{2}} \leq 0$.

In order to prove that $0 \leq G(a,b)$ over $S_1 \cup S_2$, we will prove the following statements.

i. The function $G(a,b)$ is not negative over the boundary ∂S of $S_1 \cup S_2$.
ii. $\lim\limits_{b\to+\infty} G(a,b) \geq 0 \quad \forall a \in R. \geq 0.$
iii. There exist no local minima of $G(a,b)$ over the interior S^{o} of $S_1 \cup S_2$.

Statements (i) and (ii) will be shown via Lemma A2 and Lemma A3, respectively, after which, statement (iii) will be proven.

Lemma A2. *Non-negativity of $G(a,b)$ over ∂S:*

$$\forall\ (a,b) \in \partial S, G(a,b) \geq 0.$$

Proof of Lemma A2. The proof of this lemma will be conducted by showing that

$$\lim_{b\to a} G(a,b) = 0;\ \min_{0\leq b} G(-b,b) = 0,$$

and that

$$\max_{0\leq b} G(-b,b) = G(-b_*,b_*) > 0,$$

where b_* is the unique solution that satisfies $\sqrt{2\pi}\left(1 - 2\mathrm{erf}\left(\frac{b_*}{\sqrt{2}}\right)\right) - \left(2b_* - 2b_*^3\right)e^{\frac{-b_*^2}{2}} = 0$.

One can easily confirm that $\lim_{b\to a} G(a,b) = 0$. Next, we show that $\min_{0\leq b} G(-b,b) = 0$ and that $\max_{0\leq b} G(-b,b) > 0$. One can obtain the following:

$$G(-b,b) = \left(1 - \mathrm{erf}\left(\frac{b}{\sqrt{2}}\right)\right)\mathrm{erf}\left(\frac{b}{\sqrt{2}}\right) - \frac{1}{\pi}b^2\mathrm{e}^{-b^2}$$

$$\frac{dG(-b,b)}{db} = \frac{\mathrm{e}^{\frac{-b^2}{2}}}{\pi}\left(\sqrt{2\pi}\left(1 - 2\mathrm{erf}\left(\frac{b}{\sqrt{2}}\right)\right) - \left(2b - 2b^3\right)\mathrm{e}^{\frac{-b^2}{2}}\right)$$

Let us denote $\ell(b) = \left(\sqrt{2\pi}\left(1 - 2\mathrm{erf}\left(\frac{b}{\sqrt{2}}\right)\right) - \left(2b - 2b^3\right)\mathrm{e}^{\frac{-b^2}{2}}\right)$. Note that the signs of $\ell(b)$ are equivalent to the signs of $\frac{dG(-b,b)}{db}$ and note that $\frac{d}{db}\ell(b) = -2\left(b^2 - 1\right)\left(b^2 - 3\right)\mathrm{e}^{-b^2/2}$. It is easy to confirm that

$$\ell(0) = \sqrt{2\pi} > 0$$

$$\ell(1) = \sqrt{2\pi}\left(1 - 2\mathrm{erf}\left(\frac{1}{\sqrt{2}}\right)\right) < 0$$

$$\ell\left(\sqrt{3}\right) = \sqrt{2\pi}\left(1 - 2\mathrm{erf}\left(\sqrt{\frac{3}{2}}\right)\right) + 4\sqrt{3}\mathrm{e}^{-\frac{3}{2}} < 0.$$

Therefore, it has been shown that unique b_* exists in $(0,1)$ such that $G(-b,b)$ increases on the interval $[0,b_*]$ and decreases on the interval $[b_*,\infty)$. Since

$$G(0,0) = 0 \quad \text{and} \lim_{b\to\infty} G(-b,b) = 0,$$

$G(-b,b) \geq 0$ holds. Furthermore, $\max_{0\leq b} G(-b,b) = G(-b_*,b_*)$ and b_* satisfies

$$\sqrt{2\pi}\left(1-2\mathrm{erf}\left(\frac{b_*}{\sqrt{2}}\right)\right)-\left(2b_*-2b_*^3\right)\mathrm{e}^{\frac{-b_*^2}{2}}=0.$$

□

Lemma A3. *Non-negativity of $G(a,b)$ at infinity:*

$$\lim_{b\to+\infty} G(a,b) \geq 0 \quad \forall a \in R.$$

Proof of Lemma A3. Let $G_{b=\infty}(a)$ denote $\lim_{b\to\infty} G(a,b)$. $G_{b=\infty}(a)$ is expressed as

$$G_{b=\infty}(a) = \tfrac{1}{4}\left(1-\mathrm{erf}^2\left(\tfrac{a}{\sqrt{2}}\right)\right)-\tfrac{1}{4\pi}(2+a^2)\mathrm{e}^{-a^2}.$$

