Adaptive experimental design using Bayesian optimization to improve the cost efficiency of small-plot field trials

Yuji Saikai, Vivak Patel, Lucía Gutiérrez, Brian Luck, Jed Colquhoun, Shawn Conley, and Paul Mitchell

September 24, 2019

Abstract

In most agricultural experiments, factorial design is the standard choice of experimental design. It is intuitive, easy to use, and can be effective in some problems. However, what is hidden in these advantages is inefficiency in choosing factor levels to investigate. This is particularly problematic when the objective of experiments is to estimate the optimal value of some response function, rather than to estimate the overall shape of the response function. In the literature, this distinction is rarely made explicit, and factorial designs are used for both purposes. In this paper, we propose a novel approach, called Bayesian optimization, to adaptively designing agricultural experiments for optimization and demonstrate significant gain in the cost efficiency of small-plot field trials. Results show that the annual difference between two estimated profits by the factorial design and our approach can be as high as \$272/ha despite the fact that our approach uses less than half of the plot numbers used in the factorial design. Considering the efficiency gain in both estimated profits and required number of plots, we think that Bayesian optimization can benefit many agricultural researchers in their own experiments.

Keywords: Experimental design, Cost efficiency, Factorial design, Bayesian optimization, APSIM

1 Introduction

In many branches of the agricultural sciences, small-plot field experimentation is the fundamental technique to determine optimal input use and management practices across different production environments (Mead et al., 1993; Petersen, 1994). For example, small-plot field trials have been used for determining optimal fertilizer application rates (Buresh et al., 2019; Cela et al., 2011; Huang et al., 2008; Zewen Jin et al., 2019; Li et al., 2018; Rens et al., 2018; Storer et al., 2018; H. Wang et al., 2017; W. Wang et al., 2012), fungicide dose rates (Lynch et al., 2017), plant varieties (Huang et al., 2008; Lynch et al., 2017; Storer et al., 2018), seeding rates (Dai et al., 2013), and plant densities (Khan et al., 2017; Ren et al., 2017). Unfortunately, since any experiment is subject to limited resources and incurs economic costs, exhaustive field trials are infeasible. Indeed, most studies including those cited above involve only a handful of levels for each factor, and the total number of factors is very small. When involving many factors, each factor usually allows for only two levels (Orlowski et al., 2016). Thus, in determining optimal management, small-plot field experiments must be carefully designed to maximize scarce resources.

In most cases, small-plot field experiments are based on factorial designs, in which all possible combinations of the levels of every factor are investigated and replicated over several years (Montgomery, 2017). As an example, Zewen Jin et al. (2019) conducted field experiments based on

a factorial design of four biochar rates (0,5,20,40) and four nitrogen fertilizer rates (0,60,90,120), which were replicated from 2011 to 2016. As another example, Lynch et al. (2017) conducted field experiments based on a factorial design of two fungicide types (azole, azole+SDHI), five applications rates (0, 0.25, 0.50, 1.0, 2.0), and three varieties (SR5, SR7, SR8), which were replicated from 2012 to 2015. Although these and many other field experiments that employ factorial designs provide qualitative insights into dependencies between factors and clues about optimal levels, in terms of quantitatively maximizing scarce resources, they are problematic for several reasons:

1. With sufficient domain knowledge, hand-picked factor levels may work. But, they are less likely to work if growing environments and/or factors are considerably different from the existing ones on which such domain knowledge is based.
2. Even with well-chosen factor levels, the fixed number of levels limits exploration of factors that incorporate continuous values.
3. Finally, static experimental designs by definition preclude adaptation of designs that reflects the information obtained in the preceding years to explore more promising values.

In addition to the inefficiency in data collection, the standard approaches to agricultural experimentation suffer from model misspecification that can potentially have a negative impact on profitability. After collecting data, researchers next estimate some statistical model of yield/profit function and maximize this estimated function to determine optimal management choices. The standard approaches typically use relatively simple statistical models of limited flexibility. Therefore, if the underlying function takes a complex shape, which is quite possible in cases involving many variables as we show in our experiments, such models will fail to capture the function's major structure and provide poor guidance on profit maximization. In contrast, our approach uses a sufficiently flexible statistical model to avoid model misspecification.

In summary, small-plot field experiments that use factorial designs for optimizing input use and management practices make inefficient use of scarce resources dedicated to experimentation. Moreover, simple statistical models that typically accompany factorial designs may suffer from model misspecification, further reducing efficiency. In this work, we propose an adaptive, flexible experimental designs for optimization. Our approach, based on Bayesian optimization (BO) (Brochu et al., 2010; Shahriari et al., 2016), reflects information from the preceding years in the current design, so as to make the most of scarce resources, and has the ability to capture potentially complex profit functions.

We demonstrate the advantage of our approach over the traditional factorial designs by using two distinct simulation environments: one constructed from real field-trial data and the other from the existing cropping systems simulator. As a result of our methodology, small-plot field experiments can be conducted more efficiently, allowing farmers to determine optimal input use and management practices with fewer experimental resources and, thereby, reducing the costs.

2 Materials and methods

We compare two regimes of experimental designs, BO design and factorial design, in six different simulation environments. Each simulation environment is characterized by an "oracle," a conceptual device that tells us the crop yield for a set of specific levels of production factors. Using the yield obtained from the oracle, we then calculate a profit using an output price and input costs. Once we set up the simulation environments, we carry out simulations over 1- to 10-year time horizons and compare their performances on profit maximization based on both years of experiment and number of sampling plots required to reach a certain level of profit.

2.1 Oracle and simulation environment

An oracle returns a yield when we query a set of specific values of production factors. It is a yield response function corrupted by random noise. Mathematically, yield (y) is defined as follows:

$$y = f(x) + \varepsilon,$$

where $f : \mathcal{X} \to \mathbb{R}_+$, a function that takes a set of production factors $x \in \mathcal{X}$ and returns a nonnegative yield $y \in \mathbb{R}_+$. Note that we assume $\varepsilon \sim N(0, \sigma^2)$, that is, normally distributed noise with standard deviation of σ. For each yield (y), profit (π) is defined as follows:

$$\pi = py - c \cdot x,$$

where p is an output price and c is a set of input prices. To characterize the effects of simulation environments on the performance of designs, we construct two distinct types of simulation environments: low-dimensional with discrete choice and high-dimensional.

2.1.1 Scenario A (low-dimensional with discrete choice)

This environment is meant to capture a traditional field study that involves a small number of both continuous and discrete factors. We base this simulation study on the following story.

> Imagine a researcher who tries to determine optimal seeding rates and seed treatments of a new seed variety under 16 different soil pHs $\{6.0, 6.1, \ldots, 7.5\}$. There are two types of seed treatments, F (fungicide) and FI (fungicide and insecticide), in addition to UTC (untreated control), where UTC is cheapest and FI is most expensive. Optimality is defined as maximizing the expected profit. Overall, through small-plot field trials, the researcher tries to determine 16 optimal combinations of seeding rate and seed treatment, one for each soil pH. Given the economic costs associated with each trial, the researcher wants to make efficient use of the limited experimental resources. Which combinations in what order should the researcher use to proceed with the experiment?

We consider this to be a reasonable setting, albeit stylized, because the ability to adjust soil pH is usually constrained both financially and physically such as buffer pH (Camberato, 2014). Hence, it is valuable to have a recommendation for each soil pH.

Let x_1 denote seeding rate (in 1,000 seeds/ac), x_2 denote soil pH, and τ denote seed treatment. Then, we have $x_1 \in [0, 150]$, $x_2 \in \{6.0, 6.1, \ldots, 7.5\}$, and $\tau \in \{UTC, F, FI\}$. The oracle for this setting is a collection of three yield response functions (f_τ for each τ). In other words, when we query a combination (τ, x_1, x_2), the oracle returns a noise-corrupted yield,

$$y = f_\tau(x_1, x_2) + \varepsilon.$$

We construct these three response functions using real-world data from soybean field trials (Gaspar et al., 2015). Specifically, for each seed treatment (τ), we fit a nonparametric regression, namely a local linear regression (Hastie et al., 2016). Then, for each yield, we calculate the corresponding profit (π) using a soybean price and seed cost for that treatment. Formally,

$$\pi = py - c_\tau x_1,$$

where p is the soybean price and c_τ is the seed cost for treatment τ, all of which are found in Gaspar et al. (2015).

