A/B Testing with Fat Tails*

Eduardo M. Azevedo† Alex Deng‡ José Luis Montiel Olea§
Justin Rao¶ E. Glen Weyl‖

First version: April 30, 2018
This version: August 9, 2019

Abstract

Large and statistically powerful A/B tests are increasingly popular to screen new business and policy ideas. We study how to use scarce experimental resources to screen multiple potential innovations by proposing a new framework for optimal experimentation, that we term the A/B testing problem. The main departure from the literature is that the model allows for fat tails. The key insight is that the optimal experimentation strategy depends on whether most gains accrue from typical innovations or from rare and unpredictable large successes that can be detected using tests with small samples. We show that, if the tails of the unobserved distribution of innovation quality are not too fat, the standard approach of using a few high-powered "big" experiments is optimal. However, when this distribution is very fat tailed, a "lean" experimentation strategy consisting of trying more ideas, each with possibly smaller sample sizes, is preferred. We measure the relevant tail parameter using experiments from Microsoft Bing's EXP platform and find extremely fat tails. Our theoretical results and empirical analysis suggest that even simple changes to business practices within Bing could increase innovation productivity.

*We are grateful to Sylvain Chassang, Navin Kartik, Bobby Kleinberg, Qingmin Liu, Konrad Menzel, David Pearce, Isabelle Perrigne, Andrea Prat, Debraj Ray, Azeem Shaikh, Quang Vuong, five anonymous referees, and to workshop participants at the UCLA, the University of Chicago, Columbia University, the Federal Reserve Bank of Dallas, Microsoft Research, Northwestern University, UPenn, and New York University for useful comments and feedback. We would also like to thank Michael Kurish, Surya Ierkomos, Ruby Steedle, and Amilcar Velez for excellent research assistance.

†Wharton: eazevedo@wharton.upenn.edu, http://www.eduardomazevedo.com.

‡Microsoft Corporation: shaojie.deng@microsoft.com, http://alexdeng.github.io/.

§Columbia University: montiel.olea@gmail.com, http://www.joseluismontielolea.com/.

¶HomeAway: justinmrao@outlook.com, http://www.justinmrao.com.

‖Microsoft Research, 641 Avenue of the Americas, New York, NY 10011 and Department of Economics, Princeton University: glenweyl@microsoft.com, http://www.glenweyl.com.

1 Introduction

Randomized experiments are increasingly important for innovation in many fields. In the high-tech sector, major companies run thousands of experiments (called A/B tests) each year, and use the results to screen most product innovations.[1] In policy and academic circles, randomized experiments are used to evaluate social programs and shape public policy.[2] A striking fact in many of these settings is that there are anecdotes of "black swans": outlier ideas with large unexpected effects. These outliers are commonly thought to be important for innovation.[3]

This paper studies how to use scarce experimental resources (participants in a randomized experiment) to screen different potential innovations. To do so, we develop a simple model of optimal experimentation.[4] The model's details are based on how modern technology companies use online, large-scale experiments to guide their product innovation decisions. The key departure from previous work is that we allow for the distribution of innovation quality to have *fat tails*, thus allowing for outlier ideas. Despite the importance

[1]For example, when a user makes a search on a search engine like Bing or Google, the user is placed in about ten experiments, that are run in parallel. Experiments vary aspects such as different parts of the search algorithm and different parts of the user interface. Bing uses these experiments, called A/B tests, to evaluate almost all product innovations. This is now a common practice throughout the technology industry. A/B tests rose in popularity in cloud-based products that run on servers, like Bing and Google, where the costs of experimentation are low. But A/B tests are now increasingly used in traditional software products and other areas of business.

[2]See Duflo et al. (2007), Imbens (2010), Athey and Imbens (2017), and Deaton (2010).

[3]In technology, Kohavi et al. (2013) describe a team of Microsoft Bing engineers that increased the length of the description of each advertisement, by providing links to the target site. This idea required almost no coding effort, but went untested for a long time because other engineers did not expect it to be successful. Eventually the idea was tested, and shown to increase revenue by over tens of million dollars per year. Such anecdotes of black swans are common in many other companies. Other evidence suggestive of fat tails in innovation processes are the distribution of patent valuations (Silverberg and Verspagen, 2007) and the distribution of citations (Redner, 1998, Clauset et al., 2009). In behavioral interventions, there are examples such as defaults that have arguably generated very large effects (Madrian and Shea, 2001; Johnson and Goldstein, 2003).

[4]There are two broad strands of the theoretical literature on optimal experimentation. One strand follows the sequential decision problem proposed by Wald (1947). He and Arrow et al. (1949) considered the problem of acquiring costly information over time, and then making a decision. Recent contributions include Moscarini and Smith (2001), Fudenberg et al. (2017), Che and Mierendorff (2016), and Morris and Strack (2017). McClellan (2019) considers agency problems in this setting, and has a review of the literature on these agency problems. Banerjee et al. (2017) consider an optimal experimentation model to argue that randomized experiments are a good way to persuade an audience with heterogenous beliefs. The other strand follows the multi-armed bandit problem proposed by Thompson (1933) and Robbins (1985). They considered the problem of choosing which number of "arms" to pull over time, with arms having known payoff distributions. This sparked a fruitful literature, mostly in computer science. Bubeck and Cesa-Bianchi (2012) is an excellent overview. Bergemann and Valimaki (2008) give an overview of economic applications. Kasey and Sautmann (2019) study the problem of designing a dynamic experiment to choose between one of a number of designs, and develop an algorithm following the bandits literature on the best-arm identification problem.

of fat tails, they are assumed away in almost all of the optimal learning literature.[5]

Our main result is that the thickness of the tail of innovation quality is crucial for how to experiment. With sufficiently thin tails, the optimal strategy is similar to "big data" approaches commonly used by large technology companies. These companies carefully triage ideas, and conduct large-scale, statistically powerful experiments, that can detect even small benefits of a given product innovation. With sufficiently fat tails, however, the optimal strategy is similar to "lean" approaches used by many start-ups and entrepreneurs. These consist of running many small-scale experiments, and discarding any innovation without outstanding success.[6] More broadly, going beyond the internet setting, our results suggest that fat tails are important for optimal learning and experimentation models, at least in some settings.

We apply our model to data from A/B tests conducted at one of the largest experimentation platforms in the world at Microsoft Bing's search engine. We find evidence for fat tails, and that these fat tails have important consequences for how to run and interpret experiments.[7] Before going into details, we provide an overview of the paper.

Section 2 introduces the model. A risk-neutral firm has a set of ideas and a set of users. The quality of each idea is uncertain and is drawn independently from a prior distribution. To learn about the value of an idea, the firm can run an experiment on a subset of the users. The experiment produces a noisy signal of the quality of the idea. The firm's problem is how to assign its total budget of available users to the different ideas, and to then select which ideas to implement. We term this the A/B testing problem.

Section 3 derives our theoretical results. The decision of which ideas to implement is simple. The firm should use Bayes' rule to calculate the posterior mean of the quality of each idea, and implement ideas with a positive posterior mean quality (Proposition 1). The de-

[5] In the literature on the value of information the main results in Chade and Schlee (2002), Moscarini and Smith (2002), and Keppo et al. (2008) use bounded utility. In the optimal learning literature, the main results in Arrow et al. (1949); Fudenberg et al. (2017); Che and Mierendorff (2016) assume either a finite number of states or normally distributed utility. In the bandits literature, Bubeck et al. (2013) state "The vast majority of authors assume that the unknown distributions [...] are sub-Gaussian". They develop algorithms with asymptotically similar losses as the standard UCB algorithms for distributions with at least 2 moments, but worst bounds if no second moment exists. See also their survey Bubeck and Cesa-Bianchi (2012) section 2.4.7. The bandits literature is concerned with algorithms that achieve certain regret bounds in complex models where the optimal strategy is too complex to analyze, whereas we study properties of the optimal solution in a simple model.

[6] This is referred to as the lean startup methodology, and closely related to agile software development frameworks (Ries, 2011; Blank, 2013; Kohavi et al., 2013). The idea is to quickly and cheaply experiment with many ideas, abandon or pivot from ideas that do not work, and scale up ideas that do work.

[7] This part of out paper contributes to the recent academic literature that studies A/B testing in the tech industry. Goldberg and Johndrow (2017); Coey and Cunningham (2019) consider how to use data from many experiments to improve estimates from each experiment. They and Peysakhovich and Lada (2016); Peysakhovich and Eckles (2017) give evidence of fat tails in the quality of innovations. Feit and Berman (2018) considers a model of for how long to test new advertisements and how long to run them, and and Berman et al. (2018) give evidence of p-hacking in firms that use off-the-shelf A/B testing services.

cision of how to experiment depends on what we call the production function. We define the production function of an idea as the expected gain to the firm of allocating a number of users to experiment on the idea. It can be shown that the firm should allocate users to maximize the sum of production functions of all ideas (Proposition 2). Whether the production function has increasing or decreasing marginal returns determines the productivity of big and lean experiments.

Our main theoretical result relates the tails of the of the prior distribution of innovation quality to the shape of the production function. We assume that the prior distribution has tails that are approximately a power law with coefficient α.[8] We show that, for a relatively small number of users n, the marginal product of data is approximately proportional to

$$n^{\frac{\alpha-3}{2}}$$

(Theorem 2). This suggests that the tail coefficient α is the key parameter for understanding the shape of the production function, and the marginal returns of lean A/B tests.[9]

We use the theorem to formalize how the tails affect the optimal experimentation strategy (Corollary 1). With relatively fat tails ($\alpha < 3$), lean experimentation strategies are optimal. The intuition is that a large share of the gains from experimentation comes from finding black swans, and these outliers can be detected even with small experiments. In contrast, with relatively thin tails ($\alpha > 3$), and somewhat limited data, big data experimentation is optimal. The intuition is that, with thin tails, it is very unlikely that a small experiment will move the prior sufficiently to affect decision making. Therefore, it is best to run fewer but more precise experiments.

Besides fat tails, there is a simpler reason for lean experimentation. Allocating a large amount of users to test a single idea has eventually decreasing returns. Therefore, when experimental resources are sufficiently abundant, it is better to experiment on as many ideas as possible (remark 1). We also present results on asymptotics for large experiments, showing that marginal product decreases as $1/n^2$ (Theorem 1).

Section 4 applies our model to data from experiments at Microsoft Bing's EXP experimentation platform. EXP runs thousands of experiments per year, with the average experiment in our data having about 20 million users. Our theoretical framework leads to an Empirical Bayes problem (Robbins, 1964) in which the unobserved distribution of innovation quality can be nonparametrically identified from these data. We find both

[8] That is, the probability of an observation exceeding δ is roughly proportional to $\delta^{-\alpha}$.

[9] The production function is often referred to as the value of information (Radner and Stiglitz, 1984; Chade and Schlee, 2002; Moscarini and Smith, 2002; Keppo et al., 2008). A common finding in this literature is that the value of information is often convex close to zero. We contribute to this literature by showing that the tails of the distribution of innovation quality are a key determinant of whether this convexity result holds. With thin tails, a small experiment is of limited value, as it is unlikely to move the experimenter away from here prior beliefs. With fat tails, even small experiments can be valuable, because they can discover outliers.

reduced-form and structural evidence of fat tails. Our benchmark estimates for the key metric used to evaluate innovations is of a tail coefficient considerably smaller than 3. This result is statistically significant, robust to a number of alternative specifications, and consistent with evidence from similar products. Thus, the data suggest that fat tails can be important, even in large mature products like Bing.

The estimates have three sets of implications for experimentation in this setting. First, the fat tails affect the proper Bayesian updating of quality of an idea given experimental results. Ideas with small t-statistics should be shrunk aggressively, because they are likely to be lucky draws, whereas outlier ideas are likely to be real. In particular, the top 2% of ideas are responsible for 74.8% of the historical gains. This is an extreme version of the usual 80-20 Pareto rule. Second, the marginal value of data for experimentation is an order of magnitude lower than the average value, but is not negligible. Third, there are large gains from moving towards a lean experimentation strategy. We consider a counterfactual where Bing experiments on 20% more ideas, with the marginal ideas having the same quality distribution, while keeping the same number of users. We find that productivity would increase by 17.05%. Naturally, whether these gains can be attained depends on the costs of running additional experiments. We perform a back-of-the envelope calculation using Bing's monetary valuation for quality improvements. We find that moving towards lean experimentation would be profitable even if the fixed costs of one experiment were of the order of hundreds of thousands of dollars per year.

Section 5 presents additional results and robustness checks. First, we use the empirical estimates to understand when each of our asymptotic approximations is relevant. We find that the small sample size asymptotics of Theorem 2 provide a reasonable approximation to the production function in the Bing application.[10] Second, we provide a number of theoretical extensions: additional costs of experimentation (such as fixed costs, variable costs, and user experience costs), mutually exclusive ideas, different payoffs after implementation (such as a hypothesis-testing payoff), an elastic supply of ideas, and alternative assumptions about experimental noise. Third, we consider a limited data on triage procedures used at Bing to check whether marginal discarded ideas are worse than the average ideas tested. We find no evidence that the offline tests predict online performance. Finally, we report that the empirical results are robust to several alternative specifications.

[10] This is at first surprising, given that the typical sample sizes are of millions of users. But, in practice, user behavior is sufficiently noisy relative to effect sizes that Bing is far from the case of perfect information. Instead, outliers are important, which is the key approximation used in the asymptotic results for small sample sizes. In particular, in our application, this suggests that the force pushing towards lean experimentation is fat tails, as opposed to abundance of data.

2 The A/B Testing Problem

2.1 Model

A firm considers implementing potential *innovations* (or *ideas*) $I = \{1, \ldots, I\}$. The *quality* of innovation i is unknown and equals a real-valued random variable Δ_i, whose values we denote by δ_i. The distribution of the quality of innovation i is G_i. Quality is independently distributed across innovations.

The firm selects the number of users allocated to innovation i, n_i in $\mathbb{R}^+$, for an experiment (or A/B test) to evaluate it.[11] If $n_i > 0$, the experiment yields an estimator or *signal* equal to a real-valued random variable $\hat{\Delta}_i$, whose value we denote by $\hat{\delta}_i$. Conditional on the quality δ_i of the innovation, the signal has a normal distribution with mean δ_i and variance σ_i^2/n_i. The signals are assumed to be independently distributed across innovations. The firm faces the constraint that the total amount of allocated users $\sum_{i=1}^{N} n_i$ is at most equal to the number of users N available for experimentation. The firm's *experimentation strategy* is defined as the vector $\boldsymbol{n} = (n_1, \ldots, n_I)$.