The derivative $\frac{d}{da}G_{b=\infty}(a)$ is expressed as

$$\tfrac{d}{da}G_{b=\infty}(a) = \tfrac{\mathrm{e}^{\frac{-a^2}{2}}}{2\pi}\left((a^3+a)\mathrm{e}^{\frac{-a^2}{2}}-\sqrt{2\pi}\mathrm{erf}\left(\tfrac{a}{\sqrt{2}}\right)\right).$$

It is easy to see that

$$\tfrac{d}{da}\left((a^3+a)\mathrm{e}^{\frac{-a^2}{2}}-\sqrt{2\pi}\mathrm{erf}\left(\tfrac{a}{\sqrt{2}}\right)\right) = -\mathrm{e}^{-\frac{a^2}{2}}\left(a^2-1\right)^2 \leq 0.$$

Also, it is easy to see that $\frac{d}{da}G_{b=\infty}(0)=0$. Therefore,

$$\begin{cases} \frac{d}{da}G_{b=\infty}(a) > 0 & (a<0) \\ \frac{d}{da}G_{b=\infty}(a) < 0 & (0<a) \end{cases}.$$

In other words, $G_{b=\infty}(a)$ has its maximum at $a=0$. It is easy to confirm that

$$G_{b=\infty}(0) = \tfrac{1}{4}-\tfrac{1}{4}\tfrac{2}{\pi} > 0$$

and

$$\lim_{a\to\infty} G_{b=\infty}(a) = \lim_{a\to-\infty} G_{b=\infty}(a) = 0 \ .$$

Hence, $G_{b=\infty}(a) \geq 0$ for all a. □

To prove statement (iii), one should note that the necessary conditions for (a,b) to become a local minimum are as follows:

$$g(a,b)=0;\ \tfrac{\partial g(a,b)}{\partial a} \geq 0$$

$$h(a,b)=0;\ \tfrac{\partial h(a,b)}{\partial b} \geq 0$$

Let us now introduce the following three sets: $S_3 = \left\{(a,b)\in \mathrm{R}\middle| a\leq -\sqrt{3}, -a<b\right\}$, $S_4 = \left\{(a,b)\in \mathrm{R}^2\middle| -\sqrt{3}\leq a<-1, -a<b\right\}$, and $S_5 = \left\{(a,b)\in \mathrm{R}^2\middle| -1\leq a<0, -\tfrac{1}{a}\leq b\right\}$. We will first prove that $g(a,b)>0$ for S_3, S_4, and S_5 in Lemma A4, Lemma A5, and Lemma A6, respectively.

Lemma A4. *$g(a,b)>0$ for $(a,b)\in S_3 = \left\{(a,b)\in R^2\middle| a\leq -\sqrt{3}, -a<b\right\}$.*

Proof of Lemma A4. Firstly,

$$g(a,-a) = -\sqrt{2\pi}(1-2q(a,-a)) + \left(2a^3-2a\right)\mathrm{e}^{-\frac{a^2}{2}},$$

and it is easy to confirm that $g\left(-\sqrt{3},\sqrt{3}\right) = 2\sqrt{2\pi}\mathrm{erf}\left(\sqrt{\frac{3}{2}}\right) - \sqrt{2\pi} - 4\sqrt{3}\mathrm{e}^{-3/2} > 0$. Since $\frac{\partial}{\partial a}g(a,-a) = -2\left(a^2-3\right)\left(a^2-1\right)\mathrm{e}^{-a^2/2}$, it is shown that $g(a,-a) \geq 0$ for $\forall a \leq -\sqrt{3}$. In addition, $\frac{\partial g(a,b)}{\partial b}(a,b) \geq 0$ for $(a,b) \in S_3$ holds since

$$\widetilde{g}_b(a,-a) = \left(a^2-1\right)\left(a^2-3\right) \geq 0$$

and

$$\frac{\partial}{\partial b}\widetilde{g}_b(a,b) = 2\left(a^2-1\right)b + 2a \geq 2\left(a^2-1\right)b - 2b = 2\left(a^2-2\right)b > 0.$$

Therefore, $g(a,b) > 0$ for $(a,b) \in S_3$. □

Lemma A5. *$g(a,b) > 0$ for $(a,b) \in S_4 = \left\{(a,b) \in \mathbb{R}^2 \middle| -\sqrt{3} \leq a < -1, -a \leq b\right\}$.*

Proof of Lemma A5. For each fixed $a \in \left[-\sqrt{3},-1\right)$, $g(a,b)$ attains the minimum at $b = b_c(a)$, where

$$b_c(a) = \frac{-a+\sqrt{a^4-3a^2+3}}{a^2-1}.$$

Note that when $a \in \left[-\sqrt{3},-1\right)$,

$$\sqrt{3} \leq b_c(a) \leq \frac{-2a}{a^2-1}$$

holds. Subsequently,

$$2\sqrt{2\pi}q(a,b_c(a)) \geq 2\sqrt{2\pi}q\left(a,\sqrt{3}\right)$$
$$-\left(2a+a^2b_c(a)-b_c(a)\right) \geq 0 \ .$$