The following figures depict three profit response surfaces (without noise). As you can see, each surface is not globally concave.

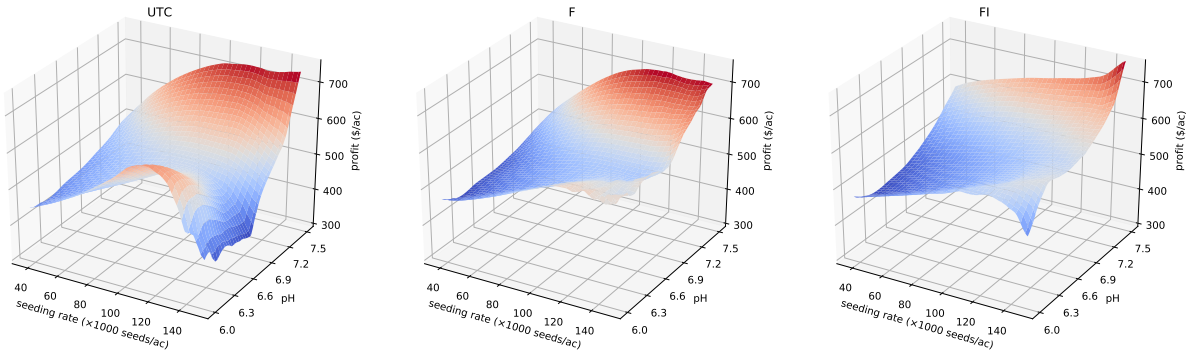


Figure 1: Profit response surfaces

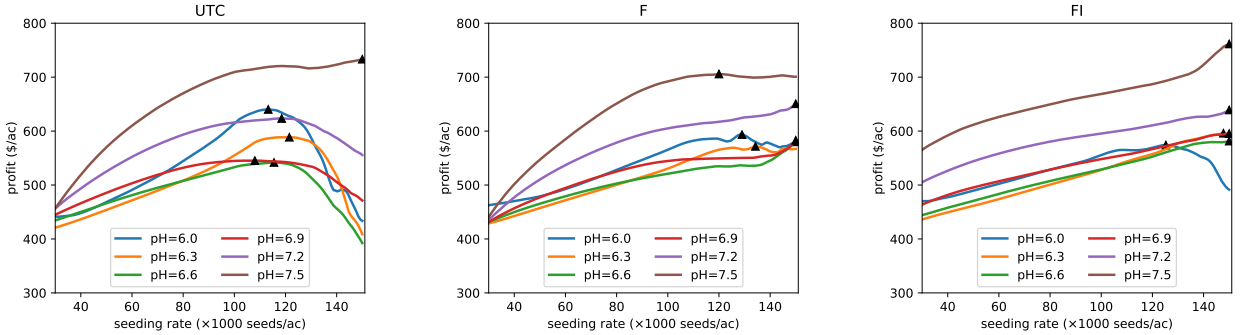


Figure 2: Profit response curves at the selected pHs

▲ indicates the maximum profit for each pH. The following table contains specific values.

pH	6.0	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0	7.1	7.2	7.3	7.4	7.5
Treatment	UTC	UTC	UTC	FI	FI	FI	FI	FI	FI	FI	F	F	F	F	FI	FI
Seeding rate	113	119	112	148	148	147	150	150	150	150	150	150	150	150	150	150

Table 1: Optimal seed treatments and seeding rates for each pH

To investigate the effects of noise, we implement three different noise levels, $\sigma \in \{10, 30, 50\}$, which corresponds to respectively about 10%, 30%, and 50% of the maximum possible yield.

2.1.2 Scenario B (high-dimensional)

This environment is meant to serve as a more realistic scenario in which farmers make many decisions on input use and management practices. In particular, there are more continuous choices than the low-dimensional case. Scenario B is relevant not only in practice but also has attracted attention in the literature (Orlowski et al., 2016). Since there rarely exist datasets densely covering a high-dimensional continuous space, to create a suitable oracle, we make use of the Agricultural Production Systems sIMulator (APSIM), a highly advanced simulator of agricultural systems (Holzworth et al., 2014) being widely used for generating hypothetical datasets (Zhenong Jin et al., 2018, 2019, 2017; Lobell et al., 2013, 2014, 2015).

We simulate a maize production system in Ames, Iowa, using the weather data from 2013. In the APSIM maize module, we identify six production factors that are suitable for our purpose:

- x_1 : sowing density (seeds/m^2)
- x_2 : sowing depth (mm)
- x_3 : row spacing (m)
- x_4 : N fertilizer amount before sowing (kg/ha)
- x_5 : N fertilizer amount at sowing (kg/ha)
- x_6 : N fertilizer amount for top dressing (kg/ha)

Thus, this oracle returns a yield (y) based on the levels of six factors $x = (x_1, \ldots, x_6)$

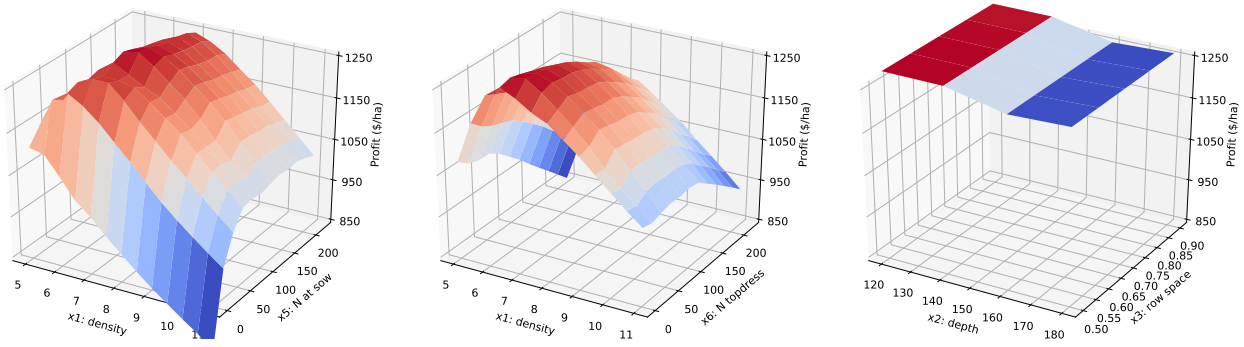
$$y = f(x) + \varepsilon.$$

Profit is calculated in a similar way:

$$\pi = py - c \cdot x.$$

The output price (p) is obtained from Duffy (2013), and the input costs (c) are obtained from Johanns (2019). We assume no cost for sowing depth and row spacing, which implies the cost vector $c = (c_1, 0, 0, c_4, c_5, c_6)$. It turns out that the maximum profit is \$1,288/ha, achieved by the management choices $x = (7, 120, 0.5, 0, 100, 20)$. Due to the six-dimensional space, it is impossible to directly plot the profit surface. We instead plot six pairs of interest, (x_1, x_5), (x_1, x_6), (x_2, x_3), (x_4, x_5), (x_4, x_6), and (x_5, x_6), while fixing the other four factors at the optimal levels. As seen below, the profit function takes a complex shape as a mixture of flat, convex and concave portions.

To investigate effects of noise, we implement three different noise levels, $\sigma \in \{500, 1000, 1500\}$, which corresponds, respectively, to about 5%, 10%, and 15% of the maximum possible yield. Notice that this is a much smaller variation than seen in scenario A. The reason is that, due to the higher dimensionality, the effect of noise is stronger and 15% of noise is large enough to make learning very difficult.