After seeing the results of the experiments, the firm selects a *subset S of innovations to implement* (or to "ship") conditional on the signal realizations of the innovations that were tested. Formally, the subset S of innovations that are implemented is a random variable whose value is a subset of I, and is measurable with respect to the signal realizations. We also refer to S as the firm's *implementation strategy*.

The *firm's payoff*, which depends on both the experimentation and implementation strategies, is the sum of the quality of implemented innovations. The *A/B testing problem* is to choose an experimentation strategy $\boldsymbol{n}$ and an implementation strategy S to maximize the *ex ante expected payoff*

$$\Pi(\boldsymbol{n}, S) \equiv \mathbb{E}\left[\sum_{i \in S} \Delta_i\right]. \tag{1}$$

2.2 Discussion

One way to gain intuition about the model is to think about how it relates to our empirical application: the Bing search engine. The potential innovations I correspond to the thousand innovations that engineers propose every year. Bing triages these innovations, and selects a subset that makes it to A/B tests (by setting $n_i > 0$). These innovations are typically A/B tested for a week, with the average n_i of about 20 million users.[12] The number

[11] The number of users is assumed to be in the positive real line, because we are interested in experiments with sample sizes in the millions.

[12] It is common practice to require the duration of the experiments to be a multiple of weeks in order to avoid fishing for statistical significance and multiple testing problems; see Kohavi et al. (2013) p. 7. Also

N of users available for experimentation is constrained by the total flow of user-weeks in a year.[13]

We now discuss three important modeling assumptions. First, the gain from implementing multiple innovations is additive. This is a simplification because, in principle, there can be interactions in the effect of different innovations. Section 5 shows that allowing for mutually exclusive ideas does not change the main message of the paper. Interactions between ideas was the subject of an early debate at the time when A/B testing started being implemented in major technology companies (Tang et al., 2010; Kohavi et al., 2013). One proposal was to run multiple parallel experiments, and to analyze them in isolation, to increase sample sizes. Another proposal—based on the idea that interactions between innovations could be important—was to use factorial designs that measure all possible interactions. While both positions are theoretically defensible, the industry has moved towards parallel experiments because it was found that interactions were of second-order importance relative to the value of parallelization in running more precise experiments.

Second, there is no cost of running an experiment, so that the scarce resources are innovation ideas and data for experimentation. This assumption is for simplicity. Section 5 shows that introducing costs of experimentation does not change the main message of the paper. However, some readers may find it counter-intuitive that data is scarce, given the large sample sizes in major platforms. This point was raised in early industry discussions about A/B testing, where some argued that "there is no need to do statistical tests because [...] online samples were in the millions" (Kohavi et al., 2009b p. 2). Despite this intuitive appeal, this position has been discredited, and practitioners consider data to be scarce. For example, Deng et al. (2013) say that "Google made it very clear that they are not satisfied with the amount of traffic they have [...] even with 10 billion searches per month." And parallelized experiments are viewed as extremely valuable, which can only be the case if data is scarce (Tang et al., 2010; Kohavi et al., 2013). Data is scarce because large, mature platforms pursue innovations with small effect sizes, often of a fraction of a percent increase in performance (Deng et al., 2013).

Third, experimental errors are normally distributed. This is a reasonable assumption in our main application because the typical estimator for the unknown quality is a difference between sample means with i.i.d. data, and treatment/control groups are in the millions.

---

treatment effects often vary with the day of the week, so industry practitioners have found an experiment to be more reliable if it is run for whole multiples of a week (Kohavi et al., 2009a). While the timing in our model is simpler than reality, it is closer to practice than the unrestricted dynamic experimentation in bandit problems.

[13] Our model is related to the standard multi-armed bandit problem. The potential innovations I corresponds to the bandit arms. The number of available users N corresponds to the number of periods in the bandit problem. There are three key differences. First, the A/B testing problem ignores the payoffs during the experimentation phase because, in practice, they are dwarfed by payoffs after implementation. Second, multiple innovations can be implemented. Third, the timing of the A/B testing problem is simpler: there are no dynamics.

2.3 Assumptions and Notation

We assume that the distribution G_i has a finite mean, a smooth density g_i with bounded derivatives of all orders, and that $g_i(0)$ is strictly positive. These assumptions will be maintained throughout the paper, unless otherwise stated.

We use the following notation. Two functions h_1 and h_2 are *asymptotically equivalent* as n converges to n_0 if

$$\lim_{n \to n_0} \frac{h_1(n)}{h_2(n)} = 1.$$

This is denoted as $h_1 \sim_{n_0} h_2$, and we omit n_0 when there is no risk of confusion.

Given a sample size $n_i > 0$ for experiment i and signal realization $\hat{\delta}_i$, denote the *posterior mean of the quality* Δ_i of innovation i as

$$P_i(\hat{\delta}_i, n_i) \equiv \mathbb{E}[\Delta_i | \hat{\Delta}_i = \hat{\delta}_i \, ; \, n_i].$$

If $n_i = 0$, we abuse notation and define $P_i(\hat{\delta}_i, n_i)$ as the unconditional mean of Δ_i.

Because the experimental noise is normally distributed, it is known that $P_i(\cdot, n_i)$ is smooth and strictly increasing in the signal provided $n_i > 0$. Moreover, there is a unique *threshold signal* $\delta_i^*(n_i)$ such that $P_i(\delta_i^*(n_i), n_i) = 0$ (see Lemma A.1).

3 Theoretical Results

3.1 The Optimal Implementation Strategy

The optimal implementation strategy is simple. The firm observes the signal $\hat{\delta}_i$, calculates the posterior mean $P_i(\hat{\delta}_i, n_i)$ using Bayes' rule, and implements innovation i if this posterior mean is positive. We formalize this immediate observation as the following proposition.

Proposition 1 (Optimal Implementation Strategy)**.** *Consider an arbitrary experimentation strategy $\boldsymbol{n}$ and an implementation strategy S^* that is optimal given $\boldsymbol{n}$. Then, with probability one, innovation i is implemented iff the posterior mean innovation quality $P_i(\hat{\delta}_i, n_i)$ is positive.*

The proof of this proposition and all of our theoretical results are collected in the Appendix. In practice, the most common implementation strategy is to implement an innovation if it has a statistically significant positive effect at a standard significance level, typically 5%. Other versions of this strategy adjust the critical value to account for multiple hypothesis testing problems. Proposition 1 shows that these approaches are not optimal. Instead, the optimal is to base implementation decisions on the posterior mean.

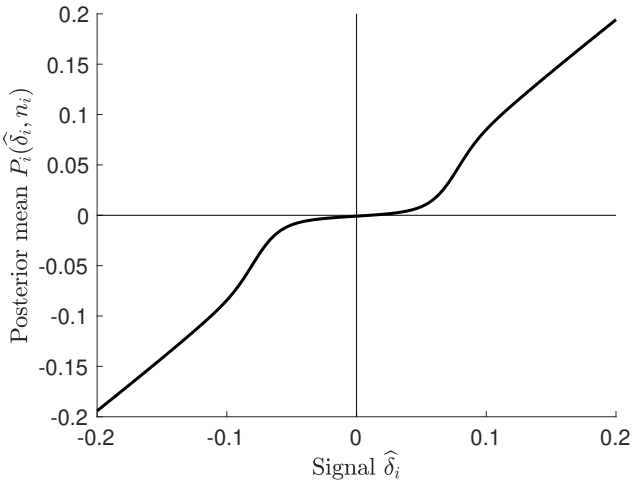


Figure 1: The posterior mean function $P_i(\hat{\delta}_i, n_i)$

Notes: The figure assumes a prior with a Student t-distribution and parameters equal to our benchmark empirical estimates from Section 4. The parameter σ_i is set to the average in the data and n_i is set to the typical value of 20 million users, so that the standard error $\sigma_i/\sqrt{n_i}$ is 0.022.

Fat tails can have a substantial impact on the posterior mean function. To illustrate this point, Figure 1 plots the posterior mean function for a fat-tailed prior with a t-distribution with a small negative mean. With this fat-tailed prior, the posterior mean function is flat close to zero, so that typical observations with small t-statistics should be considerably shrunk. The intuition is that typical observations with good results are likely to be due to lucky experimental draws. However, for large outliers, the posterior mean function approaches the diagonal. The intuition is that, with fat tails but normal experimental noise, large outliers are likely to be due to idea quality, and not to experimental noise.[14] We will return to this issue when we discuss our empirical findings.

3.2 The Production Function

The value of potential innovation i with no data equals its mean, provided that it is positive,

$$\mathbb{E}[\Delta_i]^+.$$

If the firm combines innovation i with data from n_i users, the firm can run the experiment, and only implement the idea if the posterior mean quality is positive. By Proposition 1, the total value of A/B testing innovation i is the expected value of the positive part of the posterior mean; this is

$$\mathbb{E}[P_i(\hat{\Delta}_i, n_i)^+].$$

[14] See Morris and Yildiz (2016) for similar results in a coordination game.

Thus, the value of investing data from n_i users into potential innovation i equals

$$f_i(n_i) \equiv \mathbb{E}[P_i(\hat{\Delta}_i, n_i)^+] - \mathbb{E}[\Delta_i]^+. \tag{2}$$

We term $f_i(n_i)$ the *production function* for potential innovation i. We term $f_i'(n_i)$ as the *marginal product of data* for i. With this notation, the firm's payoff can be immediately decomposed as follows.

Proposition 2 (Production Function Decomposition)**.** *Consider an arbitrary experimentation strategy $\boldsymbol{n}$ and an implementation strategy S that is optimal given $\boldsymbol{n}$. Then the firm's expected payoff is*

$$\Pi(\boldsymbol{n}, S) = \underbrace{\sum_{i \in I} \mathbb{E}[\Delta_i]^+}_{\text{value of ideas with no data}} + \underbrace{\sum_{i \in I} f_i(n_i)}_{\text{additional value from data}}.$$

That is, the payoff equals the sum of the gain from innovations that are profitable to implement even without an experiment, plus the sum of the production functions of the data allocated to each experiment. The production functions are differentiable for $n_i > 0$.

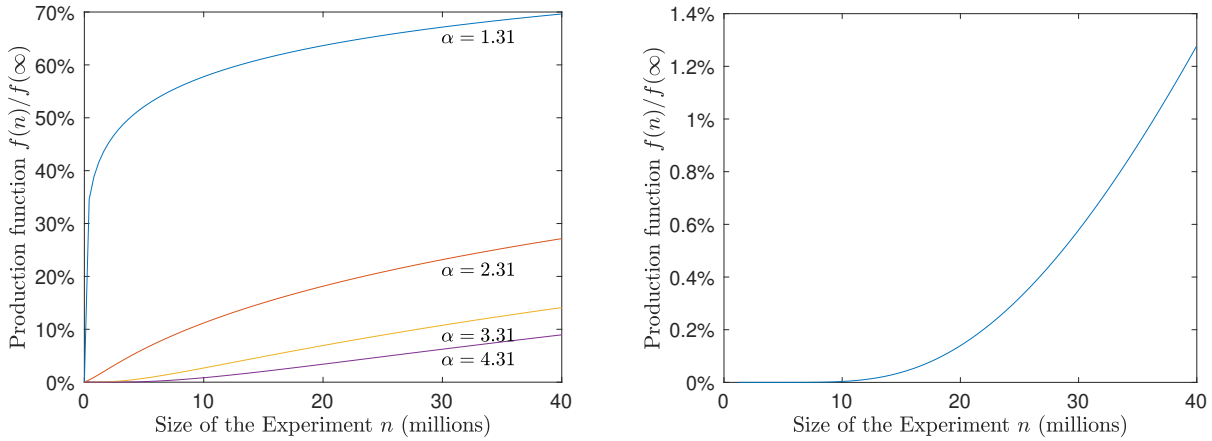
This straightforward decomposition reduces the A/B testing problem to constrained maximization of the sum of the production functions. Therefore, the shape of the production function is a crucial determinant of the optimal innovation strategy. Figure 2 plots the production function with illustrative model primitives for sample sizes up to 40 million users. Panel B depicts the case of a normal prior. Panel A depicts the case of a fat-tailed t-distribution, for varying tail coefficients. The figure shows that the production function can have either increasing or decreasing returns to scale, and that the shape of the production function depends on the tail coefficients of the prior distribution.

3.3 Main Results: Shape of the Production Function

This section develops our main theoretical results, which characterize the shape of the production function (and consequently speak to the optimal experimentation strategy). Throughout this subsection, we consider a single innovation, and omit the subscript i for clarity. To describe the optimal implementation strategy, define the *threshold t-statistic* $t^*(n)$ as the t-statistic associated with the threshold signal, $t^*(n) = \delta^*(n)/(\sigma/\sqrt{n})$. We remind the reader that $\delta_i^*(n_i)$ is defined as the unique threshold signal such that $P_i(\delta_i^*(n_i), n_i) = 0$.

We establish two theorems. The first theorem characterizes the production function for very large sample sizes (large-n approximation), in the limit where the experiment is much more informative than the prior.

Theorem 1 (Production Function for Large n)**.** *Consider n converging to infinity. We have the following.*



(a) Student t prior (b) Normal prior

Figure 2: The production function $f(n)$ with and without fat tails

Notes: The figures plot the production function as a fraction of the value of perfect information, $f(n)/f(\infty)$. Panel A depicts a Student t prior, and Panel B depicts a normal prior. The t-distribution parameters correspond to the benchmark empirical estimates in Section 4.5 for our main metric, but with varying tail coefficients. The normal prior has the same mean and scale parameters. σ_i is set to the average value in the data.