Therefore, for $(a,b) \in S_4$,

$$g(a,b) \geq g(a,b_c(c)) \geq -\sqrt{2\pi}\left(1-2q\left(a,\sqrt{3}\right)\right) + \left(a+a^3\right)\mathrm{e}^{-\frac{a^2}{2}}.$$

Here, $\frac{d}{da}\left(-\sqrt{2\pi}\left(1-2q\left(a,\sqrt{3}\right)\right) + \left(a+a^3\right)\mathrm{e}^{-a^2/2}\right) = -\left(a^2-1\right)^2\mathrm{e}^{-a^2/2} \leq 0.$

Hence,

$$\begin{aligned} g(a,b) &\geq g(a,b_c(c)) \\ &\geq -\sqrt{2\pi}\left(1-2q\left(a,\sqrt{3}\right)\right) + \left(a+a^3\right)\mathrm{e}^{-\frac{a^2}{2}} \\ &\geq -\sqrt{2\pi}\left(1-2q\left(-1,\sqrt{3}\right) - 2\mathrm{e}^{-\frac{1}{2}}\right. \\ &> 0. \end{aligned}$$

□

Lemma A6. *$g(a,b) > 0$ for $(a,b) \in S_5 = \left\{(a,b) \in \mathbb{R}^2 \middle| -1 \leq a < 0, -\frac{1}{a} \leq b\right\}$.*

Proof of Lemma A6. When $b \geq -\frac{1}{a}$, the inequality $\frac{\partial g(a,b)}{\partial b} \leq 0$ holds because

$$\widetilde{g}_b\left(a,-\tfrac{1}{a}\right) = 4 - \left(a+\tfrac{1}{a}\right)^2 \leq 0$$

and

$$\frac{\partial}{\partial b}\widetilde{g}_b(a,b) = 2\big((a^2-1)b+a\big) \le 2\Big((a^2-1)\Big(-\tfrac{1}{a}\Big)+a\Big) = \tfrac{2}{a} < 0.$$

Therefore,

$$g(a,b) \ge g_{b=\infty}(a) > g_{b=\infty}(0) = 0 \ .$$

□

In order to prove statement (iii), let us first note that $\frac{\partial g(a,b)}{\partial a} < 0$ when $ab > 1$. Therefore, condition $ab \le -1$ is necessary for a local minimum of $G(a,b)$. It is easy to confirm that the set $S^o \cap \left\{(a,b) \in \mathrm{R}^2 \middle| ab \le -1\right\}$ is a subset of $S_3 \cup S_4 \cup S_5$, which means that a local minimum of $G(a,b)$ would need to be somewhere in the set $S_3 \cup S_4 \cup S_5$. However, Lemmas A4–A6 show that no points $(a,b) \in S_3 \cup S_4 \cup S_5$ can be a local minimum of $G(a,b)$. Therefore, statement (iii) holds.

Proof of Theorem 1. Statement (iii) shows that no points (a,b) in S^o can be the minimum of $G(a,b)$. Thus, the minimum of $G(a,b)$ exists only on ∂S, or at infinity points. To that end, statement (i) and statement (ii) show that $G(a,b) \ge 0$ on ∂S and at infinity points, respectively. Hence, $G(a,b)$ is non-negative over $S_1 \cup S_2$, which completes the proof of Theorem 1. □

Proof of Proposition 1. When the true mean of the outgoing product quality is in the center of the specification interval, the asymptotic difference of $\hat{p}_1$ and $\hat{p}_2$ is represented as $G(-b,b),\ b \ge 0$. Lemma A2 shows that $\max_{0 \le b} G(-b,b) = G(-b_*,b_*)$, where b_* is the unique solution that satisfies

$$\sqrt{2\pi}\left(1 - 2erf\left(\frac{b_*}{\sqrt{2}}\right)\right) - \left(2b_* - 2b_*^3\right)e^{\frac{-b_*^2}{2}} = 0.$$

This completes the proof of Proposition 1. □

Appendix D. Ablation Study of BO Configurations with Yield Estimators

We conduct an ablation study on the problem introduced in Section 3.1 to isolate the impact of our proposed yield estimator from other BO design choices. We compare Bayesian Optimization using the proposed estimator versus the classical estimator across different acquisition functions and kernel choices. For each configuration, we report quantiles of performance at each iteration. Although acquisition and kernel choices affect absolute performance, the results show that the proposed estimator consistently improves sample efficiency and final yields across the tested configurations. Detailed experimental settings and code are provided in the Supplemental Codebase.