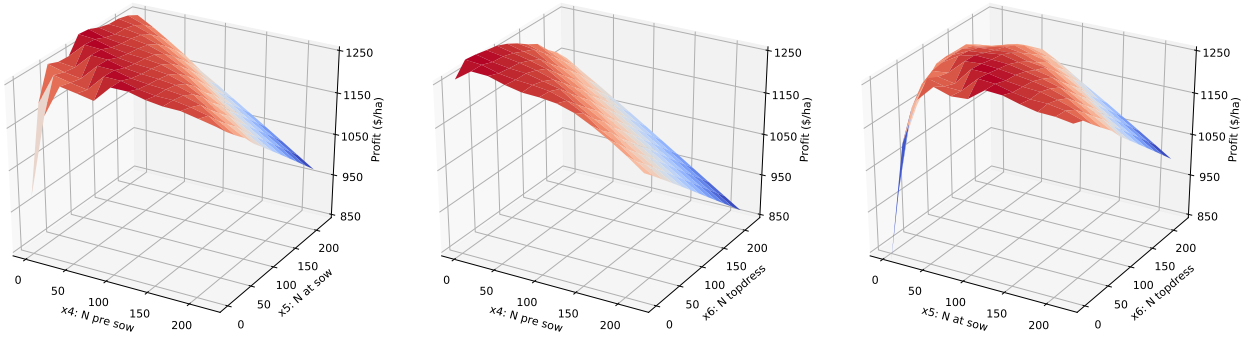


Figure 3: Selected profit surfaces of the high-dimensional oracle around the maximum

2.2 Experimental designs

Every experimental design either explicitly or implicitly assumes an accompanying statistical model that researchers estimate using the data collected by implementing the design (Montgomery, 2017). In agricultural experiments, it is common to use simple concave functions including quadratic (Bachmaier and Gandorfer, 2009; Meyer-Aurich et al., 2010; Whelan et al., 2012), negative exponential (Edwards and Purcell, 2005; Gaspar et al., 2015), and piecewise linear (Ouedraogo and Brorsen, 2018; Park et al., 2018). In the BO literature, (adaptive) designs are called "acquisition functions," and typically accompanied by a Gaussian process statistical model (Rasmussen and Williams, 2006).

Each experiment proceeds on a batch-sampling basis over years. Let M denote a batch size. Then, every year a researcher employs M plots and collect M samples. For scenario A, an optimal seeding rate for each pH is searched for over $x_1 \in [30, 150]$. For scenario B, an optimal level of the six factors is searched for over $x_1 \in [5.0, 11.0]$, $x_2 \in [120, 180]$, $x_3 \in [0.5, 0.9]$, $x_4 \in [0, 220]$, $x_5 \in [0, 220]$, and $x_6 \in [0, 220]$.

2.2.1 Factorial design

For scenario A, a factorial design consists of $\{45, 75, 105, 135\}$ in seeding rate and $\{6.0, 6.5, 7.0, 7.5\}$ in pH for each of three seed treatments, making up $M = 4 \times 4 \times 3 = 48$. For scenario B, we choose two levels for each factor, making up $M = 2^6 = 64$. Specifically, $x_1 \in \{7.0, 9.0\}$, $x_2 \in \{140, 160\}$, $x_3 \in \{0.6, 0.8\}$, $x_4 \in \{60, 140\}$, $x_5 \in \{60, 140\}$, and $x_6 \in \{60, 140\}$. Intuitively, two levels for each factor may be too coarse to detect good optimal values over the continuous search space. However, with three levels, it would be $3^6 = 729$ and too large to be practical. This already indicates a fundamental problem of factorial design in a high-dimensional space, a problem dubbed the "curse of dimensionality" (Bellman, 2015). As is standard in the literature, the design in each scenario is fixed over the years. Finally, for an accompanying statistical model, we assume a quadratic model, which is in the case of $x = (x_1, x_2)$:

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2,$$

and in the case of $x = (x_1, x_2, x_3, x_4, x_5, x_6)$:

$$f(x) = \sum_{k_1 + k_2 + \cdots + k_6 \leq 2} \beta_k \prod_{j=1}^{6} x_j^{k_j},$$

where $k_j \in \{0, 1, 2\}$ for all $j \in \{1, 2, \ldots, 6\}$, and $k = (k_1, k_2, k_3, k_4, k_5, k_6)$ denotes one of 28 indices for the coefficient β jointly defined by $k_1, k_2, \ldots, k_6$.

2.2.2 Bayesian optimization design

In contrast, the BO design adaptively chooses any combinations recommended by the algorithm. Each time after obtaining M samples, it updates the Gaussian process statistical models and reflects these updates on the next year's sampling recommendation. Hence, the adaptive experimental design. The algorithm is named "batch expected improvement," which is an extension of the standard expected improvement (EI) algorithm (Jones et al., 1998; Moćkus et al., 1978). The details of the algorithms for both scenario A and B are found in Appendix. The basic mathematics for and the specifications of Gaussian process are also found in Appendix.

While the standard EI algorithm is a strictly sequential algorithm designed to recommend only a single combination for the next sampling occasion, in agricultural experiments, we test multiple combinations at each occasion (typically, a year). Hence, we need to extend it so as to recommend multiple combinations at a time. To this end, we run the standard EI algorithm M times within a year by assuming $M-1$ hypothetical samples, as if they were sequentially observed. We adopt the idea from Ginsbourger et al. (2010) and assume each hypothetical sample to be equal to the lowest value observed so far (see line 7 in Algorithm 1). In other words, we assume within a year the algorithm makes a bad recommendation (remember the objective of the profit maximization—searching for a combination that returns the highest value). This heuristic allows the EI algorithm to do more exploration than exploitation; otherwise, the EI algorithm would tend to over-exploit suboptimal choices without sufficient exploration (Berk et al., 2018).

2.3 Performance metric

To compare alternatives, we need a performance metric. For this purpose, we use a notion of "regret" defined as a percentage of missed profit—a relative difference between the maximum possible profit and a profit computed using a statistical model of the environment estimated with the samples collected by following a design.

$$\text{Regret} = \frac{\text{True maximum profit} - \text{Estimated maximum profit}}{\text{True maximum profit}} \times 100.$$

For scenario A, since a regret can be calculated for each of 16 pHs, we use their average as the regret of a design.

2.4 Overall procedure

Our simulation study places the researcher in two distinct scenarios, each of which has three subscenarios characterized by different noise levels. In each subscenario, experiments last for T years. In each year, a design does the following:

1. Picks M testing combinations
2. Queries each combination to the oracle
3. Receives corresponding M yield observations and computes their profits
4. Adds them to the sample pool
5. Updates the statistical model for the underlying profit function

Note that step 5 is meaningful only for the BO design because the factorial design is static and independent of the statistical model. After T years, a regret is calculated as a performance metric of the design. To reduce the effect of the randomness involved in both yield noise and the algorithms themselves, we run multiple Monte Carlo simulations for each subscenario and report their sample mean and standard deviation as a final result.

3 Results

3.1 Scenario A

The study for scenario A was conducted with 1,000 Monte Carlo simulations. Besides the fact that the factorial design had $M = 48$, the BO version also used the same number for the batch size. The effects of noise are illustrated separately for the factorial and BO designs in the figures below. As expected, the greater the noise level, the more difficult it is to learn profit functions and optimize them. With a greater noise level, clearly, the regret in any time horizon T is higher. In addition, at noise level 50, the BO version hardly learns anything even after 10 years, but the factorial does no better, nonetheless.

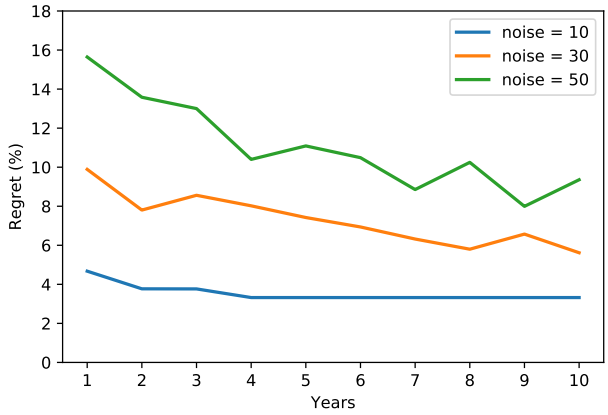


Figure 4: Effect of noise on learning (Factorial)

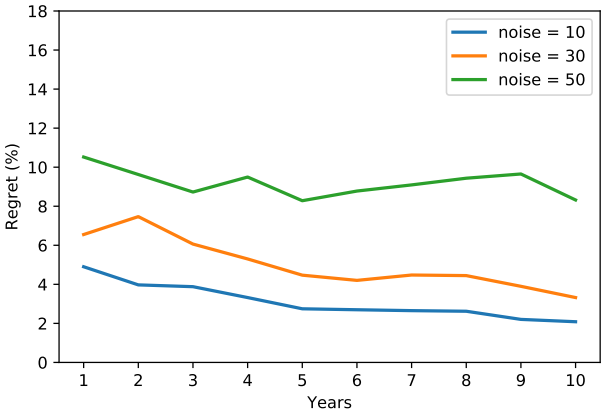


Figure 5: Effect of noise on learning (BO)

To show more detail, the following three figures plot the results separately for each noise level. Solid lines trace the mean regret at each time horizon (T), and shaded areas represent 1 standard deviation around the means. Overall, on average, the BO design demonstrates higher performance for most time horizons in all the subscenarios. However, the differences may not be seen as economically significant, and moreover, their distributions are largely overlapped, indicating even weaker appeal of the higher mean performances.