1. *The threshold t-statistic $t^*(n)$ converges to 0. Moreover, if $g'(0) \neq 0$,*

$$t^*(n) \sim -\frac{\sigma}{\sqrt{n}} \cdot \frac{g'(0)}{g(0)}.$$

2. *Marginal products converge to 0 at a rate of $1/n^2$. More precisely,*

$$f'(n) \sim \frac{1}{2} \cdot g(0) \cdot \sigma^2 \cdot \frac{1}{n^2}. \tag{3}$$

The theorem shows that, for very large samples, the marginal product of additional data declines rapidly. Moreover, this holds regardless of details about the distribution of ideas, which only affects the asymptotics up to a multiplicative factor. The intuition is that additional data only helps to resolve edge cases, where the value of an innovation is close to 0. Mistakes about these cases are not very costly, because even if the firm gets them wrong the associated loss is small.[15]

The intuition of the proof is as follows. Lemma A.2 in the appendix shows that, for all n,

[15]This argument echoes themes developed by Vul et al. (2014) for more special distributions and by Fudenberg et al. (2017) in dynamic learning context. Moscarini and Smith (2002); Keppo et al. (2008) find that, in models with a finite number of states of the world, the value of information decays exponentially for large n. The reason for the difference is that we have an infinite number of states, and a strictly positive density $g(0)$.

marginal product equals

$$f'(n) = \frac{1}{2n} \cdot m(\delta^*(n)\ ,\ n) \cdot \text{Var}\left[\Delta|\hat{\Delta} = \delta^*(n)\right], \tag{4}$$

where $m(\cdot\ ,\ n)$ is the marginal distribution of the signal $\hat{\Delta}$. The marginal product depends in an intuitive way on the elements of this formula. It is more likely that additional data will be helpful if the existing estimate has few data points n, if the likelihood $m(\delta^*(n), n)$ is large, and if there is a lot of uncertainty about quality conditional on the marginal signal. The proof gives further intuition of why the exact formula holds. We then proceed to show that $m(\delta^*(n)\ ,\ n) \cdot \text{Var}\left[\Delta|\hat{\Delta} = \delta^*(n)\right] \sim g(0)\sigma^2/n^2$. Intuitively, this result can be thought of as a consequence of the Bernstein-von Mises theorem, which says that Bayesian posteriors are asymptotically normal with the same mean and variance as the maximum likelihood estimator. This implies that the threshold $\delta^*(n)$ is close to zero, and the conditional variance in equation (4) is σ^2/n. Thus, the general formula (4) simplifies to the asymptotic formula (3). The formula for the threshold t-statistic is derived from an asymptotic version of Tweedie's formula (Corollary A.1).

Theorem 1 may only hold for extremely large sample sizes. For example, in Figure 2, even experiments with tens of millions of users only generate a fraction of the value of perfect information. The theorem implicitly relies on a Bernstein-von Mises type approximation where there is so much data that the prior is uninformative. This only happens when the experiments are much more precise than the variation in the quality of ideas. Even large platforms like Bing are far below this scale, as in the anecdotal evidence cited in section 2.2, and in the empirical evidence we give below.

Matters are different for small n, where the shape of g has dramatic effects on the shape of f. The next theorem shows that, if the ex ante distribution of idea quality has Pareto-like tails,[16] the marginal product is determined by the thickness of the tails.

Theorem 2 (Production Function for Small n)**.** *Assume that the distribution of innovation quality satisfies $g(\delta) \sim \alpha c(\delta) \cdot |\delta|^{-(\alpha+1)}$ as δ converges to $\pm\infty$, where $c(\delta)$ is a slowly varying function, $\alpha > 1$ and $\alpha \neq 3$. Assume there is a constant $C > 0$ such that $c(\delta) > C$ for large enough $|\delta|$. Assume also that $\mathbb{E}[\Delta] < 0$ and consider n converging to 0. We have the following.*

1. *The threshold t-statistic $t^*(n)$ converges to infinity at a rate of $\sqrt{\log 1/n}$. More precisely,*

$$t^*(n) \sim \sqrt{2(\alpha - 1)\log \frac{\sigma}{\sqrt{n}}}.$$

[16] The p.d.f.s covered by Theorem 2 include the generalized Pareto density of Pickands (1975), affine transformations of the t-distribution (which is the model used in our empirical application), and any distribution where the tails are Pareto, Burr, or log gamma.

2. *Marginal products are roughly proportional to* $n^{\frac{\alpha-3}{2}}$*. More precisely,*

$$f'(n) \sim \frac{1}{2} \cdot \alpha c(\delta^*(n)) \cdot (\sigma t^*(n))^{-(\alpha-1)} \cdot n^{\frac{\alpha-3}{2}}.$$

3. *If the tails of* g *are sufficiently thick so that* $\alpha < 3$*, then the marginal product at* $n = 0$ *is infinity.*

4. *If the tails of* g *are sufficiently thin so that* $\alpha > 3$*, the marginal product at* $n = 0$ *is zero.*

The theorem states that, for small n, $f'(n)$ behaves as approximately proportional to $n^{\frac{\alpha-3}{2}}$. This behavior determines the marginal returns of the production function in small A/B tests. Much like in neoclassical producer theory, this behavior is crucial for the optimal experimentation strategy. With relatively thin tails $\alpha > 3$, marginal products are increasing (and zero at $n = 0$), and we have increasing returns to scale. With relatively thick tails, marginal products are decreasing (and infinite at $n = 0$), so that we have decreasing returns to scale. These cases are illustrated in Figure 2.

The intuition for the theorem is as follows. If g is not sufficiently fat tailed, $\alpha > 3$, then a small bit of information is unlikely to change the optimal action as it is too noisy to overcome the prior. A bit of information is therefore nearly useless. Only once the signal is strong enough to overcome the prior does information start to become useful. This makes the value of information convex for small sample sizes. This intuition has been formalized in a classic paper by Radner and Stiglitz (1984). They consider a setting that is, in some ways, more general, but that precludes the possibility of fat tails. Because they assumed away fat tails, they concluded that the value of information is generally convex for small n. Our theorem shows that their conclusion is reversed in the fat tailed case.[17]

Our theorem shows that if $\alpha < 3$, most of the value of experimentation comes from a few outliers and even extremely noisy signals will suffice to detect them. More precise signals will help detect smaller effects, but if most of the value is in the most extreme outliers, such smaller effects have quickly diminishing value. Thus, the value of information is concave for small n.

At first sight, it is not clear why the dividing line is $\alpha = 3$. As it turns out, $\alpha = 3$ can be explained with a simple heuristic argument. Consider a startup firm that uses a lean

[17] One reason for the difference between our result and Radner and Stiglitz (1984) is the units of information that they use. In their leading example (p. 40), they measure the quantity of information by the correlation between a signal and the state, which is roughly proportional to $\sqrt{n}$. The value of information in their example is roughly proportional to n, so that, as they explain, the convexity result depends on the units. However, allowing for fat tails is the main reason for the different results. Chade and Schlee (2002) explain that, even when information is measured by n, the convexity result depends on substantial assumptions. Nevertheless, they show that convexity does hold with some generality. And Keppo et al. (2008) find a convex value of information in a natural model with thin tails. Our result clarifies that one practically important reason why we can have a concave value of information close to 0 is the existence of fat tails.

experimentation strategy. The firm tries out many ideas in small A/B tests, in hopes of finding one idea that is a big positive outlier. Even though the A/B tests are imprecise, the firm knows that, if a signal is several standard errors above the mean, it is likely to be an outlier. So the firm decides to only implement ideas that are, say, 5 standard errors above the mean. This means that the firm will almost certainly detect all outliers that are more than, say, 7 standard errors above the mean. This yields value

$$f(n) \propto \int_{7\sigma/\sqrt{n}}^{\infty} \delta g(\delta)\, d\delta \propto \int_{7\sigma/\sqrt{n}}^{\infty} \delta\delta^{-(\alpha+1)}\, d\delta = \int_{7\sigma/\sqrt{n}}^{\infty} \delta^{-\alpha}\, d\delta.$$

Integrating we get

$$f(n) \propto \frac{1}{\alpha-1}(7\sigma/\sqrt{n})^{-(\alpha-1)} \propto n^{\frac{\alpha-1}{2}}.$$

Thus, the marginal product is proportional to $n^{\frac{\alpha-3}{2}}$, as in the theorem.

The proof of the theorem formalizes and generalizes this heuristic. The starting point is to show that the first order condition for the optimal threshold, and the marginal products can be written as integrals. These integrals are dominated by regions where either quality is in the mean of its distribution, but the signal is extreme, or where the signal is in the middle of its distribution, but true quality is extreme. This implies that these integrals can then be approximated by closed-form expressions, due to the power law assumption. The theorem can then be derived with calculations in the lines of the heuristic argument.

The heuristic argument clarifies why the small-n approximation can be useful for the sample sizes in real applications, that often include millions of users. The key approximation used in the proof of Theorem 2 is that outlier ideas are responsible for a large share of the gains. As long as this is true, the theorem can provide a useful approximation. Indeed, Section 5.1 shows that, in our empirical application, the approximations in Theorem 2 are useful for the typical sample sizes of tens of millions of users.

3.4 The Optimal Experimentation Strategy

We now use the results to understand the optimal experimentation strategy. The simplest case is when experimental resources are abundant, so that we are close to having perfect information. That is, the case where N converges to infinity.

Remark 1. If N is large enough, then it is optimal to run experiments on all ideas (that is, it is optimal to "go lean").

With plentiful data, it is optimal to experiment on every idea because the returns to data are eventually decreasing. This follows because the production function is increasing and bounded by the value of perfectly observing idea quality. For example, all of the production functions in Figure 2 are eventually bounded by 100% of the value of perfect

information. So allocating a large number of users to a single idea eventually has small returns, and it is better to experiment on as many ideas as possible.

Assuming that experimental resources are abundant is equivalent to assuming that there are enough users to make the noise of each A/B test almost negligible. The following corollary characterizes the optimal experimentation strategy when there the outcome of an A/B test does not fully reveal the innovation quality.

Corollary 1. *[Optimal Experimentation Strategy]Assume that all ideas have the same prior distribution of quality, that this distribution satisfies the assumptions of Theorem 2, and that there is more than one idea.*

- *If the distribution of quality is sufficiently thick-tailed, $\alpha < 3$, it is optimal to run experiments on all ideas (that is, it is optimal to "go lean").*
- *Suppose in addition that the slowly varying function in Theorem 2 satisfies $c(\delta) \to c$ as $\delta \to \infty$. If the distribution of quality is sufficiently thin-tailed, $\alpha > 3$, and if N is sufficiently small, the firm should allocate all experimental resources to one idea (that is, it is optimal to "go big").*

The corollary relates the experimentation strategy to the tail of the distribution of innovation quality. If the distribution of innovation quality is sufficiently thin-tailed, most ideas are marginal improvements. The production function is convex close to $n = 0$, because obtaining a small amount of data is not sufficient to override the default implementation decision. In this case, it is optimal to choose only one idea, and run a large, high-powered experiment on it. We call this strategy "big data A/B testing" as it involves ensuring the experiment has a large enough sample to detect fairly small effects. This strategy is in line with common practice in many large technology companies, where ideas are carefully triaged, and only the best ideas are taken to online A/B tests.

If the distribution of innovation quality is sufficiently thick tailed, a few ideas are large outliers, with very large negative or positive impacts. These are commonly referred to as black swans, or as big wins when they are positive. The production function is concave and has an infinite derivative at $n = 0$. The optimal innovation strategy in this case is to run many small experiments, and to test all ideas. We call this the "lean experimentation" strategy, as it involves running many cheap experiments in the hopes of finding big wins (or avoiding a negative outlier). This strategy is in line with the lean startup approach, which encourages companies to quickly iterate through many ideas, experiment, and pivot from ideas that are not resounding successes (Blank, 2013).

4 Empirical Application

4.1 Setting

We now apply the model to a major experimentation platform, Microsoft's EXP. This is an ideal setting because we have detailed data on thousands of A/B tests that have been performed in the last few years. We can use the data to estimate the ex ante distribution of innovation quality, and thus understand the importance of fat tails and the optimal experimentation strategy.

EXP was originally part of the Bing search engine, but has since expanded to help several products within Microsoft run A/B tests. This expansion coincides with the rise of A/B testing throughout the technology industry, due to the large increase in what is known as cloud-based software. Traditional client-based software, like Microsoft's Word or Excel, runs locally in users' computers. Innovations used to be evaluated offline by product teams, and implemented in occasional updates. In contrast, cloud based software, like Google, Bing, Facebook, Amazon, or Uber, mostly runs on server farms. In these cloud-based products, most innovations are evaluated using A/B tests, and are developed and implemented at scale in an agile workflow. These practices have spread, and even traditional software products like Microsoft Office now use A/B testing.

EXP runs thousands of A/B tests per year. Our empirical analysis focuses on Bing, which is itself a large product. Bing has revenues of the order of 7 billion USD per year, and makes comparable investments in engineering. Bing serves over 12 billion monthly searches worldwide. In the US, Bing had 136 million unique searchers in 2017, with about 1/3 market share in the PC market.[18]

There are three key empirical challenges to obtain reliable estimates of the distribution of innovation quality. First, the distribution g_i represents the prior information about idea i. Thus, to estimate g_i, even with perfect observations of the realized true quality δ_i, we need many observations of ideas that engineers see as coming from the same distribution. To illustrate this problem, imagine that engineers test a set of ideas that look good, and have a distribution g_1, and ideas that look bad and have a distribution g_2. If we do not observe which ideas are good and which are bad, we would incorrectly think that the ex-ante distribution of ideas is an average of g_1 and g_2.

The second challenge is that many online A/B tests have potential data issues: they are experimental flukes. These data issues arise because running many parallel A/B tests in a major cloud product is a difficult engineering problem. The simplest examples are failures of randomization, which can be detected when there is a statistical difference between the

[18] Based on 2017 Comscore data and on Microsoft's form 10Q for the quarter ending on March 31, 2018.

number of users in treatment and control groups.[19] This kind of measurement error can bias estimates of the distribution of innovation quality. For example, if true effects are normally distributed, but experimental flukes produce a few large outliers, a researcher may incorrectly conclude that the distribution of true effects is fat tailed.

The third challenge is that our model assumes that innovations can be identified by a single quality metric, that is additive across different innovations. In practice, there are multiple possible performance measures that can be used. Also, it is not unreasonable to think that some innovations can be complements or substitutes.

4.2 Data Construction

We have detailed data on the universe of experiments EXP ran on Bing from January 2013 to June 2016. The data includes dates, areas experimented, experimental results of thousands of metrics disaggregated across dimensions such as geography, language, and device types, details about the experimental procedure, comments, and identity of the owner of the experiment.