Appendix D.1. Difference in Acquisition Functions

In this subsection, we investigate the effect of the acquisition function. We repeated the experiments from Section 3.1 using Expected Improvement (EI) as the acquisition function; detailed settings are available in the Supplemental Codebase. We consider four GPR configurations: *DIST-EI*, *DIST-NEI*, *PROB-EI*, and *PROB-NEI.* The prefixes *DIST* and *PROB* denote the GPR target as in the main text, while the suffixes *EI* and *NEI* denote the acquisition functions—Expected Improvement and Noisy Expected Improvement. Hence, *DIST-EI* uses estimated distribution parameters with EI; *DIST-NEI* uses estimated distribution parameters with NEI; *PROB-EI* uses empirical yield with EI; and *PROB-NEI* uses empirical yield with NEI. The results are shown in Figure A3 and Table A1. Focusing on acquisition function differences, Noisy EI outperforms standard EI in this problem

setting. This is expected because the task inherently includes observation noise, and Noisy EI explicitly accounts for that noise when evaluating improvement.

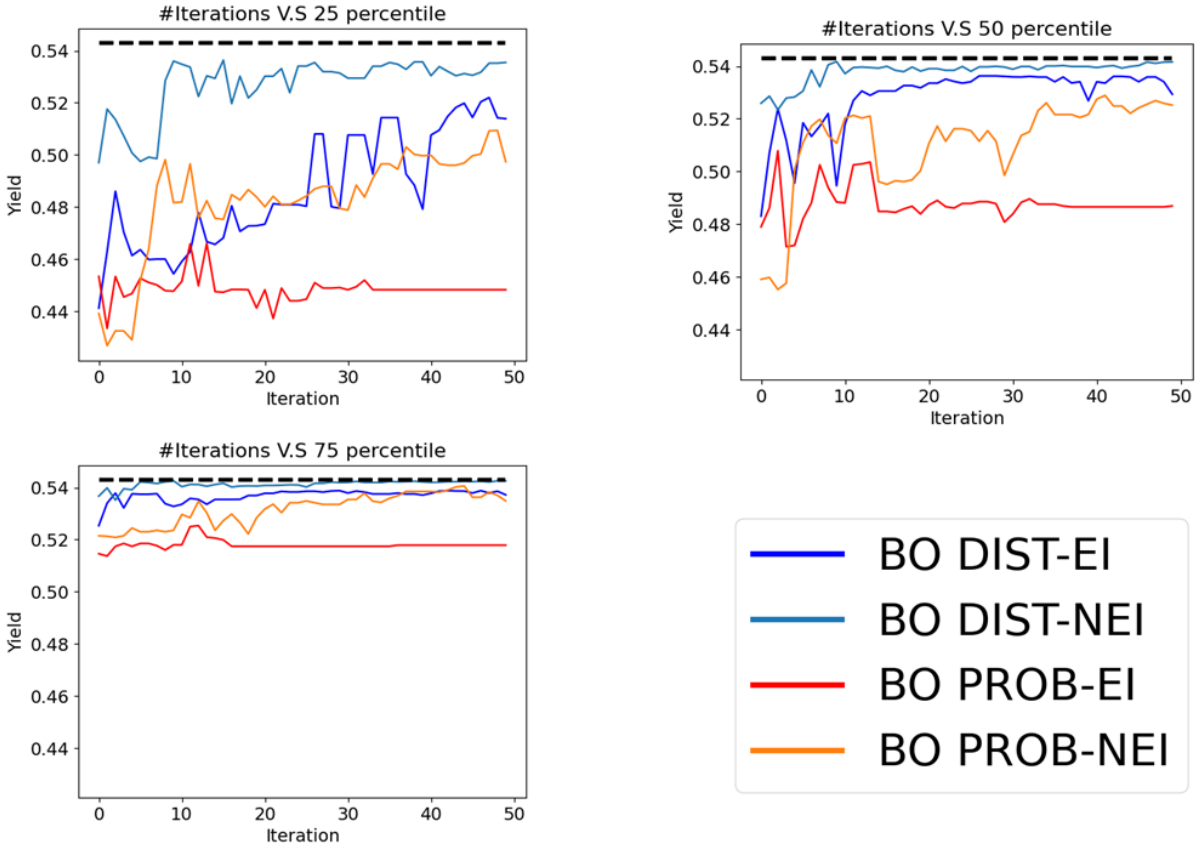


Figure A3. The progression of the 25th, 50th and 75th percentiles of yields obtained for the control parameters suggested over 16 random replications of each algorithm. The black dashed line corresponds to the highest possible yield in this synthetic problem.

Table A2 reports the computation times for the experiments, measured on the standard workstation. As noted in Section 2.2-C, NEI requires substantially more computation than EI because NEI relies on Monte Carlo simulation to evaluate expected improvement. Similarly, the proposed estimator is more computationally demanding than the classical estimator because it uses Monte Carlo methods to compute posterior distributions. In these experiments, NEI required approximately 1.5× the computation time of EI, while GPRs using the proposed yield estimator required about 5× the computation time of the classical yield estimator.

Table A1. Yields suggested by BOo methods as iterations progressed.