3.1.1 10% noise

Years	1	2	3	4	5	6	7	8	9	10
Factorial	6.8	5.4	4.8	4.1	3.8	4.0	3.6	3.5	3.5	3.4
(std)	(5.0)	(4.6)	(3.9)	(2.4)	(1.4)	(2.1)	(1.2)	(1.2)	(1.2)	(1.3)
BO	7.4	4.9	4.2	3.7	3.1	3.1	2.7	2.6	2.2	2.2
(std)	(6.0)	(2.3)	(1.6)	(1.4)	(1.4)	(1.7)	(1.5)	(1.5)	(1.0)	(1.2)
Difference	-0.6	0.5	0.6	0.4	0.7	0.9	0.9	0.9	1.3	1.2

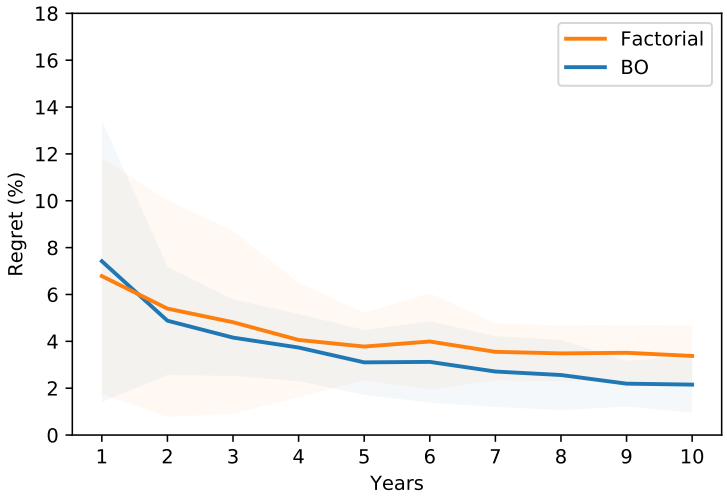


Figure 6: Regret under noise = 10%, M = 48

3.1.2 30% noise

Years	1	2	3	4	5	6	7	8	9	10
Factorial	11.9	9.6	9.7	8.9	8.2	8.1	7.6	6.8	7.2	6.7
(std)	(7.3)	(5.7)	(6.4)	(5.1)	(5.7)	(5.1)	(5.0)	(4.4)	(4.7)	(4.0)
BO	9.9	9.5	8.5	8.7	6.7	6.7	6.3	5.5	5.5	4.2
(std)	(7.0)	(6.0)	(6.8)	(7.7)	(6.4)	(6.4)	(6.4)	(3.5)	(4.7)	(2.7)
Difference	2.0	0.1	1.2	0.2	1.5	1.4	1.3	1.3	1.7	2.5

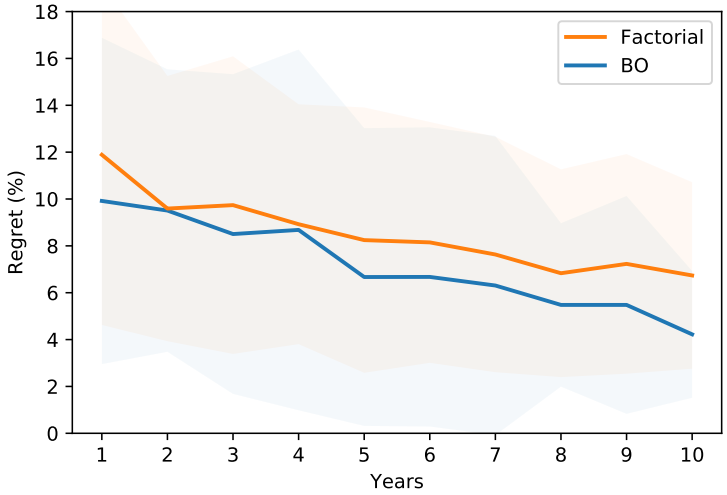


Figure 7: Regret under noise = 30%, M = 48

3.1.3 50% noise

Years	1	2	3	4	5	6	7	8	9	10
Factorial	15.7	14.5	13.0	11.9	10.8	11.3	10.5	10.2	9.3	10.6
(std)	(7.9)	(7.7)	(7.5)	(6.9)	(6.5)	(6.4)	(6.3)	(5.7)	(5.6)	(6.1)
BO	10.7	10.6	9.3	8.9	9.1	9.7	9.3	10.9	9.7	8.3
(std)	(3.6)	(5.4)	(5.0)	(2.9)	(3.7)	(4.2)	(2.0)	(4.5)	(2.5)	(1.7)
Difference	5.0	3.9	3.7	3.0	1.7	1.6	1.2	-0.7	-0.4	2.3

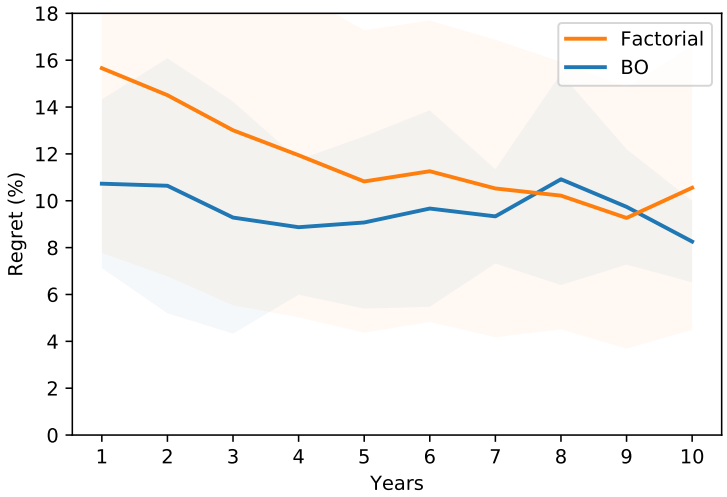


Figure 8: Regret under noise = 50%, M = 48

3.2 Scenario B

The study for scenario B was conducted with 1,000 Monte Carlo simulations. As opposed to $M = 64$ of the factorial design, the BO design here uses only $M = 30$ for the batch size. The reason for the smaller batch size is because it turned out $M = 30$ is sufficient to produce convincing results, and the computational costs grow increasingly with batch size. First, the effects of noise are illustrated separately for the factorial and BO designs in the figures below. Similar to scenario A, there is a pattern that the greater the noise level, the more difficult it is to learn profit functions and optimize them.