To alleviate the empirical problems above, we restricted our sample in four ways. First, we restricted attention to relatively homogeneous areas because engineers consider the prior on ideas to be ex ante homogeneous.[20] Second, we restricted attention to experiments that are similar to the basic version of our model.[21] Third, to ensure high data quality we restricted attention to the US market and dropped A/B tests with signs of experimental problems. Fourth, we restricted attention to user experience areas so idea quality is well summarized by a few key metrics. See the supplementary appendix for further details.

The main quality measure we consider is a proprietary success metric that we call **session success rate** or simply **success rate**. The success rate for a user is the proportion of queries where the user found what she was looking for. This measure is calculated

[19] Many other, more complex experimental problems commonly happen. For example, Bing caches the first few results of common queries. For the experiments to be valid, every user has to cache the data for all the versions of all the experiments that she takes part on, even for the treatments that she will not be exposed to. This both creates a cost of the experimentation platform, since it slows down the website as a whole, and creates a challenge to run a valid experiment. As a final example, consider a treatment that slows down a website. This treatment could cause a instrumentation issue if it makes it easier for clicks to be detected. So, even if the treatment worsens user experience, it could seem to be increasing engagement, only because it made it easier to detect clicks (Kohavi and Longbotham, 2011).

[20] Engineers currently view ideas in a relatively even footing because of their previous experience with A/B tests. Previous A/B tests revealed that it is very hard to predict which innovations are effective ex-ante, and sometimes the best innovations come from unexpected places. Kohavi et al. (2009b, 2013) describe their experience running experiments at Bing as "humbling." One of their major tenets is that "we are poor at assessing the value of ideas." They give several examples of teams in other companies that have reached similar, if not even more extreme, conclusions.

[21] We dropped, for example, experiments with multiple treatments or that only applied to narrow areas.

from detailed data on user behavior. The success rate is a good overall measure of performance, and plays a key role in the decision to implement an idea at scale. While our main analysis uses success rate, we will consider some of these other metrics in robustness and placebo analyses. First, we consider three alternative metrics based on short-run user interactions, much like success rate. We refer to them as **alternative short-run metrics #1, #2, and #3**. These metrics help us validate our methodology, because qualitative results should be similar to the results for success rate. We consider two long-run metrics, that measure overall user engagement. We refer to them as placebo **long-run metrics #1 and #2**. Engineers consider the long-run metrics more important. However, it is hard to detect movements in these metrics, which is why most shipping decisions are based on short-run metrics such as success rate. We use these metrics as placebos to validate our methodology, because they should have a small amount of signal relative to the experimental noise, and experiments should have a low value. All of the metrics we use are measured at the user level, which is also the level of randomization of the experiments. Although these metrics use different units, engineers commonly consider percentage improvements. We define the **delta** of a metric in an experiment as the raw effect size divided by the control mean, defined in percent. In the remainder of the paper, we will use deltas to analyze experiments across all metrics. We refer to the sample delta in a metric in an experiment, or signal, as the sample estimate of the percentage improvement. This corresponds to the signal $\hat{\delta}_i$ in the theoretical model.

Finally, we performed a detailed audit on the data. We contacted the owners of all experiments in the tails of the outcome distribution, and of a random sample of experiments in the head of the distribution. We used the audit to weed out experiments that were minor tweaks of the same idea, or that owners considered to be unreliable because of potential data issues. We found that 50% of the audited observations had concerns. To be conservative, we estimated our model in two ways. The first is a standard maximum likelihood estimator that ignores the potential data concerns (except for dropping invalid observations). The second is a weighted maximum likelihood strategy that weights each observation by its reliability (where the reliability is estimated using the audit data).[22] The two estimators produce similar results (Supplementary appendix Figure B.3). For that reason, we report the standard maximum likelihood estimator in the body of the paper. Supplementary Appendix B derives the theoretical properties of the alternative estimator, gives further details on the audit data, and reports the results.

[22] In the appendix we show that the weighted maximum likelihood estimator is consistent and asymptotically normal in a model where the data is a mixture of the distribution of interest an a non-parametric distribution for the observations with data reliability concerns.

4.3 Descriptive Statistics

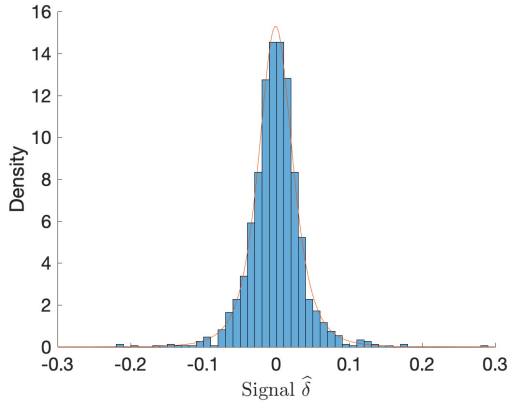
Table 1 displays summary statistics, at the level of experiments. The table reveals three striking facts. First, Bing conducts large experiments, with the average experiment having about 20 million subjects. This reflects both the fact that Bing has a substantial number of active users, and also the fact that experiments are highly parallelized. These large sample sizes are translated in precise estimation of all metrics. For example, the average standard error for success rate is of only 0.029%.

Table 1: Summary statistics: experiments

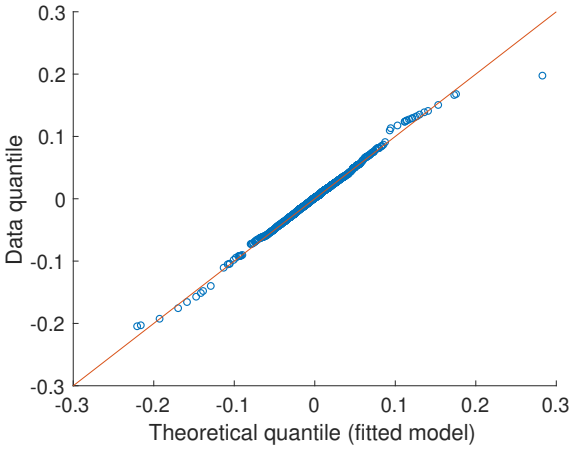
	Mean	Min	Max	Standard deviation	Interquartile range
All experiments (N = 1,450)					
Number of subjects	19,447,892	2,005,051	125,837,134	16,539,352	
Duration (days)	10.84	7.00	28.00	4.69	
Probability valid	0.52	0.25	1.00	0.10	
Sample delta					
Success rate	−0.001%	−0.220%	0.283%	0.036%	0.035%
Short-run metric #1	−0.003%	−0.234%	0.345%	0.035%	0.033%
Short-run metric #2	−0.019%	−11.614%	3.289%	0.434%	0.139%
Short-run metric #3	−0.004%	−0.465%	0.504%	0.066%	0.058%
Long-run metric #1	0.001%	−2.157%	0.669%	0.157%	0.153%
Long-run metric #2	0.002%	−0.565%	0.432%	0.084%	0.090%
Sample delta standard error					
Success rate	0.029%	0.009%	0.091%	0.013%	
Short-run metric #1	0.025%	0.009%	0.072%	0.011%	
Short-run metric #2	0.103%	0.035%	0.271%	0.040%	
Short-run metric #3	0.044%	0.012%	0.120%	0.020%	
Long-run metric #1	0.158%	0.045%	0.459%	0.075%	
Long-run metric #2	0.092%	0.030%	0.255%	0.044%	

The second fact is that effect sizes of the studied interventions are also small. The mean sample deltas are close to zero, for all metrics. The standard deviation of the sample delta for success rate is of only 0.036%. This reflects the fact that Bing is a mature product, so that it is hard to make innovations that have, on their own, a very large impact on overall performance. Even though the effects are small in terms of metrics, they are considered important from a business perspective. Practitioners consider that the value of a 1% improvement in success rate is of the order of hundreds of millions of dollars. Thus, even gains of the order of 0.1% are substantial, and worth considerable engineering effort.

Third, the summary statistics suggest that the distribution of measured effects is fat tailed. Many experiments have very small measured deltas, while a handful show substantial



(a) Histogram of success rate and estimated density.



(b) Q–Q plot of model fit.

Figure 3: Model fit

Notes: A histogram (panel A) and Q–Q plot (panel B) showing the raw data on measured deltas in success rate and the fit of the benchmark model estimates.

gains. This can be seen in the histogram in Figure 3a. The summary statistics display signs of fat tails: for all metrics, the largest absolute value deltas are several standard deviations away from the mean.

Figure 4 displays a log-log plot of the tail distribution of sample delta. That is, the log of the rank of each observation versus the log of the observation. Log-log plots are a standard way to visualize fat-tailed distributions. If the variable $\hat{\delta}_i$ has a Pareto distribution with tail parameter α, then, with infinite data, the log-log plot is a straight line with a slope of negative α. Consistent with fat tails, Figure 4 displays relatively low slopes. Indeed, we simulated data with the same sample size and variance, but coming from a normal distribution, and found substantially higher slopes. Log-log plots suffer from well-known problems in finite samples (Clauset et al., 2009). For that reason, we use the log-log plots to transparently describe the data, but the slopes are not a reliable way to estimate the precise tail coefficients. Thus, we will use a maximum likelihood procedure for our benchmark estimates.

4.4 Identification and Maximum Likelihood Estimation

Fix a metric of interest (for example, success rate). We would like to estimate the metric's ex ante distribution of innovation quality, which is the key parameter of our model (and which we will henceforth denote succinctly by the density g). In the A/B testing problem there are at least two functions of g that are empirically relevant. The first one is the posterior mean of δ_i given $\hat{\delta}_i$, which provides the optimal estimator (from a decision-theoretic

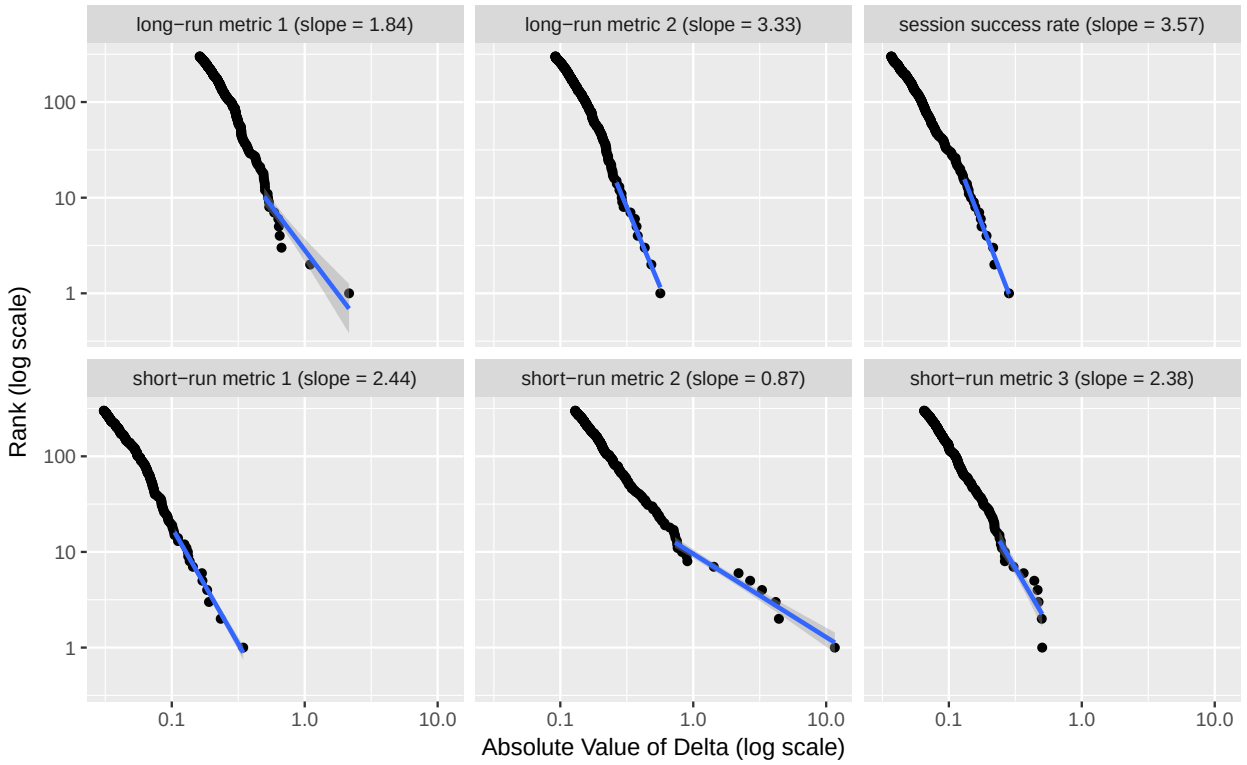


Figure 4: Log-log plots of the tails of the distribution of sample deltas

Notes: Each figure plots, in a log-log scale, the rank of the absolute value of sample deltas, versus the absolute value of sample delta $|\hat{\delta}_i|$. Each panel corresponds to a metric. Each plot has 150 observations, and the 15 largest observations are used to calculate the slope.

perspective) for the unobserved quality of idea i. The second one is the optimal experimentation strategy $\mathbf{n}(g)$, which is defined as the experimentation strategy that maximizes the firm's expected payoff subject to user availability.

Our empirical strategy is to use the outcomes of the different A/B tests to estimate g. We then construct 'plug-in' estimators of the functions of g that are relevant in the A/B testing problem. For example, if $\hat{g}$ denotes the estimator of g, then $\mathbf{n}(\hat{g})$ is the plug-in estimator of the optimal experimentation strategy. These type of procedures are usually called Empirical Bayes estimators, as the prior is estimated from the data.[23]

To formalize our strategy, we start by summarizing each A/B test i affecting the corresponding metric using the triplet

$$(\hat{\delta}_i, \sigma_i, n_i), \tag{5}$$

where $\hat{\delta}_i$ denotes the estimated *delta* of idea i, $\sigma_i/\sqrt{n_i}$ is the estimated standard error, and n_i is the sample size.[24]

[23] Azevedo et al. (2019) review parametric and nonparametric Empirical Bayes approaches to estimate the distribution of unobserved quality given previous outcomes of A/B tests.