Percentile	**25**				**50**			
Method	**BO DIST-EI**	**BO DIST-NEI**	**BO PROB-EI**	**BO PROB-NEI**	**BO DIST-EI**	**BO DIST-NEI**	**BO PROB-EI**	**BO PROB-NEI**
Iteration								
4	0.461	0.501	0.447	0.429	0.496	0.528	0.472	0.501
9	0.454	0.536	0.448	0.482	0.495	0.542	0.488	0.511
14	0.466	0.529	0.447	0.476	0.531	0.539	0.485	0.496
19	0.473	0.525	0.441	0.484	0.532	0.538	0.484	0.5

Table A1. *Cont.*

Percentile	25				50			
Method	**BO DIST-EI**	**BO DIST-NEI**	**BO PROB-EI**	**BO PROB-NEI**	**BO DIST-EI**	**BO DIST-NEI**	**BO PROB-EI**	**BO PROB-NEI**
Iteration								
24	0.481	0.534	0.444	0.482	0.534	0.54	0.488	0.516
29	0.479	0.531	0.449	0.479	0.536	0.54	0.481	0.498
34	0.514	0.534	0.448	0.497	0.536	0.54	0.488	0.526
39	0.479	0.536	0.448	0.5	0.527	0.54	0.487	0.522
44	0.52	0.531	0.448	0.497	0.536	0.54	0.487	0.522
49	0.514	0.535	0.448	0.497	0.529	0.542	0.487	0.525

Percentile	75			
Method	**BO DIST-EI**	**BO DIST-NEI**	**BO PROB-EI**	**BO PROB-NEI**
4	0.538	0.539	0.517	0.524
9	0.533	0.543	0.518	0.524
14	0.535	0.541	0.521	0.524
19	0.537	0.541	0.517	0.528
24	0.538	0.541	0.517	0.534
29	0.539	0.542	0.517	0.534
34	0.538	0.542	0.517	0.534
39	0.537	0.542	0.518	0.539
44	0.539	0.542	0.518	0.541
49	0.537	0.543	0.518	0.535

Table A2. Calculation time for 1 trial (50 iterations)/mean ± STD.

BO DIST-EI	BO DIST-NEI	BO PROB-EI	BO PROB-NEI
23.447 ± 3.705	31.102 ± 3.105	4.670 ± 0.638	6.148 ± 0.609

Appendix D.2. Difference in Kernel Functions

To illustrate differences between kernel functions, we compare the RBF kernel and the exponential kernel. Both are special cases of the Matern kernel: the exponential kernel corresponds to Matern with $\nu = 1/2$, while the RBF kernel is obtained when $\nu \to \infty$. Because ν controls smoothness, this comparison provides an overview of how performance depends on the assumed kernel smoothness.

When comparing variants, we consider four GPR configurations: *DIST-EXP*, *DIST-RBF*, *PROB-EXP*, and *PROB-RBF*. Here, the prefix *DIST* and *PROB* represent the target of GPR as in the main text. The suffix *EXP* denotes the exponential kernel and *RBF* denotes the radial basis function kernel. Thus, *DIST-EXP* uses estimated distribution parameters with an exponential kernel, *DIST-RBF* uses estimated distribution parameters with an *RBF* kernel, *PROB-EXP* uses empirical yield with an exponential kernel, and *PROB-RBF* uses empirical yield with an *RBF* kernel. The results are shown in Figure A4 and Table A3.

When we examine the trajectories at the 25th and 50th percentiles, the RBF kernel appears to perform slightly better than the exponential kernel. Although acceptability (i.e., whether quality characteristics fall within specification) is discontinuous with respect to the quality metrics, the relationships between control parameters and yield are relatively smooth in this problem, as shown in Figure 8. This may partly explain the RBF kernel's slight advantage. Nonetheless, the variation in performance due to the chosen GPR target is larger than that due to the kernel choice, as evidenced by the 25th-percentile trajectory.