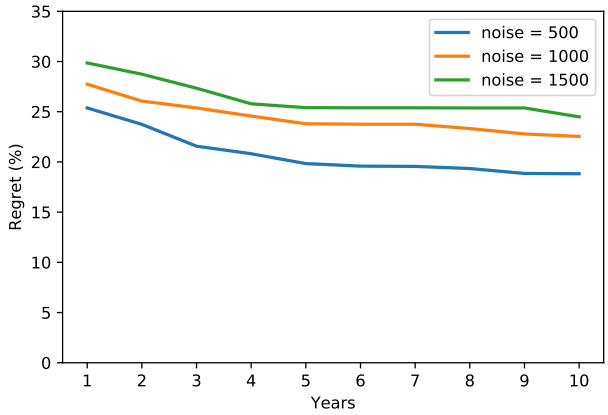


Figure 9: Effect of noise on learning (Factorial)

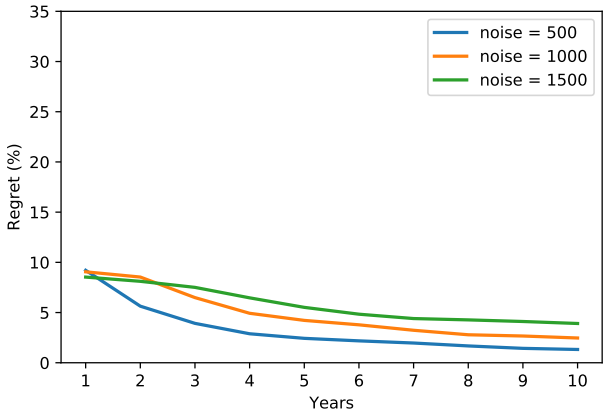


Figure 10: Effect of noise on learning (BO)

To show this in more detail, the following three figures plot the results separately for each noise level. Solid lines trace the mean regret at each time horizon (T), and shaded areas represent the corresponding 1 standard deviations around the means. It is evident that, on average, the BO design outperformed the factorial design by a wide margin, at least 12.9 percentage points and larger with longer time horizons. Besides, as indicated by the two largely separated shaded areas in each figure, the mean regrets of the BO design is also statistically distant from that of the factorial. Moreover, it is probably a more practical question to ask how often the BO design "beat" the factorial design. So, at the bottom of each table, we also show the percentage of simulation

runs when the BO had lower regrets than the factorial at each time horizon. To be fair, we gave both designs the same random realization of noise at each simulation run (and of course different randomness across 1,000 runs). As implied by the large mean differences, the BO beat the factorial at least 83.3% of times and generally more as the horizon became longer.

3.2.1 5% noise

Years	1	2	3	4	5	6	7	8	9	10
Factorial	25.5	23.6	22.4	21.7	21.4	20.9	20.6	20.4	20.3	19.9
(std)	(9.3)	(7.7)	(6.2)	(5.7)	(5.1)	(4.8)	(4.4)	(4.1)	(4.0)	(3.9)
BO	12.6	9.6	6.9	4.8	3.7	2.9	2.5	2.1	1.8	1.7
(std)	(11.0)	(9.2)	(7.8)	(6.3)	(4.6)	(3.8)	(3.1)	(2.8)	(2.1)	(2.0)
Difference	12.9	14.0	15.5	16.9	17.7	18.0	18.1	18.3	18.5	18.2
%wins	83.3	87.4	92.5	95.8	97.9	98.2	99.0	99.3	99.6	99.8

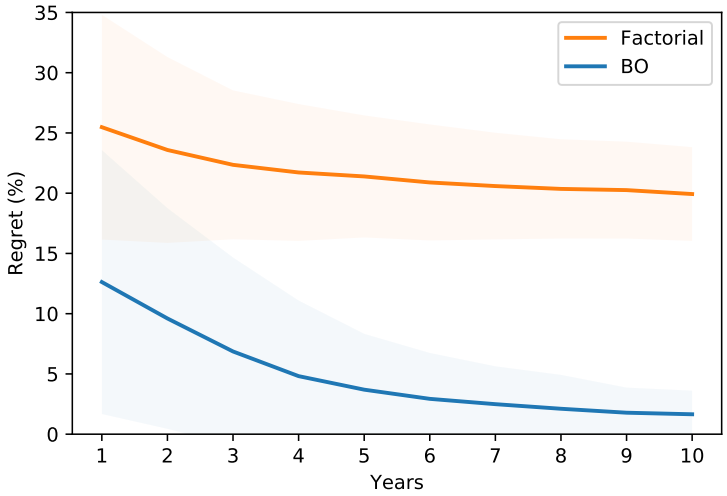


Figure 11: Regret under noise = 5%, M = 30, 64

3.2.2 10% noise

Years	1	2	3	4	5	6	7	8	9	10
Factorial	29.4	27.8	26.0	25.3	24.9	24.1	24.1	23.7	23.6	23.3
(std)	(12.2)	(11.0)	(9.7)	(9.1)	(8.4)	(8.1)	(8.0)	(7.2)	(6.8)	(6.8)
BO	11.3	11.9	10.3	8.2	6.4	5.4	4.7	3.8	3.5	3.1
(std)	(8.7)	(10.4)	(9.8)	(8.5)	(6.6)	(5.8)	(5.8)	(4.5)	(4.0)	(3.2)
Difference	18.1	15.9	15.7	17.1	18.5	18.7	19.4	19.9	20.1	20.2
%wins	89.0	88.4	89.4	91.3	95.2	96.9	96.3	98.5	98.8	99.0

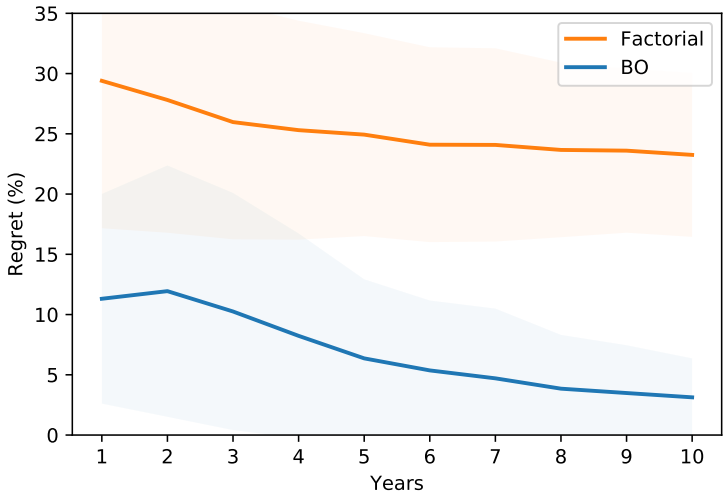


Figure 12: Regret under noise = 10%, M = 30, 64

3.2.3 15% noise

Years	1	2	3	4	5	6	7	8	9	10
Factorial	31.7	30.0	28.8	27.7	26.8	26.5	26.4	25.9	26.0	25.4
(std)	(13.1)	(12.5)	(11.9)	(10.9)	(10.4)	(10.2)	(10.1)	(9.8)	(9.2)	(8.9)
BO	10.6	10.5	10.5	9.6	8.6	7.8	6.6	5.8	5.5	5.2
(std)	(8.3)	(8.3)	(8.9)	(8.6)	(8.3)	(7.8)	(6.3)	(5.5)	(5.7)	(5.3)
Difference	21.1	19.5	18.3	18.1	18.2	18.7	19.8	20.1	20.5	20.2
%wins	92.8	92.2	90.2	90.0	92.1	93.2	95.6	95.9	96.6	96.8

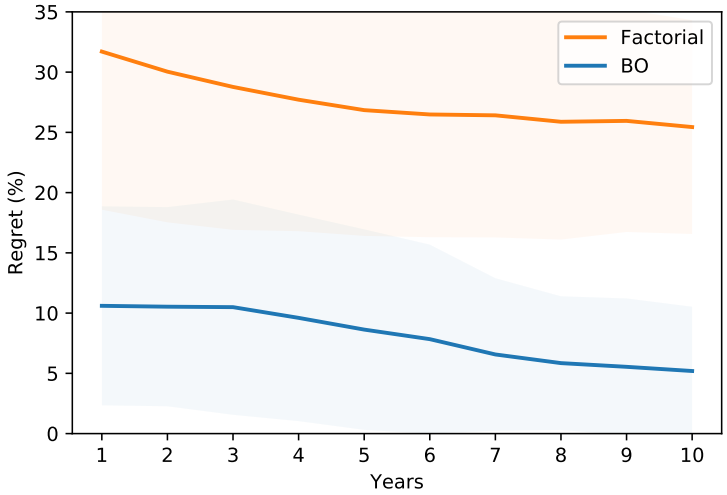


Figure 13: Regret under noise = 15%, M = 30, 64

4 Discussion

We created two distinct simulation environments and tested optimization performances of factorial designs and adaptive designs based on BO. Two environments were constructed to highlight conditions in which the BO design could offer the most benefits—high-dimensional environments involving many management and environmental variables. The common results in both designs

were that the noisier the production systems, the more difficult it is to learn and optimize. These were intuitive results and reiterated the difficulty involved in optimizing agricultural systems, where a considerable amount of noise is typically present. In terms of relative performance, the BO designs outperformed the factorial designs on average, indicated by the blue curves located below the orange curves for most time horizons in Figures 6-8 and 11-13. Recall that one of the reasons for replication is to reduce the variability of estimator or standard error (Mead et al., 1993). However, as the similar sizes of the shaded areas in each figure indicated, the variability of the factorial design in both scenarios is by no means smaller than that of the BO design at large and even larger in many cases (e.g. Figures 8, 12, 13). So, the factorial design in this case actually sacrificed the quality (i.e., average regret) of the estimator without gaining anything. Besides the common characteristics of the results between the two scenarios, there are notable differences between scenario A and scenario B.