[24] For notational simplicity—and given that we will estimate the ex ante distribution of idea quality sepa-

Following the theoretical analysis from Section 3, the distribution of $\hat{\delta}_i$ is given by a two-stage hierarchical model:[25]

$$\delta_i \text{ is distributed according to } g, \tag{6}$$

$$\hat{\delta}_i|\delta_i \text{ is distributed as a } \mathcal{N}(\delta_i, \sigma_i^2/n_i). \tag{7}$$

That is, the estimator $\hat{\delta}_i$ is normally distributed with known variance given the true quality δ_i. This is a reasonable assumption because of the large sample sizes in each experiment. This makes the errors approximately normally distributed, and the standard estimate for the sample variance is consistent and precisely estimated relative to treatment effects.

NONPARAMETRIC IDENTIFICATION OF g: The prior g is nonparametrically identified. To see this, note that the unconditional distribution of $\hat{\delta}_i$ equals the sum of two independent random variables:

$$\hat{\delta}_i = \delta_i + (\sigma_i/\sqrt{n_i})\epsilon, \text{ where } \delta_i \text{ has p.d.f. } g, \ \epsilon_i \text{ is } \mathcal{N}(0,1), \text{ and } \delta_i \perp \epsilon_i.$$

If we let $\psi_X(t)$ denote the characteristic function of X at point t, it is straightforward to see that:

$$\psi_\delta(t) = \psi_{\hat{\delta}_i}(t) \Big/ \exp\left(-\frac{1}{2}\frac{\sigma_i^2}{n_i}t\right). \tag{8}$$

It is a well-known fact that any probability distribution is characterized by its characteristic function (Billingsley (1995), Theorem 26.2, p. 346). Consequently, g is non-parametrically identified from the unconditional distribution of $\hat{\delta}_i$, which in principle can be estimated using data for different A/B tests with similar σ_i.[26]

MAXIMUM LIKELIHOOD ESTIMATION: Although the ex ante distribution of idea quality, g, is non-parametrically identified, we estimate our model imposing parametric restrictions

rately for each metric—we omit the use of subscript m throughout this section.

[25] Hierarchical models are used extensively in Bayes and Empirical Bayes statistical analysis (see Chapters 2 and 3 in Carlin and Louis (2000)). Two-stage hierarchical models are also known as *mixture models* (Seidel (2015)), where g is typically called the *mixing* distribution.

[26] The identification argument above has been used extensively in the econometrics and statistics literature; see Diggle and Hall (1993) for a seminal reference. If, contrary to our assumption, the distribution of $(\sigma/\sqrt{n})\epsilon$ were unknown, non-parametric identification of g would not be possible unless additional data is available or additional restrictions are imposed; see for example Li and Vuong (1998).

on g.[27] In particular, we assume that

$$\delta \sim M + s \cdot t_\alpha, \tag{9}$$

where $M \in \mathbb{R}$, $s \in \mathbb{R}_+$, and t_α is a t-distributed random variable with α degrees of freedom. This means that we can write the second stage of our hierarchical model as

$$\delta \text{ has distribution } g(\cdot;\beta), \text{ with } \beta \equiv (M, s, \alpha)',$$

and the parametric likelihood of each estimate $\hat{\delta}_i$ as the mixture density

$$m(\hat{\delta}_i|\beta;\sigma_i,n_i) = \int_{-\infty}^{\infty} \phi\left(\hat{\delta}_i;\delta,\sigma_i/\sqrt{n_i}\right) g(\delta,\beta)d\delta. \tag{10}$$

In the equation above $\phi(\cdot;\delta,\sigma_i/\sqrt{n_i})$ denotes the p.d.f of a normal random variable with mean δ and variance σ_i^2/n_i.

Now, we will write the likelihood for the results of n different A/B tests

$$\widehat{\boldsymbol{\delta}} = (\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_n).$$

If we assume that each estimator $\hat{\delta}_i$ is an independent draw of the model in (10), then the log-likelihood of $\widehat{\boldsymbol{\delta}}$ given the parameter β and the vector of standard errors: $\boldsymbol{\sigma} \equiv (\sigma_1/\sqrt{n_1}, \sigma_2/\sqrt{n_2}, \ldots, \sigma_n/\sqrt{n_n})$ is given by

$$\log m(\widehat{\boldsymbol{\delta}}|\beta;\boldsymbol{\sigma}) \equiv \sum_{i=1}^{n} \log m(\hat{\delta}_i|\beta;\sigma_i,n_i). \tag{11}$$

The Maximum Likelihood (ML) estimator, $\widehat{\beta}$, is the value of β that maximizes the equation above. Note that the likelihood in (11) corresponds to a model with independent, not identically distributed data. Sufficient conditions for the asymptotic normality of the ML estimator for β are given in Hoadley (1971).[28,29]

[27] The default approach for doing nonparametric estimation of g in the mixture model given by equations (6)-(7) is the infinite-dimensional Maximum Likelihood estimation routine suggested by Kiefer and Wolfowitz (1956), and refined recently by Jiang and Zhang (2009). It is known, see Theorem 2 in Koenker and Mizera (2014), that the nonparametric Maximum Likelihood estimator of g given a sample of size n is an atomic probability measure with no more than n atoms. The tails of an atomic probability measure are never fat, even if the true tails of g are. Because of this reason, we decided to follow a parametric approach for the estimation of g.

[28] The conditions in Hoadley (1971) essentially require that the first and second derivatives of the log-likelihood with respect to β to be well-defined.

[29] It is also possible to follow a fully Bayesian approach for the estimation of the parameters of the model. To do this, one could complete the Bayesian hierarchy by choosing a prior for the hyper-parameters of the model $\beta \equiv (M, s, \alpha)$. The outcomes of the A/B tests then allow us to get posterior distributions over β. Since each β is associated to a posterior mean for Δ_i (for each treatment), the following procedure provides

4.5 Estimation Results

The model fits the data well. Figure 3a displays a Q-Q plot and figure 3b compares the fitted and actual histograms for success rate. The estimated parameters are given in Table 2 and Figure 5.

Table 2: Maximum Likelihood Estimates

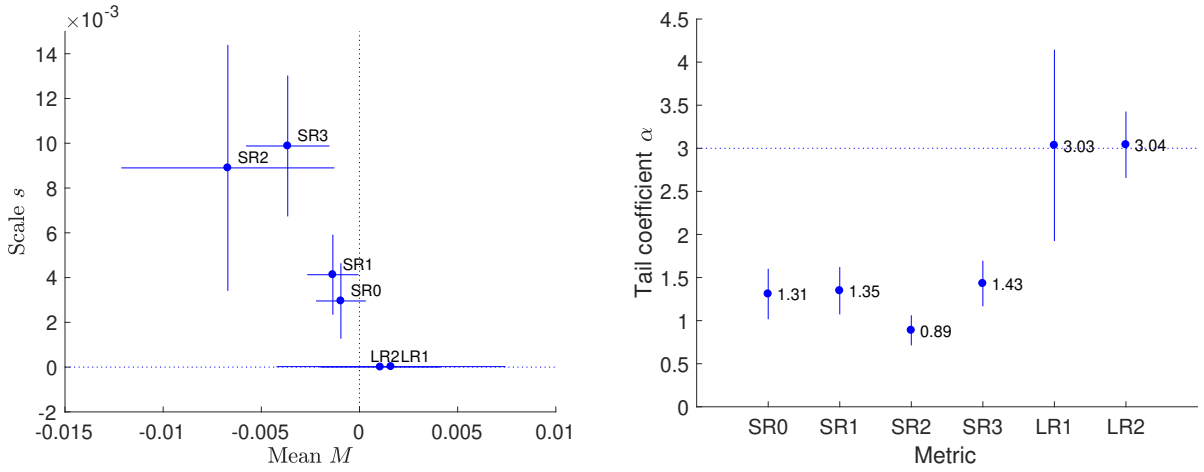
	α	M	s
Success Rate (SR0)	1.31**	-0.000946	0.00296
	(0.149)	(0.000647)	(0.000861)
Short-Run Metric #1	1.35**	-0.00136	0.00413
	(0.14)	(0.000667)	(0.000912)
Short-Run Metric #2	0.887**	-0.0067	0.0089
	(0.089)	(0.00277)	(0.0028)
Short-Run Metric #3	1.43**	-0.00365	0.00988
	(0.135)	(0.00108)	(0.00161)
Long-run metric #1	3.03	0.00161	2.21e-05
	(0.14)	(0.000667)	(0.000912)
Long-run metric #2	3.04	0.00106	2.51e-06
	(0.0916)	(0.00209)	(0.00277)

Notes: The table displays the maximum likelihood estimates of the parameters M, s, and the tail coefficient α. Standard errors are reported in parentheses. Asterisks are used to denote the magnitude of p values based on a one-sided t-tests for the hypothesis $\alpha < 3$ ($^{*}p < 1\%$ and $^{**}p < .1\%$).

Our main empirical result is that idea quality is fat tailed, with a tail coefficient far below the $\alpha = 3$ threshold in Theorem 2. The tail coefficient for success rate is of 1.31. The hypotheses that $\alpha = 3$ and $\alpha = 2$ are both rejected with a p-value of <0.001. This result is supported by three additional facts. First, the tail coefficients are similar for all short-term metrics. Second, this is consistent with findings of fat tails in data from similar large cloud products, from Facebook and Ebay experiments (Coey and Cunningham, 2019; Peysakhovich and Lada, 2016; Peysakhovich and Eckles, 2017; Goldberg and Johndrow, 2017). Third, we show in Supplementary Appendix D that the key finding of fat tails is robust to disaggregating the data across a number of dimensions, such as across budget areas. This assuages concerns that the results are driven by our estimation incorrectly pooling ideas that engineers see as being ex ante different.

There are three facts that suggest that the coefficient estimates are reasonable. First, the magnitudes are all consistent with the reduced form statistics. Second, the estimates are

a natural Bayesian estimator for the effect of a specific treatment: generate posterior draws of β, evaluate the posterior mean of Δ_i for each draw, and then average across draws.



(a) Mean and scale parameters.

(b) Tail coefficients.

Figure 5: Parameter estimates

Notes: The figure displays the maximum likelihood estimates of the parameters M and s (panel a) and the tail coefficient α (panel b). SR0 is success rate. SR1, SR2, and SR3 are the alternative short-run metrics. LR1 and LR2 are the long-run metrics. The solid lines are 95% confidence intervals.

qualitatively similar across all short-run metrics. The mean idea quality M is always a small negative number, with a similar slope as a function of the scale parameter s. Third, the placebo long-run metrics have parameters that imply that there is very little signal relative to noise. This can be seen from the low values of the scale parameter s.[30] Supplementary appendix E shows that, for long-run metrics, typical experiments should hardly update the posterior mean, and that the value of experimentation is very low. Meanwhile, the short-run metrics have qualitatively similar results as success rate. This is consistent with long-run user behavior being more difficult to change than short-run user behavior. It is also consistent with Bing engineers' view that it is hard to generate detectable movements in long-run metrics.

4.6 Implications

We now discuss three key implication of the theoretical and empirical results. We use the benchmark parameter estimates for success rate. Throughout this section, we will illustrate results with a typical experiment, which means the estimated prior, 20 million users, and the average value of σ_i, so that the standard error $\sigma_i/\sqrt{n_i}$ is 0.022. These are the parameters used in Figure 1.

[30]These estimates should be interpreted carefully because the estimated values of s for long-run metrics are close to zero, which is the boundary of the parameter space. The results in Andrews (1999) suggest that the standard errors based on the Fisher Information matrix might be conservative.

4.6.1 Implication 1: Shrinking Experimental Estimates with Small t Statistics

The fat tails imply that measured deltas with small t statistics should be shrunk aggressively, whereas measured deltas that are outliers should not be shrunk very much. The intuition is that marginally statistically significant deltas are likely to be due to a lucky experiment, whereas large outliers are likely to be real. This can be seen from the shape of the posterior mean function of a typical experiment. Figure 1 displays the posterior mean function $P_i(\hat{\delta}_i|n_i)$. A marginally significant delta equal to 0.044 (which has a t statistic of 2) has a posterior mean of only 0.006. However, an outlier experimental result of 0.088 (which has a t statistic of 4), has a posterior mean of 0.066.

This shrinkage implies that the black swan ideas correspond to a large share of the gains from innovation. To see this, consider the following "p-value" implementation strategy: to implement all ideas with positive statistically significant at the 5% level measured deltas. The p-value strategy is the most commonly used implementation strategy in practice. We can evaluate the historical gains of this policy by adding the posterior mean quality of the ideas in the data that would be implemented. We find that 74.8% of the historical gains come from the top 2% ideas. This is an extreme version of the Pareto 80/20 principle. This extreme Pareto principle arises from the combination of large outliers and the Bayesian shrinkage.

Finally the Bayesian shrinkage has implications for the optimal implementation strategy. In the typical experiment in Figure 1, the threshold for implementation is only 0.010, which corresponds to a t statistic of 0.472. The reason is that, due to the small negative prior mean, a relatively weak positive experimental result already pushes an idea into the region of positive posterior mean. This is similar to the findings of Goldberg and Johndrow (2017) in Ebay data. The optimal policy generates a historical gain in our data of 2.3% improvement in success rate. This is 28.16% more than the gain of the p-value policy. We note that practitioners use a more strict implementation threshold because there are costs of implementing each feature, and of making the codebase more complex. In the typical experiment of Figure 1, the p-value policy is optimal if the implementation cost equals a 0.0055% gain in success rate.

4.6.2 Implication 2: Gains from Lean Experimentation

The estimated tail coefficient α of 1.31 is well below the theoretical threshold $\alpha = 3$ of Theorem 2. Consistent with the theorem, the estimated production function has decreasing returns to scale close to 0 (figure 2a). This suggests that, in Bing's setting, there are considerable gains in moving towards a lean experimentation strategy.

To understand the value of moving towards lean experimentation, consider the following numerical example. A firm tests I innovations in a total of N users. Innovations are

homogeneous, and the firm splits users equally across innovations, so that there are $n = N/I$ users in each experiment. Total production Y is then

$$Y = I \cdot f\left(\frac{N}{I}\right) = I \cdot f(n). \tag{12}$$

We begin by computing the gain of testing a larger number I of ideas, keeping the total amount of data N fixed. We continue to use the benchmark empirical estimates for G_i and n equal to 20 million users. The computations show that the total production Y grows almost linearly with the number of ideas tested. An increase in the number of ideas by 10% increases total production by 8.59%, and an increase in the number of ideas by 20% increases total production by 17.05%.