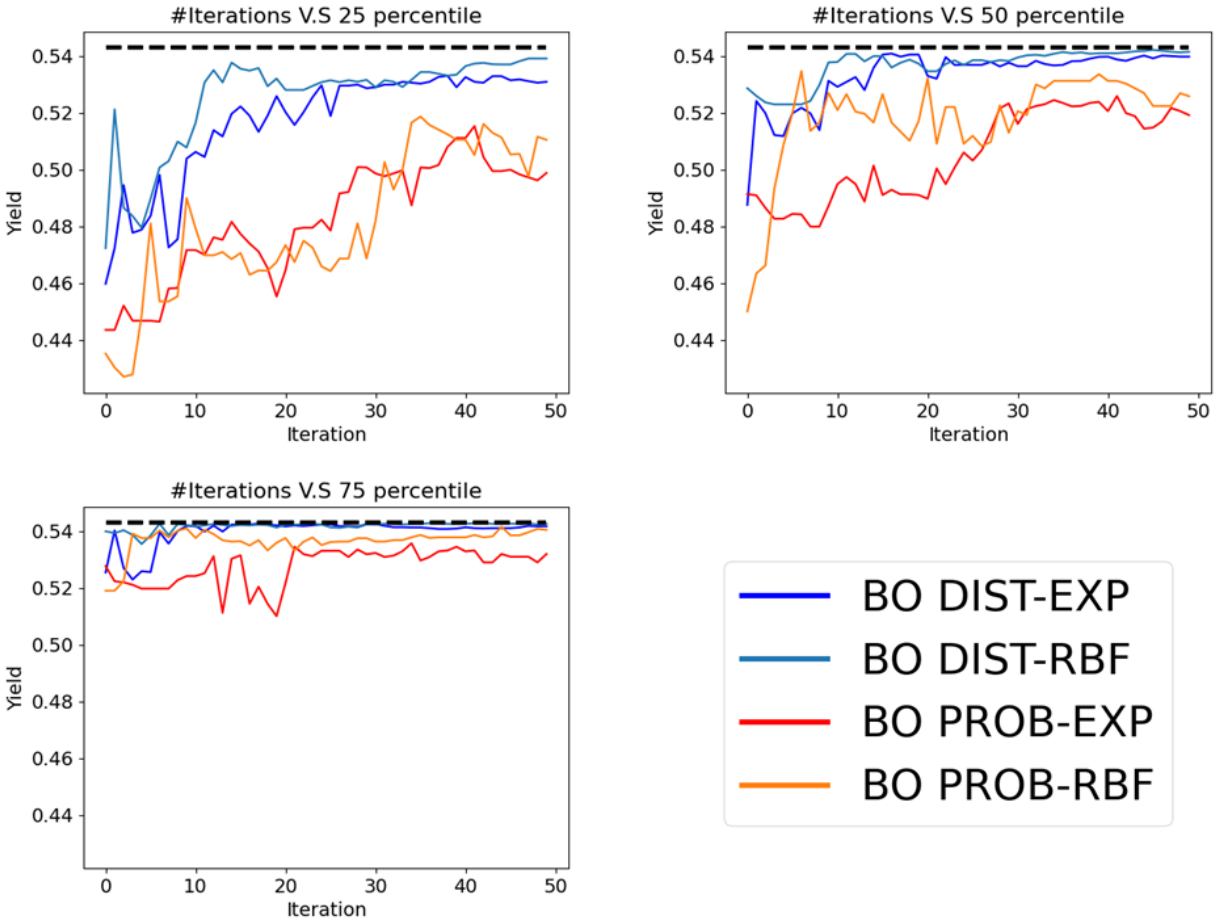


Figure A4. The progression of the 25th, 50th and 75th percentiles of yields obtained for the control parameters suggested over 16 random replications of each algorithm. The black dashed line corresponds to the highest possible yield in this synthetic problem.

Table A3. Yields suggested by BO methods as iterations progressed.

Percentile	**25**				**50**			
Method	**BO DIST-EXP**	**BO DIST-RBF**	**BO PROB-EXP**	**BO PROB-RBF**	**BO DIST-EXP**	**BO DIST-RBF**	**BO PROB-EXP**	**BO PROB-RBF**
Iteration								
4	0.479	0.479	0.447	0.449	0.512	0.523	0.483	0.509
9	0.504	0.508	0.472	0.490	0.531	0.538	0.487	0.527
14	0.520	0.538	0.482	0.468	0.536	0.540	0.501	0.517

Table A3. *Cont.*

Percentile	**25**				**50**			
Method	**BO DIST-EXP**	**BO DIST-RBF**	**BO PROB-EXP**	**BO PROB-RBF**	**BO DIST-EXP**	**BO DIST-RBF**	**BO PROB-EXP**	**BO PROB-RBF**
Iteration								
19	0.526	0.532	0.455	0.467	0.540	0.537	0.491	0.517
24	0.530	0.531	0.482	0.466	0.537	0.537	0.506	0.509
29	0.529	0.532	0.501	0.469	0.538	0.538	0.523	0.513
34	0.531	0.532	0.487	0.517	0.537	0.541	0.524	0.531
39	0.529	0.533	0.511	0.510	0.540	0.541	0.524	0.534
44	0.533	0.537	0.499	0.511	0.540	0.542	0.514	0.527
49	0.531	0.539	0.499	0.510	0.540	0.541	0.519	0.526

Percentile	**75**			
Method	**BO DIST-EXP**	**BO DIST-RBF**	**BO PROB-EXP**	**BO PROB-RBF**
4	0.526	0.535	0.520	0.538
9	0.542	0.542	0.524	0.541
14	0.542	0.542	0.530	0.536
19	0.542	0.541	0.510	0.536
24	0.542	0.542	0.533	0.535
29	0.542	0.542	0.532	0.537
34	0.541	0.542	0.536	0.538
39	0.541	0.542	0.534	0.538
44	0.541	0.543	0.532	0.541
49	0.542	0.543	0.532	0.540