In scenario A, while the overall performance of the BO design is higher in most time horizons, the performance differences are small, ranging from -0.7 to 5.0 percentage points across all the subscenarios (Figures 6-8). As a result, these differences may not be seen as economically significant. Moreover, the distributions of the two designs at each time horizon are largely overlapped, indicating even weaker statistical appeal of the BO design's higher average performances. This was attributed to the low dimensionality of the scenario. Specifically, while there were technically three dimensions in scenario A—seed treatment, seeding rate, and pH—the seed treatment dimension was discrete containing only three elements $\{UTC, F, FI\}$. Therefore, it was still "easy" to cover the dimension by the factorial design, reducing the relative advantage of the BO version.

In contrast, in scenario B, the BO design overwhelmingly outperformed the factorial design by a wide margin, ranging from 12.9 to 21.1 percentage points. With the maximum possible profit of \$1,288/ha, 20 points of difference in regret are equivalent to \$258/ha, which is of considerable economic benefits. To examine this in more detail, consider a five-year experimental horizon. The differences in regret across three subscenarios are 17.7 (5% noise), 18.5 (10% noise), and 18.2 (15% noise), corresponding to \$228/ha, \$238/ha, and \$234/ha respectively. Remember that the BO design used only 30 samples each year, less than half of the 64 samples used by the factorial design. Thus, the addition of experimental cost saving made the economic benefits of the BO version even more significant. Furthermore, as clearly seen under the 5% noise subscenario (Figure 11), the performance difference in some cases widened as the time horizon became longer, indicating that the BO design learned more than the factorial.

Despite the promising ability and performance of BO experimental designs, there are several clarifications and limitations to note for real-world implementations and future research.

- We do not claim that the constructed simulation environments are particularly accurate or realistic in every detail. Instead, what we tried to create are independent and impartial devices that allow us to systematically test and compare alternative experimental designs.
- We implicitly assumed that, within a subscenario characterized by an oracle and a noise level, the noise distribution is stationary over the time horizon. For example, given the strong influence of weather on agricultural yield, it may be necessary to reflect weather patterns over multiple years.
- We assumed homoskedasticity—a common noise level across all values of x within a subscenario. With appropriate data, we can test that. Then, if violated, we can estimate heteroskedastic errors and reflect them in simulations.
- Our BO technique is limited to one-shot optimization, in other words, that researchers choose all the factor levels at the beginning of the year and wait to see results until the end of the

year. In agriculture, many management choices are made within the year. To handle such more realistic situations, we need to construct dynamic models in which feedback information and learning take place both within and across years.

- When variables were theoretically continuous (e.g., seeding rate), we assumed that it was possible to choose any arbitrary level and run the BO algorithm accordingly. In reality, however, choices of continuous variables are constrained for various practical reasons. For example, for a technical reason, we cannot change a fertilizer amount precisely by a fraction of a kilogram. Even though technically possible, it is economically infeasible to change row spacing every time an algorithm recommends a different value. Therefore, for real-world implementations, we need to modify those continuous factors into discrete ones.

5 Conclusions

The goal of our work is to introduce BO techniques to agricultural audiences so that they can apply them to their own problems. To this end, we have described an experimental design based on BO and demonstrated its potential by conducting simulation studies. In general, the benefits of BO will likely be significant when the dimensionality of management and environmental variables is high, and the nature of production factors is continuous. As we try to optimize more variables, the number of combinations to examine increases exponentially. Consequently, fixed factorial designs or any brute-force search strategy is less effective, relatively increasing the advantage of sophisticated designs such as those based on BO. In addition, the more variables involved in defining the yield response function, the higher the chances of interactions between them (D. S. Bullock and D. G. Bullock, 2000; Ruffo et al., 2006), leading to a complex shape of the profit function, which is difficult to optimize by using the traditional experimental designs. Thus, it is safer and likely beneficial for researchers to use more flexible statistical models (e.g., Gaussian process used in BO) and guard against potential model mispecification and, thereby, substantial loss. In summary, experimental designs based on BO address these issues and potentially create significant economic benefits.

References

Bachmaier, Martin and Markus Gandorfer (2009). "A conceptual framework for judging the precision agriculture hypothesis with regard to site-specific nitrogen application". In: *Precision agriculture* 10.2, p. 95.

Bellman, Richard E (2015). *Adaptive control processes: a guided tour*. Vol. 2045. Princeton university press.

Berk, Julian, Vu Nguyen, Sunil Gupta, Santu Rana, and Svetha Venkatesh (2018). "Exploration Enhanced Expected Improvement for Bayesian Optimization". In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 621–637.

Brochu, Eric, Vlad M Cora, and Nando De Freitas (2010). "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning". In: *arXiv preprint arXiv:1012.2599*.

Bullock, David S and Donald G Bullock (2000). "From agronomic research to farm management guidelines: A primer on the economics of information and precision technology". In: *Precision Agriculture* 2.1, pp. 71–101.

Buresh, Roland J., Rowena L. Castillo, Judith Carla Dela Torre, Eufrocino V. Laureles, Marianne I. Samson, Philip Joshua Sinohin, and Marlon Guerra (2019). "Site-specific nutrient management for rice in the Philippines: Calculation of field-specific fertilizer requirements by Rice Crop Manager". In: *Field Crops Research* 239, pp. 56–70.

Camberato, Jim (2014). *Low soil pH and limestone recommendations for mineral soils: The confusion of 'Buffer pH'*. URL: https://ag.purdue.edu/agry/extension/SiteAssets/soilfertilityassets/Buffer-pH.pdf.

Cela, Sebastián, Francisca Santiveri, and Jaime Lloveras (2011). "Optimum nitrogen fertilization rates for second-year corn succeeding alfalfa under irrigation". In: *Field Crops Research* 123.2, pp. 109–116.

Dai, Xinglong, Xiaohu Zhou, Dianyong Jia, Lili Xiao, Haibo Kong, and Mingrong He (2013). "Managing the seeding rate to improve nitrogen-use efficiency of winter wheat". In: *Field Crops Research* 154, pp. 100–109.

Duffy, Michael (2013). *Estimated Costs of Crop Production in Iowa - 2013*. Estimated Costs of Crop Production in Iowa | Ag Decision Maker. URL: http://econ2.econ.iastate.edu/faculty/duffy/documents/EstimatedCostsofCropProduction2013.pdf.

Edwards, Jeffrey T and Larry C Purcell (2005). "Soybean yield and biomass responses to increasing plant population among diverse maturity groups". In: *Crop Science* 45.5, pp. 1770–1777.

Gaspar, Adam P, Paul D Mitchell, and Shawn P Conley (2015). "Economic risk and profitability of soybean fungicide and insecticide seed treatments at reduced seeding rates". In: *Crop Science* 55.2, pp. 924–933.

Ginsbourger, David, Rodolphe Le Riche, and Laurent Carraro (2010). "Kriging is well-suited to parallelize optimization". In: *Computational intelligence in expensive optimization problems*. Springer, pp. 131–162.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2016). *The elements of statistical learning*. 2nd ed. New York: Springer.