This computation assumes that there is a costless way to increase the number of ideas and that marginal ideas are just as productive as the current ideas. Therefore, a key question is whether there are practical ways to move towards lean experimentation. That is, whether there are additional ideas to be tested that have much greater benefits than costs.

In the case of Bing, one practical way to increase the number of ideas would be to reduce offline tests and triage procedures. Triage procedures take place before online A/B tests. In Section 5.3, we describe some of the current triage procedures, and use the best available data to evaluate them. The data shows that a substantial number of ideas seem to be eliminated in the offline tests. Moreover, we tried to measure the quality of the marginal ideas, by checking how offline tests predict online performance. We find no evidence that marginal ideas are worse. This indicates that there may be gains moving towards lean experimentation.

A potential objection to this point is that there might be high fixed costs of experimentation, so that it would not be worth it to run smaller experiments. We can use a simple back-of-the-envelope calculation to show that this would require implausibly large costs. For this analysis, we have to consider the value of an experiment in dollars. While our analysis is based on success rate, Microsoft uses a proprietary "monetary overall evaluation criterium," which is a conversion rate between metrics and revenue. This conversion rate is used to make decisions that involve tradeoffs between key metrics, such as a change that increases ad revenue but hurts user experience. This conversion rate is proprietary, so that we cannot use it in our paper to convert the results to dollars. However, we can use a ballpark value to understand the value of having a typical idea and testing it in a typical experiment. The value of a typical idea in an experiment with 20 million users is $f(20e6)$, or a 2.39e-03% gain in success rate. A 1% gain in success rate is valued in the order of hundreds of millions of dollars of yearly revenue. Therefore, the value of testing an additional marginal idea is in the order of a million dollars of yearly revenue increase. The ideas that are evaluated in offline tests are mostly coded. Therefore, the additional

costs of A/B testing them are likely small. It is implausible that these costs are greater than the estimated benefit, which is of the order of a million dollars of yearly revenue. In particular, this analysis suggests that the gains from moving towards lean experimentation are economically significant. See Section F.1 for a thorough discussion of alternative costs of experimentation and other caveats.

4.6.3 Implication 3: The Marginal Value of Data is Economically Significant

It is sometimes argued that, in large platforms, sample sizes are so large that the marginal value of data is close to zero. For example, in the early days of A/B testing many industry experts argued that online sample sizes were "in the millions" so that it is not necessary to use statistics (Kohavi et al., 2009b). More recently, there are claims that parallelized experimentation makes sample sizes so large that the marginal value of data is insignificant.

We can use our estimates to evaluate the merit of this view in the Bing setting. We consider the same numerical example with total production being given by equation 12. We consider the gains from increasing n by 10%, holding fixed the number of ideas. We find that this results in an increase in Y of 1.29% of the total. This increase is much smaller than 10%, due to the decreasing returns, but still economically significant.

We can gain some intuition into why data has a nontrivial marginal value from the production function. The estimated production function at an n of 20 million users is far from its maximum value, so that getting more data is still valuable. Moreover, consider the theorem 2 approximation that $f(n)$ is a constant times $n^{\frac{\alpha-1}{2}}$. This suggests that the marginal product of the data $f'(n)$ is about $\frac{\alpha-1}{2}$ times the average product $f(n)/n$. This is close to what we found when increasing n by 10%. That is, the empirical estimates suggest that the marginal value of data is an order of magnitude smaller than the average value. At the same time, the estimates are not consistent with the view that the marginal value of data is negligible.

5 Additional Results and Robustness

5.1 Relevance of the Small-n and Large-n Asymptotics

We now examine the range where the asymptotic formulas in Theorems 1 and 2 provide a relevant approximation to the value of experimentation. Figure 6 plots the production function for the Bing application (based on the parameter estimates for session success rate) along with the approximations. We plot n in a log scale, so that we cover a broad range of sample sizes.

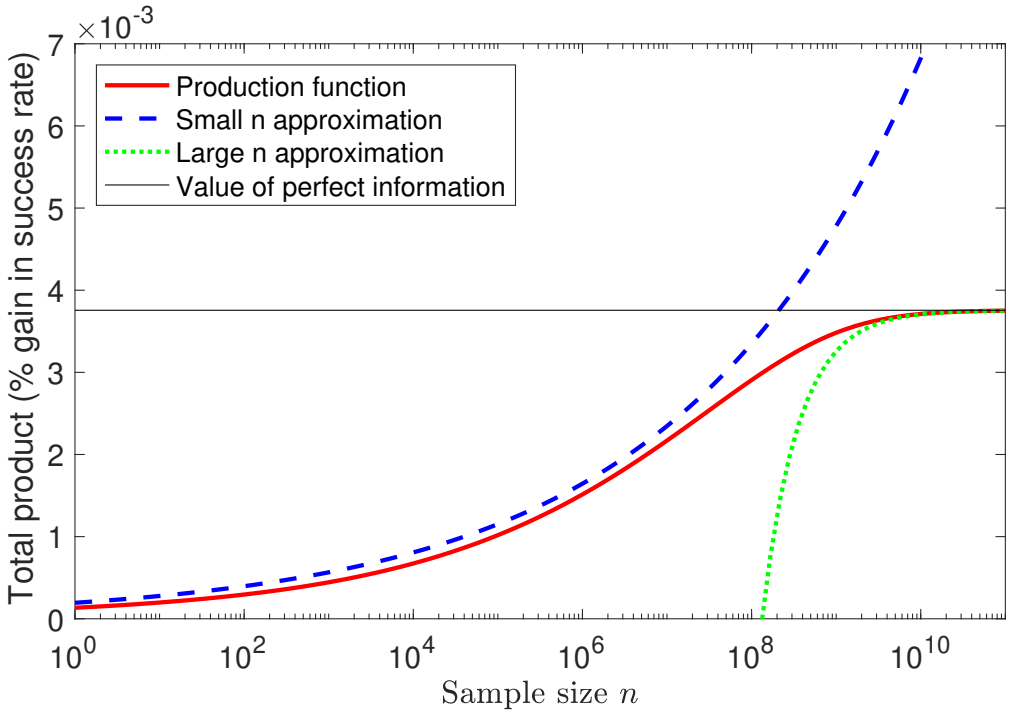


Figure 6: Empirically relevant ranges for the asymptotic approximations.

Notes: The figure depicts the production function, the large n approximation from Theorem 1, and the small n approximation from Theorem 2 under the benchmark parameter estimates. For large n, we integrate Theorem 1's formula for f' to obtain the approximation $f(n) \approx f(\infty) - 1/2g(0)\sigma^2 n^{-1}$. For small n, we approximated $t^*(n)$ by Theorem 2's asymptotic formula at $n = 1$, so that $t^*(n) \approx \sqrt{2(\alpha-1)\log\sigma}$. We took $c(\delta^*(n))$ to be the relevant constant for the t-distribution. With this further approximation, the only term in the formula for $f'(n)$ that depends on n is $n^{\alpha-3/2}$. Integration then yields $f(n) \approx \frac{1}{\alpha-1}\alpha c(\sigma t^*)^{-(\alpha-1)} n^{\frac{\alpha-1}{2}}$.

The large-n approximation in Theorem 1 starts to be useful in experiments of the order of hundreds of millions of users. It becomes a very good approximation close to one billion users. This is also the point where the experiment is precise enough to attain most of the value of perfect information, so that the production function starts to converge to this value. In principle, it seems surprising that the asymptotic approximation only becomes good at such large values, which are much greater than in realistic experiments. The reason is that the large-n approximation is based on the fact that, asymptotically, the experiment becomes so informative that we can basically ignore the prior, as in the Bernstein-von Mises theorem. However, large-scale A/B testing aims to detect small performance improvements with noisy outcome data. This means that, for realistic sample sizes, there is still considerable uncertainty, and the prior should not be ignored.

The small-n approximation in Theorem 2 gives useful estimates for experiments with at most tens of millions, or maybe one hundred million users. For larger experiments, the approximation is far off, as the approximation converges to infinity while the production function is bounded by the value of perfect information. Moreover, Figure 2 shows that the small-n approximation gives the correct qualitative shape of the production function

for the relevant sample sizes of up to tens of millions. And, moreover, the approximation correctly matches the key economic property of returns to scale and the optimal experimentation strategy. At first, it seems counter intuitive that a small-n approximation is useful for experiments with millions of users. The reason is that in the small-n approximation outlier signals $\hat{\delta}_i$ are responsible for most of the gains. In practice, mediocre signal draws still have some payoff importance, which is why the approximation is not exact for realistic experiments. Nevertheless, as shown by our estimates, outliers are still extremely important at the practically relevant sample sizes. This why the small-n approximation is relevant in realistic experiment sizes, and why the tail coefficient is a useful statistic for optimal experimentation.

5.2 Theoretical Extensions

This section considers several extensions of the baseline A/B testing problem. For the sake of exposition, we only provide a brief description of the main findings of each extension. The details are provided in Section F of the Supplementary Materials.

1. *Other costs of experimentation*. We consider three additional costs: fixed costs of testing an idea; variable costs of experimentation (as a function of the sample size); and short-term user experience costs, in which each idea i has a benefit proportional to $\Delta_i \cdot n_i$.[31] There are three main lessons derived from a model that allows for costs of experimentation. First, our production function approach is still useful to understand the value of experimentation. Second, lean experimentation need not be optimal in the presence of fixed costs because of standard arguments: if fixed costs are large, then the benefit of running a small experiment may be not be enough to cover the fixed costs. Third, fat tails are still important for determining the optimal experimentation strategy even in the presence of fixed costs. In fact, the tail coefficient of the distribution of unobserved idea quality affects how different characteristics of ideas should be evaluated.[32]

2. *Mutually exclusive ideas*. We consider a variation of the A/B testing problem where the firm can implement at most one of the I different ideas after observing the experimentation results. Thus, the payoff from implementing multiple innovations is no longer additive. This fits examples like a firm deciding between five alternative designs for a website. The main message from this variation is that the results in

[31] In the Bing example, this cost corresponds to how much the experimental platform hurts user experience.

[32] For example, with a t-prior and abundant data greater spread (s) is much more valuable the thicker the tails are. This is intuitive because, in the fat-tailed case, a larger fraction of the gains of experimentation comes from outliers.

Corollary 1 are still true: if the tails are thick enough ($\alpha > 3$), it is optimal to run experiments on all ideas even if only one can be implemented.

3. *Hypothesis-testing payoff*. We assume that if an innovation is implemented, its payoff is
$$K\mathbf{1}\left(\Delta_i > 0\right) - \mathbf{1}\left(\Delta_i \leq 0\right),$$
as opposed to the linear and unbounded payoff Δ_i. The main implication of this model is that the threshold for determining the optimality of lean experimentation becomes $\alpha = 2$. This can be derived using the same argument we used to establish Theorem 2. An intuitive derivation can be obtained using the heurisitc argument provided in p. 14.

4. *Elastic Supply of Ideas*. This variation of the model assumes that the firm has an infinitely elastic supply of identical ideas at a fixed cost per idea. We have two main results. First, under some mild conditions, the optimal scale of each experiment has to maximize average product net of fixed costs. Importantly, the optimal scale does not depend on the total number of users. Therefore, it is not optimal to grow the size of each experiment without bound as more data is available. Instead it is better to increase the number of ideas being tested. This result gives another version of the optimality of lean experimentation with a very large amount of users. Second, we consider priors of the form $\Delta = M + s\Delta_0$ and show that, if the prior is very uninformative, then greater spread (higher s) leads to leaner experimentation.

5. *A fluke model for the experimental noise.* Suppose that with probability w the signal $\hat{\delta}_i|\delta_i$ is $\mathcal{N}(0, \sigma_n^2)$ and with probability $1-w$, $\hat{\delta}_i$ has as a p.d.f $p(\cdot)$ that does not depend on δ_i. In this model the posterior distribution of δ is a convex combination of the posterior density in the model without flukes and the prior. This happens because the fluke density is not informative about the true idea quality (thus, with some probability, the prior is not updated). We want to consider the case in which the fluke density $p(\hat{\delta}_i)$ has fat-tails. We show that our small-n results will hold as long as tails of the fluke distribution are thinner than those of the prior, but the value of small experiments will be close to zero in other cases.

5.3 Quality of Marginal Ideas

Our results suggest that Bing could gain by moving towards a lean experimentation approach. One direction for improvement would be to reduce offline triage procedures, which now are used to eliminate a substantial number of ideas before they make it to an A/B test. However, this conclusion depends on the quality of the marginal ideas that are being cut in the offline procedures. It is possible the the offline triage accurately eliminates

ideas that have low quality, and that these procedures are optimal. In this Supplementary Appendix C we examine limited data on the offline triage procedure to understand whether this is the case. We caution that these data are recorded in an incomplete and potentially biased way, and that there are few observations. The analysis should be interpreted cautiously due to the data limitations.

We hand-collected data on offline triage procedures conducted by a major development team within Bing. The procedure works as follows. In an unobserved phase 0, engineers turn ideas into fully coded techniques. Candidate techniques are evaluated offline with a crowdsourcing tool similar to Amazon's Mechanical Turk. If these results look promising, engineers can make a submission to the formal phase 1 review panel. The review panel decides which techniques to move to phase 2, with the guidelines being that flat or positive offline metrics should pass in usual cases. Phase 2 is an online A/B test. The development team keeps records of a subset of the phase 1 submissions. Unfortunately, this is far from a complete record because most ideas in the dataset were ultimately implemented. Thus, this data is incomplete and biased towards successful ideas. We have 33 observations, out of which 18 were implemented. For each observation, we have the results of four offline metrics, the decision to go to phase 2, results of online A/B tests when available, and the decision of whether to implement the idea.

Supplementary Appendix C documents two main patterns. First, the data is consistent with a substantial number of ideas being turned down in triage. The data has some evidence that is consistent with the review panel roughly following its guidelines.