References

1. Brajlih, T.; Valentan, B.; Balic, J.; Drstvensek, I. Speed and accuracy evaluation of additive manufacturing machines. *Rapid Prototyp. J.* **2011**, *17*, 64–75. [CrossRef]
2. Kumar, N.; Kennedy, K.; Gildersleeve, K.; Abelson, R.; Mastrangelo, C.M.; Montgomery, D.C. A review of yield modelling techniques for semiconductor manufacturing. *Int. J. Prod. Res.* **2006**, *44*, 5019–5036. [CrossRef]
3. Katari, M.; Shanmugam, L.; Malaiyappan, J.N.A. Integration of AI and machine learning in semiconductor manufacturing for defect detection and yield improvement. *J. Artif. Intell. Gen. Sci.* **2024**, *3*, 418–431. [CrossRef]
4. Liang, Y.-Z.; Xie, P.; Chan, K. Quality control of herbal medicines. *J. Chromatogr. B* **2004**, *812*, 53–70. [CrossRef]
5. Appleton, T.; Bryan, P.; Contos, D.; Henry, T.R.; Lehmann, P.; Ohorodnik, S.; Reed, D.; Robichaud, C.; Schetter, J.; South, N.; et al. Nonclinical dose formulation: Out of specification investigations. *AAPS J.* **2012**, *14*, 523–529. [CrossRef]
6. Chen, W.; Wiecek, M.M.; Zhang, J. Quality utility—A compromise programming approach to robust design. *J. Mech. Des.* **1999**, *121*, 179–187. [CrossRef]
7. Zang, C.; Friswell, M.; Mottershead, J. A review of robust optimal design and its application in dynamics. *Comput. Struct.* **2005**, *83*, 315–326. [CrossRef]
8. Du, X.; Chen, W. Towards a better understanding of modeling feasibility robustness in engineering design. *J. Mech. Des.* **1999**, *122*, 385–394. [CrossRef]
9. Ochoa, J.S.; Cangellaris, A.C. Random-space dimensionality reduction for expedient yield estimation of passive microwave structures. *IEEE Trans. Microw. Theory Tech.* **2013**, *61*, 4313–4321. [CrossRef]
10. Styblinski, M.A.; Opalski, L.J. Algorithms and software tools for IC yield optimization based on fundamental fabrication parameters. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **1986**, *5*, 79–89. [CrossRef]