Holzworth, Dean P, Neil I Huth, Peter G deVoil, Eric J Zurcher, Neville I Herrmann, Greg McLean, Karine Chenu, Erik J van Oosterom, Val Snow, and Chris Murphy (2014). "APSIM–evolution towards a new generation of agricultural systems simulation". In: *Environmental Modelling & Software* 62, pp. 327–350.

Huang, Jianliang, Fan He, Kehui Cui, Roland J Buresh, Bo Xu, Weihua Gong, and Shaobing Peng (2008). "Determination of optimal nitrogen rate for rice varieties using a chlorophyll meter". In: *Field Crops Research* 105.1, pp. 70–80.

Jin, Zewen, Can Chen, Xiaomin Chen, Fei Jiang, Isaac Hopkins, Xiaoling Zhang, Zhaoqiang Han, Grace Billy, and Jhony Benavides (2019). "Soil acidity, available phosphorus content, and optimal biochar and nitrogen fertilizer application rates: A five-year field trial in upland red soil, China". In: *Field Crops Research* 232, pp. 77–87.

Jin, Zhenong, Elizabeth A Ainsworth, Andrew D B Leakey, and David B Lobell (2018). "Increasing drought and diminishing benefits of elevated carbon dioxide for soybean yields across the US Midwest". In: *Global Change Biology* 24.2, e522–e533.

Jin, Zhenong, Sotirios V Archontoulis, and David B Lobell (2019). "How much will precision nitrogen management pay off? An evaluation based on simulating thousands of corn fields over the US Corn-Belt". In: *Field Crops Research* 240, pp. 12–22.

Jin, Zhenong, George Azzari, and David B Lobell (2017). "Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches". In: *Agricultural and Forest Meteorology* 247, pp. 207–220.

Johanns, Ann (2019). *Iowa cash corn and soybean prices*. Cash Corn and Soybean Prices | Ag Decision Maker. URL: https://www.extension.iastate.edu/agdm/crops/html/a2-11.html.

Jones, Donald R, Matthias Schonlau, and William J Welch (1998). "Efficient global optimization of expensive black-box functions". In: *Journal of Global optimization* 13.4, pp. 455–492.

Khan, Aziz, Leishan Wang, Saif Ali, Shahbaz Atta Tung, Abdul Hafeez, and Guozheng Yang (2017). "Optimal planting density and sowing date can improve cotton yield by maintaining reproductive organ biomass and enhancing potassium uptake". In: *Field Crops Research* 214, pp. 164–174.

Li, Maona, Yunling Wang, Ardeshir Adeli, and Haijun Yan (2018). "Effects of application methods and urea rates on ammonia volatilization, yields and fine root biomass of alfalfa". In: *Field Crops Research* 218, pp. 115–125.

Lobell, David B, Graeme L Hammer, Greg McLean, Carlos Messina, Michael J Roberts, and Wolfram Schlenker (2013). "The critical role of extreme heat for maize production in the United States". In: *Nature Climate Change* 3.5, p. 497.

Lobell, David B, Michael J Roberts, Wolfram Schlenker, Noah Braun, Bertis B Little, Roderick M Rejesus, and Graeme L Hammer (2014). "Greater Sensitivity to Drought Accompanies Maize Yield Increase in the U.S. Midwest". In: *Science* 344.6183, p. 516.

Lobell, David B, David Thau, Christopher Seifert, Eric Engle, and Bertis Little (2015). "A scalable satellite-based crop yield mapper". In: *Remote Sensing of Environment* 164, pp. 324–333.

Lynch, Joseph P, Elizabeth Glynn, Steven Kildea, and John Spink (2017). "Yield and optimum fungicide dose rates for winter wheat (Triticum aestivum L.) varieties with contrasting ratings for resistance to septoria tritici blotch". In: *Field Crops Research* 204, pp. 89–100.

Mead, R, RN Curnow, and AM Hasted (1993). "Statistical Methods in Agriculture and Experimental Biology". In:

Meyer-Aurich, Andreas, Alfons Weersink, Markus Gandorfer, and Peter Wagner (2010). "Optimal site-specific fertilization and harvesting strategies with respect to crop yield and quality response to nitrogen". In: *Agricultural Systems* 103.7, pp. 478–485.

Moćkus, J, V Tiesis, and A Źilinskas (1978). "The Application of Bayesian Methods for Seeking the Extremum. Vol. 2". In: Dixon, L and G Szego. *Toward Global Optimization*. Vol. 2. Amsterdam, The Netherlands: Elsevier.

Montgomery, Douglas C (2017). *Design and analysis of experiments*. John Wiley & Sons.

Orlowski, John M, Bryson J Haverkamp, Randall G Laurenz, David Marburger, Eric W Wilson, Shaun N Casteel, Shawn P Conley, Seth L Naeve, Emerson D Nafziger, and Kraig L Roozeboom (2016). "High-input management systems effect on soybean seed yield, yield components, and economic break-even probabilities". In: *Crop Science* 56.4, pp. 1988–2004.

Ouedraogo, Frederic and B Wade Brorsen (2018). "Hierarchical Bayesian Estimation of a Stochastic Plateau Response Function: Determining Optimal Levels of Nitrogen Fertilization". In: *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie* 66.1, pp. 87–102.

Park, Eunchun, Wade Brorsen, and Xiaofei Li (2018). *How to Use Yield Monitor Data to Determine Nitrogen Recommendations: Bayesian Kriging for Location Specific Parameter Estimates*. Agricultural and Applied Economics Association.

Petersen, Roger G (1994). *Agricultural field experiments: design and analysis*. CRC Press.

Rasmussen, Carl Edward and Christopher K Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ren, Tao, Bo Liu, Jianwei Lu, Zhonghua Deng, Xiaokun Li, and Rihuan Cong (2017). "Optimal plant density and N fertilization to achieve higher seed yield and lower N surplus for winter oilseed rape (Brassica napus L.)" In: *Field Crops Research* 204, pp. 199–207.

Rens, Libby R, Lincoln Zotarelli, Diane L Rowland, and Kelly T Morgan (2018). "Optimizing nitrogen fertilizer rates and time of application for potatoes under seepage irrigation". In: *Field Crops Research* 215, pp. 49–58.

Ruffo, Matías L, Germán A Bollero, David S Bullock, and Donald G Bullock (2006). "Site-specific production functions for variable rate corn nitrogen fertilization". In: *Precision Agriculture* 7.5, pp. 327–342.

Shahriari, Bobak, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas (2016). "Taking the human out of the loop: A review of bayesian optimization". In: *Proceedings of the IEEE* 104.1, pp. 148–175.

Storer, K E, P M Berry, D R Kindred, and R Sylvester-Bradley (2018). "Identifying oilseed rape varieties with high yield and low nitrogen fertiliser requirement". In: *Field Crops Research* 225, pp. 104–116.

Wang, Hongyuan, Yitao Zhang, Anqiang Chen, Hongbin Liu, Limei Zhai, Baokun Lei, and Tianzhi Ren (2017). "An optimal regional nitrogen application threshold for wheat in the North China Plain considering yield and environmental effects". In: *Field Crops Research* 207, pp. 52–61.

Wang, Weini, Jianwei Lu, Tao Ren, Xiaokun Li, Wei Su, and Mingxing Lu (2012). "Evaluating regional mean optimal nitrogen rates in combination with indigenous nitrogen supply for rice production". In: *Field Crops Research* 137, pp. 37–48.

Whelan, B M, J A Taylor, and A B McBratney (2012). "A small strip approach to empirically determining management class yield response functions and calculating the potential financial net wastage associated with whole-field uniform-rate fertiliser application". In: *Field Crops Research* 139, pp. 47–56.

Appendices

A. Algorithms for the BO designs

In terms of notation, T is the total number of years used for experimentation, $\mathcal{S}_\tau$ is a set of samples for seed treatment τ, GP_τ is a Gaussian process regression model for the profit response function of seed treatment τ, $\min_y\{\mathcal{S}_{\tau^*}\}$ implies the minimum observed yield for τ^*, and α_τ is an acquisition function defined based on GP_τ. A pair of braces { } indicates a collection or set that contains elements, for example, $\{\alpha_\tau\}$ is a set of three acquisition functions: α_{UTC}, α_F, and α_{FI}. Notice that, in the first year, $M/3$ number of samples are randomly collected for each seed treatment and stored in $\mathcal{S}_\tau$.