Second, there is no evidence that these offline tests are predictive of online performance. Under the hypothesis that the offline tests are successful in screening promising ideas, we expect to find that the offline results are highly predictive of performance in online A/B tests. In fact, the offline tests have little to no predictive power of the results of online A/B tests. For example, Figure 7 plots the change in session success rate in an online A/B test versus each standardized offline metric. The figure suggests that there is almost no correlation between the offline and online results. This is confirmed by a series of alternative specifications reported in Supplementary Appendix C. In particular, the offline metrics seem to have almost no correlation with each other, even though some of them are supposed to be alternative measures of similar aspects of performance.

The results in this subsection should be taken with a grain of salt due to the data limitations. However, the results do show that there is no strong evidence that offline metrics are highly predictive of online results, nor is there strong evidence that the marginal ideas that are being discarded are of lower quality than the ideas that make it to online A/B tests.

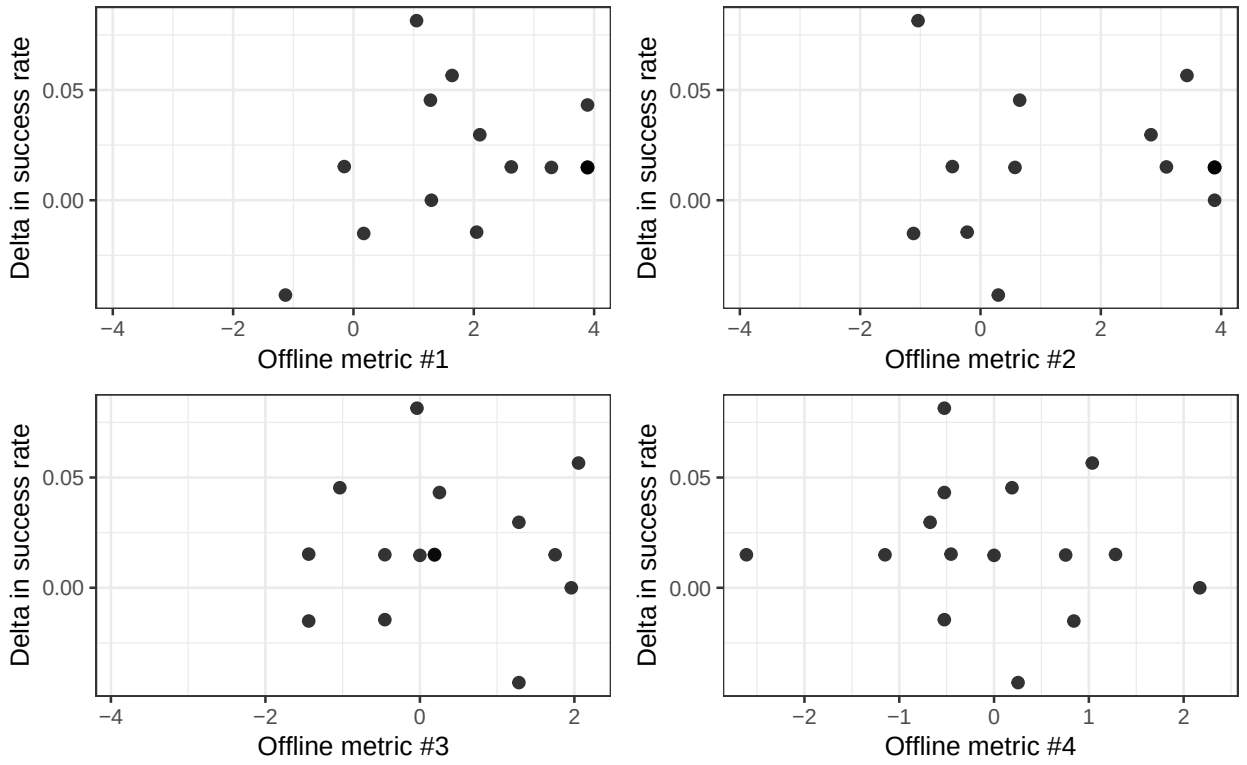


Figure 7: Performance in online A/B tests versus performance in offline tests.

Notes: The vertical axis plots the measured delta in success rate. The horizontal axes plot performance in each of the offline metrics. The horizontal axes are standardized so that the standard error of the offline experiments is 1.

5.4 Robustness Checks

Priors with bounded support.

Theorem 2 assumes a prior with unbounded support. In principle, this is at odds with our empirical application because the success rate metric is bounded between 0 and 1. In practice, this is not an issue because the prior is concentrated around 0, so that the t-distribution describes the data well in the relevant range. The probability of Δ taking a value that is not feasible is negligible. For example, under the benchmark estimates, the probability of an increase in success rate of more than 10% is 8.35e-06. For this reason, our numerical results are virtually unchanged if we replace g by a distribution that is truncated at -10% and 10%.

Heterogeneous priors and biases in tail coefficient estimates.

One potential bias in our empirical analysis is that we assumed that all ideas come from the same prior idea distribution. This assumption is partly justified because the dataset was constructed to include relatively homogeneous ideas, and that engineers consider to be ex ante similar. However, if engineers see these ideas as different, we could be incorrectly inferring that the distribution of quality is fat tailed.

We examined empirically whether our estimates are robust to this concern. To do so, we estimated the prior distribution of ideas for different subsets of the data. We split the data into different groups based on areas of experimentation, time period, experiment length, and sample size. The results of these disaggregated analyses is reported in Supplementary Appendix D. For all of these subgroups, we found low tail coefficients, in the ballpark of the coefficients in our main specification. This lends support to our central estimates not suffering from large biases due to heterogeneity in priors.

Sequential testing.

A pontential concern is that our estimates could be biased due to engineers using dynamic experimentation strategies. The most obvious concern would be if there is p-hacking, where engineers stop experiments as soon as they find a statistically possible result. Indeed, Berman et al. (2018) find evidence of p-hacking for businesses using off-the-shelf A/B testing software. However, large experimental platforms like EXP employ several statisticians, and have policies in place to eliminate p-hacking. For example, one of the rules is that experiments are usually run in multiples of one week.[33] To formally test whether our finding of fat tails is an artifact of p-hacking, Supplementary Appendix D estimates our model separately for experiments that last exactly one week (which is the most modal case), experiments that are run for longer. Both subsamples display fat tails with coefficients in the same ballpark. Our main result also holds for experiments that are run for exact multiples of a week, where sequential testing is less likely to be a concern.

6 Conclusion

A/B tests have risen in prominence with increased availability of data and with lower costs of experimentation. We considered a simple optimal learning model to understand how to use scarce experimental resources in this setting. Crucially, we contribute to the literature on optimal learning by allowing for the presence of fat tails of innovation quality.

The results suggest that the presence of fat tails are an important determinant of the optimal innovation strategy. In contexts with a thin-tailed distribution of innovation quality, it is desirable to perform thorough prior screening of potential innovations, and to run a few high-powered precise experiments. In the technology industry, this corresponds to rigorously screening innovation ideas prior to A/B tests. In research on anti-poverty programs, it corresponds to trying out only a few ideas with few but high-quality, high-powered research studies. In contexts with a fat-tailed distribution of innovation quality,

[33] The rule is not enforced perfectly, but whenever an experiment run for, say, 10 days, its final scorecard is compared with the scorecard at the end the first week to detect anomalies. In describing experimentation at Microsoft, (Kohavi et al., 2013) write: "While we allow experimenters to look at daily results, as they lead to insights and could help identify bugs early on, there is one final scorecard at the end of the experiment, which we require to be a multiple of weeks, usually two weeks."

it is advantageous to run many small experiments, and to test a large number of ideas in hopes of finding a big winner. In the technology industry, this corresponds to doing little to no screening of ideas prior to A/B tests, and to run many experiments even if this sacrifices sample sizes. In research on anti-poverty programs, it corresponds to trying out many ideas, even if particular studies have lower quality and statistical power, in hopes of finding one of the rare big winners.

We applied our model to detailed data on the experiments conducted in a major cloud software product, the Bing search engine. We find that the distribution of innovations is fat-tailed. This implies that lean innovation strategies are optimal. This suggests that large performance gains are possible in our empirical context. These gains are substantial in dollar terms. And there is suggestive evidence that some of these gains can be realized with simple changes such as reducing triage processes, and using Bayesian methods to evaluate innovations.

We stress that our results on Bing should not be taken as externally valid for all contexts. While it is plausible that these results extend to other similar products, it is quite possible that the distribution of innovations is different in different contexts. However, the Bing application illustrates that it is possible to achieve large gains by understanding the optimal innovation strategy, even in a setting that already uses cutting-edge experimentation techniques. It would be interesting to extend this analysis to other contexts, to try to increase the speed of innovation, especially in areas of high social value.

References

Andrews, Donald WK, "Estimation when a parameter is on a boundary," *Econometrica*, 1999, *67* (6), 1341–1383.

Arrow, Kenneth J., David Blackwell, and Meyer A. Girshick, "Bayes and minimax solutions of sequential decision problems," *Econometrica, Journal of the Econometric Society*, 1949, pp. 213–244.

Athey, Susan and Guido W. Imbens, "The Econometrics of Randomized Experiments," in "Handbook of Economic Field Experiments," Vol. 1, Elsevier, 2017, pp. 73–140.

Azevedo, Eduardo M., Alex Deng, José Luis Montiel Olea, and Glen E. Weyl, "Empirical Bayes Estimation of Treatment Effects with Many A/B Tests: An Overview.," *AEA Papers and Proceedings*, May 2019, (109), 43–47.

Banerjee, Abhijit, Sylvain Chassang, Sergio Montero, and Erik Snowberg, "A theory of experimenters," Technical Report, National Bureau of Economic Research 2017.

Bergemann, D and J Valimaki, "Bandit problems," in "The New Palgrave Dictionary of Economics," 2nd edition ed., Macmillan Press, 2008.

Berman, Ron, Leonid Pekelis, Aisling Scott, and Christophe Van den Bulte, "p-Hacking and False Discovery in A/B Testing," 2018.

Billingsley, P., *Probability and Measure*, 3rd ed., John Wiley & Sons, New York, 1995.

Blank, Steve, "Why the Lean Start-Up Changes Everything," *Harvard Business Review*, 2013, *91* (5), 64–68.

Bubeck, Sébastien and Nicolo Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, 2012, *5* (1), 1–122.

_ , _ , and Gábor Lugosi, "Bandits with heavy tail," *IEEE Transactions on Information Theory*, 2013, *59* (11), 7711–7717.

Carlin, B.P. and T.A. Louis, *Bayes and empirical Bayes methods for data analysis* number 2. In 'Texts in Statistical Science.', second edition ed., Chapman & Hall, 2000.

Chade, Hector and Edward Schlee, "Another look at the Radner–Stiglitz nonconcavity in the value of information," *Journal of Economic Theory*, 2002, *107* (2), 421–452.

Che, Yeon-Koo and Konrad Mierendorff, "Optimal sequential decision with limited attention," *unpublished, Columbia University*, 2016.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman, "Power-law distributions in empirical data," *SIAM review*, 2009, *51* (4), 661–703.

Coey, Dominic and Tom Cunningham, "Improving Treatment Effect Estimators Through Experiment Splitting," *The Web Conference*, 2019.

Deaton, Angus, "Instruments, randomization, and learning about development," *Journal of economic literature*, 2010, *48* (2), 424–55.

Deng, Alex, Ya Xu, Ron Kohavi, and Toby Walker, "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data," in "Proceedings of the sixth ACM international conference on Web search and data mining" ACM 2013, pp. 123–132.

Diggle, Peter J and Peter Hall, "A Fourier approach to nonparametric deconvolution of a density estimate," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1993, pp. 523–531.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using randomization in development economics research: A toolkit," *Handbook of development economics*, 2007, *4*, 3895–3962.

Efron, Bradley, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, 2011, *106* (496), 1602–1614.

Feit, Elea McDonnell and Ron Berman, "Profit-Maximizing A/B Tests," *Available at SSRN 3274875*, 2018.

Feller, W., *An Introduction to Probability Theory and Its Applications, Vol. 2*, Vol. 2 1967.

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki, "Stochastic Choice and Optimal Sequential Sampling," 2017. https://ssrn.com/abstract=2602927.

Goldberg, David and James E Johndrow, "A Decision Theoretic Approach to A/B Testing," *arXiv preprint arXiv:1710.03410*, 2017.

Hoadley, Bruce, "Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case," *The Annals of mathematical statistics*, 1971, pp. 1977–1991.

Imbens, Guido W., "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic literature*, 2010, *48* (2), 399–423.

Jiang, Wenhua and Cun-Hui Zhang, "General maximum likelihood empirical Bayes estimation of normal means," *The Annals of Statistics*, 2009, *37* (4), 1647–1684.

Johnson, Eric J and Daniel Goldstein, "Do defaults save lives?," 2003.

Karamata, Jovan, "Some theorems concerning slowly varying functions," 1962.

Kasey, Maximilian and Anja Sautmann, "Adaptive Experiments for Policy Choice," *Mimeo, Harvard University*, 2019.

Keppo, Jussi, Giuseppe Moscarini, and Lones Smith, "The demand for information: More heat than light," *Journal of Economic Theory*, 2008, *138* (1), 21–50.

Kiefer, Jack and Jacob Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *The Annals of Mathematical Statistics*, 1956, pp. 887–906.

Koenker, Roger and Ivan Mizera, "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules," *Journal of the American Statistical Association*, 2014, *109* (506), 674–685.

Kohavi, Ron, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann, "Online controlled experiments at large scale," in "Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining" ACM 2013, pp. 1168–1176.

_ **and Roger Longbotham**, "Unexpected results in online controlled experiments," *ACM SIGKDD Explorations Newsletter*, 2011, *12* (2), 31–35.

_ , _ , **Dan Sommerfield, and Randal M. Henne**, "Controlled experiments on the web: survey and practical guide," *Data mining and knowledge discovery*, 2009, *18* (1), 140–181.

Kohavi, Ronny, Thomas Crook, Roger Longbotham, Brian Frasca, Randy Henne, Juan Lavista Ferres, and Tamir Melamed, "Online experimentation at Microsoft," *Data Mining Case Studies*, 2009, *11.*

Li, Tong and Quang Vuong, "Nonparametric estimation of the measurement error model using multiple indicators," *Journal of Multivariate Analysis*, 1998, *65* (2), 139–165.

Madrian, Brigitte C and Dennis F Shea, "The power of suggestion: Inertia in 401 (k) participation and savings behavior," *The Quarterly journal of economics*, 2001, *116* (4), 1149–1187.