11. Dolecek, L.; Qazi, M.; Shah, D.; Chandrakasan, A. Breaking the simulation barrier: SRAM evaluation through norm minimization. In Proceedings of the 2008 IEEE/ACM International Conference on Computer-Aided Design, San Jose, CA, USA, 10–13 November 2008; pp. 322–329.
12. Fang, C.; Yang, F.; Zeng, X.; Li, X. BMF-BD: Bayesian model fusion on bernoulli distribution for efficient yield estimation of integrated circuits. In Proceedings of the 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 1–5 June 2014; pp. 1–6.
13. Ashby, M.F. Multi-objective optimization in material design and selection. *Acta Mater.* **2000**, *48*, 359–369. [CrossRef]
14. Sarkar, D.; Rohani, S.; Jutan, A. Multi-objective optimization of seeded batch crystallization processes. *Chem. Eng. Sci.* **2006**, *61*, 5282–5295. [CrossRef]
15. Shao, J. *Mathematical Statistics*; Springer: New York, NY, USA, 2003.
16. Sano, S.; Kadowaki, T.; Tsuda, K.; Kimura, S. Application of bayesian optimization for pharmaceutical product development. *J. Pharm. Innov.* **2020**, *15*, 333–343. [CrossRef]
17. Cho, S.; Kim, M.; Lee, J.; Han, A.; Na, J.; Moon, I. Multi-objective optimization of explosive waste treatment process considering environment via bayesian active learning. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105463. [CrossRef]
18. Iwama, R.; Kaneko, H. Design of ethylene oxide production process based on adaptive design of experiments and bayesian optimization. *J. Adv. Manuf. Process.* **2021**, *3*, e10085. [CrossRef]
19. Al-Qahtani, K.; Elkamel, A.; Ponnambalam, K. Robust optimization for petrochemical network design under uncertainty. *Ind. Eng. Chem. Res.* **2008**, *47*, 3912–3919. [CrossRef]
20. Rashidi, N.; Wang, Q.; Burgos, R.; Roy, C.; Boroyevich, D. Multi-objective design and optimization of power electronics converters with uncertainty quantification—Part I: Parametric uncertainty. *IEEE Trans. Power Electron.* **2020**, *36*, 1463–1474. [CrossRef]
21. Zhang, Y.; Apley, D.W.; Chen, W. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Sci. Rep.* **2020**, *10*, 4924. [CrossRef]
22. Chen, Y.; Hassani, H.; Krause, A. Near-optimal bayesian active learning with correlated and noisy tests. In Proceedings of the Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 223–231.
23. Cakmak, S.; Astudillo Marban, R.; Frazier, P.; Zhou, E. Bayesian optimization of risk measures. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 20130–20141.
24. Picheny, V.; Moss, H.; Torossian, L.; Durrande, N. Bayesian quantile and expectile optimisation. In Proceedings of the Uncertainty in Artificial Intelligence, Eindhoven, The Netherlands, 1–5 August 2022; pp. 1623–1633.
25. Makarova, A.; Usmanova, I.; Bogunovic, I.; Krause, A. Risk-averse heteroscedastic bayesian optimization. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17235–17245.
26. Daulton, S.; Cakmak, S.; Balandat, M.; Osborne, M.A.; Zhou, E.; Bakshy, E. Robust multi-objective bayesian optimization under input noise. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 4831–4866.
27. Wang, X.; Yan, C.; Ma, Y.; Yu, B.; Yang, F.; Zhou, D.; Zeng, X. Analog circuit yield optimization via freeze–thaw bayesian optimization technique. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2022**, *41*, 4887–4900. [CrossRef]
28. Iwazaki, S.; Inatsu, Y.; Takeuchi, I. Bayesian quadrature optimization for probability threshold robustness measure. *Neural Comput.* **2021**, *33*, 3413–3466. [CrossRef] [PubMed]
29. Inatsu, Y.; Iwazaki, S.; Takeuchi, I. Active learning for distributionally robust level-set estimation. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 4574–4584.
30. Routtenberg, T.; Tabrikian, J. General classes of bayesian lower bounds for outage error probability and MSE. In Proceedings of the 2009 15th IEEE/SP Workshop on Statistical Signal Processing, Cardiff, UK, 31 August–3 September 2009; pp. 69–72.
31. Cabilio, P. Sequential estimation in bernoulli trials. *Ann. Stat.* **1977**, *5*, 342–356. [CrossRef]
32. Rice, J.A. *Mathematical Statistics and Data Analysis*; Thomson/Brooks/Cole: Belmont, CA, USA, 2007; Volume 371.
33. Oehlert, G.W. A note on the delta method. *Am. Stat.* **1992**, *46*, 27–29. [CrossRef]
34. Casella, G.; Berger, R. *Statistical Inference*; CRC Press: Boca Raton, FL, USA, 2024.
35. Zhao, P.; Wu, J.; Liu, Z.; Wang, C.; Fan, R.; Li, Q. Differential private stochastic optimization with heavy-tailed data: Towards optimal rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Philadelphia, PA, USA, 25 February–4 March 2025, AAAI: Washington, DC, USA, 2025; Volume 39, pp. 22795–22803.
36. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer Science & Business Media: New York, NY, USA, 2006.
37. Zhang, Q.; Huang, W.; Jin, C.; Zhao, P.; Shu, Y.; Shen, L.; Tao, D. Multinoulli extension: A lossless yet effective probabilistic framework for subset selection over partition constraints. In Proceedings of the Forty-Second International Conference on Machine Learning, Vancouver, BC, Canada, 13–19 July 2025.
38. Frazier, P.I. A tutorial on bayesian optimization. *arXiv* **2018**, arXiv:1807.02811. [CrossRef]
39. Letham, B.; Karrer, B.; Ottoni, G.; Bakshy, E. Constrained bayesian optimization with noisy experiments. *Bayesian Anal.* **2019**, *14*, 495–519. [CrossRef]

40. Balandat, M.; Karrer, B.; Jiang, D.R.; Daulton, S.; Letham, B.; Wilson, A.G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020*; NeurIPS: San Diego, CA, USA, 2020; Volume 33.
41. Lee, C.; Graves, D.B.; Lieberman, M.A.; Hess, D.W. Global model of plasma chemistry in a high density oxygen discharge. *J. Electrochem. Soc.* **1994**, *141*, 1546. [CrossRef]
42. COMSOL. Coil Optimization of an ICP Reactor. Available online: https://www.comsol.jp/model/coil-optimization-of-an-icp-reactor-96771 (accessed on 10 March 2026).
43. Kokura, H.; Nakamura, K.; Ghanashev, I.P.; Sugai, H. Plasma absorption probe for measuring electron density in an environment soiled with processing plasmas. *Jpn. J. Appl. Phys.* **1999**, *38*, 5262. [CrossRef]
44. Nakamura, K.; Ohata, M.; Sugai, H. Highly sensitive plasma absorption probe for measuring low-density high-pressure plasmas. *J. Vac. Sci. Technol. A* **2003**, *21*, 325–331. [CrossRef]
45. Scharwitz, C.; Böke, M.; Hong, S.-H.; Winter, J. Experimental characterisation of the plasma absorption probe. *Plasma Process. Polym.* **2007**, *4*, 605–611. [CrossRef]
46. Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611. [CrossRef]
47. Sobol, I.M. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 86–112. [CrossRef]
48. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
49. Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **2015**, *71*, 804–818. [CrossRef]
50. Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* **1943**, *54*, 426–482. [CrossRef]
51. Willink, R. Relationships between central moments and cumulants, with formulae for the central moments of gamma distributions. *Commun. Stat.-Theory Methods* **2003**, *32*, 701–704. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.