Algorithm 1 Batch expected improvement for scenario A

require: T, M, $\{\mathcal{S}_\tau\}$, $\{GP_\tau\}$, $\{\alpha_\tau\}$
for $t \in \{2, 3, \ldots, T\}$ **do**
  $\mathcal{I}_\tau \leftarrow \{\ \}$ for all $\tau \in \{UTC, F, FI\}$
  $\widehat{GP_\tau} \leftarrow GP_\tau$ for all $\tau \in \{UTC, F, FI\}$
  for $m \in \{1, 2, \ldots, M\}$ **do**
    $(\tau^*, x^*) \leftarrow \text{argmax}_{\tau,x}\, \alpha_\tau(x)$
    $\mathcal{I}_{\tau^*} \leftarrow \mathcal{I}_{\tau^*} \cup \{(x^*, \underline{y})\}$ where $\underline{y} = \min_y\{\mathcal{S}_{\tau^*}\}$
    Update $\widehat{GP}_{\tau^*}$ with $\mathcal{S}_{\tau^*} \cup \mathcal{I}_{\tau^*}$
  for $\tau \in \{UTC, F, FI\}$ **do**
    for $(x, y) \in \mathcal{I}_\tau$ **do**
      $y \leftarrow Oracle(\tau, x)$
      $\mathcal{S}_\tau \leftarrow \mathcal{S}_\tau \cup \{(x, y)\}$
    Update GP_τ with $\mathcal{S}_\tau$
return $M \times T$ number of samples

Notice that, despite the larger factor space in scenario B, it is algorithmically simpler because of the lack of discrete choice of seed treatment that require separate GPs. Note also that the difference between the two algorithms is not because of ad hoc adaptation to the specific scenarios; instead, it is due simply to the structural difference of the two scenarios. Hence, our BO design is consistent and general.

Algorithm 2 Batch expected improvement for scenario B

require: T, M, $\mathcal{S}$, GP, α
for $t \in \{2, 3, \ldots, T\}$ **do**
  $\mathcal{I} \leftarrow \{\ \}$
  $\widehat{GP} \leftarrow GP$
  for $m \in \{1, 2, \ldots, M\}$ **do**
    $x^* \leftarrow \text{argmax}_x\, \alpha(x)$
    $\mathcal{I} \leftarrow \mathcal{I} \cup \{(x^*, \underline{y})\}$ where $\underline{y} = \min_y\{\mathcal{S}\}$
    Update $\widehat{GP}$ with $\mathcal{S} \cup \mathcal{I}$
  for $(x, y) \in \mathcal{I}$ **do**
    $y \leftarrow Oracle(x)$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(x, y)\}$
  Update GP with $\mathcal{S}$
return $M \times T$ number of samples

B. Gaussian process

Gaussian process is a Bayesian nonparametric model, and its behavior is largely dependent on a kernel (Rasmussen and Williams, 2006). A kernel is a function that returns a similarity measure $k(x, x')$ between two points x and x'. We use the Matérn kernel—a popular class of isotropic stationary kernels.

$$k_\nu(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^\nu B_\nu \left(\sqrt{2\nu} \frac{d}{\rho} \right),$$

where Γ is the gamma function, B_ν is the modified Bessel function of the second kind, and d is a metric often induced by the Euclidean norm, i.e. $d = \|x - x'\|$. The Matérn kernel is characterized by two parameters ν and ρ, which control, respectively, the smoothness and the scaling of distance. As standard in applied work, we do not estimate but rather handpick ν and write as Matérn$_\nu$ or $k_\nu(x, x')$. To simplify the notation, let r denote the scaled distance, $r = d/\rho$. An important property of the Matérn kernel is that when $\nu = p + 1/2, p \in \mathbb{N}$, it can be written as a product of an exponential and a polynomial of order p:

$$k_{p+1/2}(x, x') = \sigma^2 \exp\left(-\sqrt{2p+1} r \right) \frac{p!}{(2p)!} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} (2\sqrt{2p+1} r)^{p-i}.$$

Common choices of ν are $1/2, 3/2, 5/2$ and ∞, with each of which the kernel reduces to, respectively,

$$
\begin{aligned}
k_{1/2}(x, x') &= \sigma^2 \exp(-r) \\
k_{3/2}(x, x') &= \sigma^2 \exp\left(-\sqrt{3}r\right)\left(1 + \sqrt{3}r\right) \\
k_{5/2}(x, x') &= \sigma^2 \exp\left(-\sqrt{5}r\right)\left(1 + \sqrt{5}r + \frac{5}{3}r^2\right) \\
k_{\infty}(x, x') &= \lim_{\nu \to \infty} k_{\nu}(x, x') = \sigma^2 \exp\left(-\frac{1}{2}r^2\right).
\end{aligned}
$$

We use $\nu = 3/2$ for the BO design. Matérn$_\infty$ is also known as squared exponential kernel or radial basis function. The following figure plots $k_\nu(x, x')$ with $\sigma^2 = \rho = 1$ for $\nu \in \{1/2, 3/2, \infty\}$.

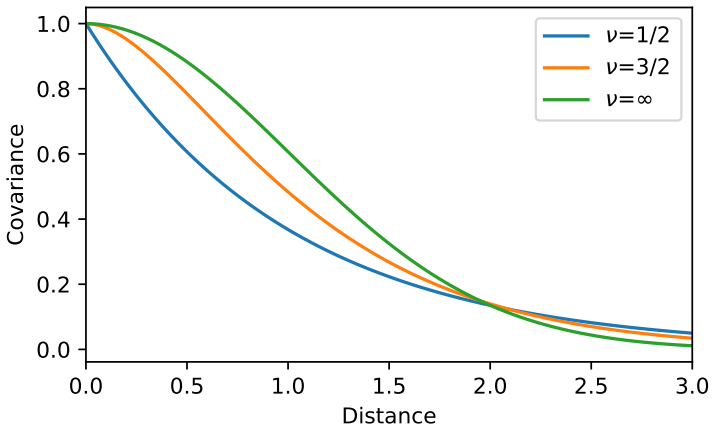


Figure 14: Matérn$_\nu$ kernels

C. Expected improvement acquisition function

In Bayesian optimization, an algorithm prescribes the next sampling point x based on how we value the mean and variance at x estimated by the accompanying GP. Specifically, the recommendation x_t for the next round t is determined by maximizing an acquisition function $\alpha(x|D_{t-1})$:

$$x_t = \operatorname*{argmax}_x \alpha(x|D_{t-1}),$$

where D_{t-1} is the data used to fit the GP at round $t-1$. The acquisition function is a reflection of the underlying utility of the next sample or our preference in selecting the next sampling point. It is heuristic and designed to trade off exploration of the search space and exploitation of the current promising areas. There are a number of acquisitions functions proposed in the literature. One of the popular acquisition functions is called expected improvement, which is constructed based on the following intuitive idea. Let y^* be the maximum value observed up until round $t-1$, i.e. $y^* = \max\{y_1, \ldots, y_{t-1}\}$. Then, we may define "improvement" at point x at round t to be

$$\max\{0, GP(x) - y^*\},$$

which is random as $GP(x)$ is a random function. Thus, the expected improvement acquisition function is defined to be:

$$\alpha_{EI}(x|D_{t-1}) = \mathbb{E}[\max\{0, GP(x) - y^*\}|D_{t-1}].$$

When using Gaussian process, at each point x in the domain, we have $GP(x) \sim \mathcal{N}(\mu(x), \sigma(x))$, which allows the expected improvement to have a closed form (Jones et al., 1998; Moćkus et al., 1978):

$$\alpha_{EI}(x|D_{t-1}) = \begin{cases} (\mu(x) - y^*)\Phi\left(\frac{\mu(x)-y^*}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{\mu(x)-y^*}{\sigma(x)}\right) & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) = 0 \end{cases},$$

where Φ is the standard normal cumulative distribution function and ϕ is the standard normal probability density function.