McClellan, Andrew, "Experimentation and Approval Mechanisms," *Mimeo, Chicago Booth*, 2019.

Morris, Stephen and Muhamet Yildiz, "Crises: Equilibrium Shifts and Large Shocks," 2016.

_ **and Philipp Strack**, "The Wald problem and the equivalence of sequential sampling and static information costs," 2017.

Moscarini, Giuseppe and Lones Smith, "The optimal level of experimentation," *Econometrica*, 2001, *69* (6), 1629–1644.

_ **and** _ , "The law of large demand for information," *Econometrica*, 2002, *70* (6), 2351–2366.

Peysakhovich, Alexander and Akos Lada, "Combining observational and experimental data to find heterogeneous treatment effects," *arXiv preprint arXiv:1611.02385*, 2016.

_ **and Dean Eckles**, "Learning causal effects from many randomized experiments using regularized instrumental variables," *arXiv preprint arXiv:1701.01140*, 2017.

Pickands, James III, "Statistical inference using extreme order statistics," *Annals of Statistics*, 1975, (3), 119–131.

Radner, Roy and Joseph E. Stiglitz, "A Nonconcavity in the Value of Information," in Marcel Boyer and Richard Kihlstrom, eds., *Bayesian Models in Economic Theory*, Amsterdam: Elsevier Science, 1984, chapter 3, pp. 33–52.

Redner, S., "How popular is your paper? An empirical study of the citation distribution," *The European Physical Journal B - Condensed Matter and Complex Systems*, Jul 1998, *4* (2), 131–134.

Ries, Eric, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*, New York: Crown Business, 2011.

Robbins, Herbert, "The Empirical Bayes approach to statistical decision problems," *The Annals of Mathematical Statistics*, 1964, pp. 1–20.

_ , "Some aspects of the sequential design of experiments," in "Herbert Robbins Selected Papers," Springer, 1985, pp. 169–177.

Seidel, Wilfried, "Mixture models," *Encyclopedia of Mathematics, http://www.encyclopediaofmath.org/index.php?title=Mixture_models& oldid=37767*, 2015.

Silverberg, Gerald and Bart Verspagen, "The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance," *Journal of Econometrics*, 2007, *139* (2), 318–339.

Small, Christopher G, *Expansions and asymptotics for statistics*, CRC Press, 2010.

Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer, "Overlapping experiment infrastructure: More, better, faster experimentation," in "Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining" ACM 2010, pp. 17–26.

Thompson, William R., "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, 1933, *25* (3/4), 285–294.

Vul, Edward, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum, "One and Done? Optimal Decisions From Very Few Samples," *Cognitive Science*, 2014, *38* (4), 599—637.

Wald, Abraham, "Foundations of a general theory of sequential decision functions," *Econometrica, Journal of the Econometric Society*, 1947, pp. 279–313.

Whitt, Ward, *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*, Springer, 2002.

A Proofs

A.1 Notation

Denote the normal cumulative distribution with mean μ and variance σ^2 as $\Phi(\cdot|\mu, \sigma^2)$ and density as $\phi(\cdot|\mu, \sigma^2)$. Denote the standard normal cumulative distribution as $\Phi(\cdot)$ and density as $\phi(\cdot)$. The density of the signal $\hat{\delta}_i$ conditional on true quality δ_i is $\phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n_i)$. Therefore, the *likelihood* of δ_i and $\hat{\delta}_i$ is $\phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n_i) \cdot g_i(\delta_i)$. The *marginal distribution of the signal* $\hat{\delta}_i$ is

$$m_i(\hat{\delta}_i, n_i) \equiv \int_{-\infty}^{\infty} \phi\left(\hat{\delta}_i \middle| \delta_i, \frac{\sigma_i^2}{n_i}\right) \cdot g_i(\delta_i)\, d\delta_i. \tag{A.1}$$

By Bayes' rule, the *posterior density* of δ_i given signal $\hat{\delta}_i$ is

$$g_i(\delta_i|\hat{\delta}_i, n_i) = \frac{\phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n) \cdot g_i(\delta_i)}{m_i(\hat{\delta}_i, n_i)}.$$

The posterior mean is

$$P_i(\hat{\delta}_i, n_i) = \int_{-\infty}^{\infty} \delta_i \cdot g_i(\delta_i|\hat{\delta}_i, n_i)\, d\delta_i = \frac{\int \delta_i \cdot \phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n) \cdot g_i(\delta_i)\, d\delta_i}{m_i(\hat{\delta}_i, n_i)}. \tag{A.2}$$

A.2 Basic Results

Lemma A.1 (Regularity Properties). *For $n_i > 0$, the marginal density $m_i(\hat{\delta}_i, n_i)$ and the posterior mean $P_i(\hat{\delta}_i, n_i)$ are smooth in both variables. The posterior mean strictly increasing in $\hat{\delta}_i$, and there exists a unique threshold signal $\delta_i^*(n_i)$ such that the posterior mean given n_i and the signal equals zero.*

Proof. By equation (A.1) and Leibniz's rule, m_i is smooth and strictly positive. Efron's equation (2.8) then implies that P_i is smooth. Efron (2011) p. 1604 shows that P_i is trictly increasing. Because of the strict monotonicity of P_i, to show that there exists a unique threshold $\delta_i^*(n_i)$, it is sufficient to show that the posterior mean is positive for a sufficiently large positive signal and negative for a sufficiently large negative signal. Consider the case of a large positive signal $\hat{\delta}_i > 1$. Because $g_i(0) > 0$, there exists δ_0 with $0 < \delta_0 < 1$ and

$g_i(\delta_0) > 0$. The numerator in the posterior mean formula (A.2) is bounded below by

$$\begin{aligned}
&\int_{-\infty}^{0} \delta_i \cdot \phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n) \cdot g_i(\delta_i)\, d\delta_i \\
&+ \int_{\delta_0}^{1} \delta_i \cdot \phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n) \cdot g_i(\delta_i)\, d\delta_i \\
\geq &\phi(\hat{\delta}_i|0, \sigma_i^2/n) \cdot \int_{-\infty}^{0} \delta_i \cdot g_i(\delta_i)\, d\delta_i \\
&+ \phi(\hat{\delta}_i|\delta_0, \sigma_i^2/n) \cdot \int_{\delta_0}^{1} \delta_i \cdot g_i(\delta_i)\, d\delta_i.
\end{aligned}$$

The fact that $g_i(\delta_0) > 0$ implies that the second integral is strictly positive. Moreover, as $\hat{\delta}$ converges to infinity, the ratio

$$\frac{\phi(\hat{\delta}_i|\delta_0, \sigma_i^2/n)}{\phi(\hat{\delta}_i|0, \sigma_i^2/n)}$$

converges to infinity, so that the posterior mean is positive. The case of a large negative signal is analogous. □

Proof of Proposition 1. The expected payoff of experimentation strategy $\boldsymbol{n}$ and implementation strategy S is given by equation (1). By the law of iterated expectations,

$$\begin{aligned}
\Pi(\boldsymbol{n}, S) &= \mathbb{E}\left(\mathbb{E}\left(\sum_{i\in S} \Delta_i \middle| \hat{\boldsymbol{\Delta}}\right)\right) \\
&= \mathbb{E}\left(\sum_{i\in S} P_i(\hat{\Delta}_i, n_i)\right).
\end{aligned}$$

This implies that, conditional on the signals, it is optimal to implement all innovations with strictly positive posterior mean, and not to implement innovations with strictly negative posterior mean. Moreover, any innovation strategy that does not do so with positive probability is strictly suboptimal, establishing the proposition. □

Proof of Proposition 2. The decomposition of the expected payoff follows from the argument in the body of the paper. The smoothness of the production function follows from equation (2) and from the smoothness of the marginal density of the signal and the posterior mean established in lemma A.1. □

A.3 Proof of the Main Theorems

Throughout this section, we omit dependence on the innovation i because the results apply to the production function for a single innovation. To avoid notational clutter, we use subscripts to denote the sample size n, as in δ_n^* and t_n^*. We denote the standard error of the experiment as $\sigma_n = \sigma/\sqrt{n}$.

We now give a formula for the marginal product, which is used in the proof of the main theorems.

Lemma A.2 (Marginal Product Formula)**.** *The marginal product equals*

$$f'(n) = \frac{1}{2n} \cdot m(\delta_n^*, n) \cdot \mathrm{Var}[\Delta | \hat{\Delta} = \delta_n^*, n]. \tag{A.3}$$

Proof. The total value of an innovation combined with data n_i equals the expectation of the value of the innovation times the probability that it is implemented. Moreover, the innovation is implemented iff the signal is above the optimally selected threshold. Therefore,

$$\begin{aligned} f(n) &= \max_{\bar{\delta}} \int \delta \cdot \Pr\{\hat{\Delta} \geq \bar{\delta} | \Delta = \delta, n\} \cdot g(\delta)\, d\delta - \mathbb{E}[\Delta]^+ \\ &= \max_{\bar{\delta}} \int \delta \cdot \Phi\left(\frac{\delta - \bar{\delta}}{\sigma_n}\right) \cdot g(\delta)\, d\delta - \mathbb{E}[\Delta]^+. \end{aligned}$$

And this expression is maximized at $\bar{\delta} = \delta_n^*$ by Proposition 1. The maximand is a smooth function of $\bar{\delta}$ and n. Therefore, by the envelope theorem and Leibniz's rule,

$$f'(n) = \int \delta \cdot \left[\frac{d}{dn}\Phi\left(\frac{\delta - \bar{\delta}}{\sigma_n}\right)\right] \cdot g(\delta)\, d\delta \Bigg|_{\bar{\delta} = \delta_n^*}.$$

Taking the derivative,

$$\begin{aligned} f'(n) &= \frac{1}{2\sqrt{n}} \int \delta \cdot (\delta - \delta_n^*) \cdot \frac{1}{\sigma} \cdot \varphi\left(\frac{\delta - \delta_n^*}{\sigma_n}\right) \cdot g(\delta)\, d\delta \\ &= \frac{1}{2n} \cdot \int \delta \cdot (\delta - \delta_n^*) \cdot \varphi\left(\delta_n^* | \delta, \sigma_n^2\right) \cdot g(\delta)\, d\delta \\ &= \frac{1}{2n} \cdot m(\delta_n^*, n) \cdot \int \delta \cdot (\delta - \delta_n^*) \cdot g(\delta | \delta_n^*, n)\, d\delta. \end{aligned}$$

Writing the integrals as conditional expectations we have

$$f'(n) = \frac{1}{2n} \cdot m(\delta_n^*, n) \cdot \left(\mathbb{E}[\Delta^2 | \hat{\Delta} = \delta_n^*, n] - \delta_n^* \mathbb{E}[\Delta | \hat{\Delta} = \delta_n^*, n]\right).$$

The result then follows because $\mathbb{E}[\Delta|\hat{\Delta} = \delta_n^*, n] = 0$ at the optimal threshold δ_n^*.

□

A.3.1 Proof of Theorem 1

Part 1: Preliminary Results We will use a standard result from Bayesian statistics, known as Tweedie's formula, which holds because of the normally distributed experimental noise. Tweedie's formula expresses the conditional mean and variance of quality using the marginal distribution of the signal.

Proposition A.1 (Tweedie's Formula)**.** *The posterior mean and variance of Δ conditional on a signal $\hat{\delta}$ and $n > 0$ are*

$$P(\hat{\delta}, n) = \hat{\delta} + \sigma_n^2 \frac{d}{d\hat{\delta}} \log m(\hat{\delta}, n) \tag{A.4}$$

and

$$\text{Var}[\Delta|\hat{\Delta} = \hat{\delta}, n] = \sigma_n^2 + \sigma_n^4 \cdot \frac{d^2}{d\hat{\delta}^2} \log m(\hat{\delta}, n).$$

Proof. See Efron (2011) p.1604 for a proof and his equation (2.8) for the formulas. □

The next lemma allows us to apply Tweedie's formula to obtain our asymptotic results.

Lemma A.3 (Convergence of the Marginal Distribution of Signals)**.** *For large n, the marginal distribution of signals is approximately equal to the distribution of true quality, and the approximation holds for all derivatives. Formally, for any $k = 0, 1, 2 \ldots$, as n converges to infinity,*

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \frac{d^k}{d\hat{\delta}^k} g(\hat{\delta}, n) + O(1/n)$$

uniformly in $\hat{\delta}$.

Proof. The kth derivative of the marginal distribution of the signal equals

$$\begin{aligned} \frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) &= \frac{d^k}{d\hat{\delta}^k} \int g(\delta) \cdot \phi\left(\hat{\delta}|\delta, \sigma_n^2\right) d\delta \\ &= \frac{d^k}{d\hat{\delta}^k} \int g(\delta) \cdot \frac{1}{\sigma_n} \phi\left(\frac{\delta - \hat{\delta}}{\sigma_n}\right) d\delta. \end{aligned}$$

With the change of variables

$$u = \frac{\delta - \hat{\delta}}{\sigma_n}$$

we have $du = d\delta/\sigma_n$ so that

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \frac{d^k}{d\hat{\delta}^k} \int g(\hat{\delta} + \sigma_n u) \cdot \phi(u)\ du.$$

The integrand and its derivatives with respect to $\hat{\delta}$ are integrable. Thus, we can use Leibniz's rule and differentiate under the integral sign, yielding

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \int \frac{d^k}{d\hat{\delta}^k} g(\hat{\delta} + \sigma_n u) \cdot \phi(u)\ du.$$

By Taylor's rule,

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \int \left[\frac{d^k}{d\hat{\delta}^k} g(\hat{\delta}) + \frac{d^{k+1}}{d\hat{\delta}^{k+1}} \cdot g(\hat{\delta})\sigma_n u + h(\sigma_n u) \cdot \frac{\sigma_n^2 u^2}{2} \right] \cdot \phi(u)\ du,$$

where the function h is bounded by $H = \sup_\delta d^{k+2} g(\delta)/d\delta^{k+2}$. H is finite by the assumption that the derivatives of g are bounded. Integrating we have

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \frac{d^k}{d\hat{\delta}^k} g(\hat{\delta}) + \int h(\sigma_n u) \cdot \frac{\sigma_n^2 u^2}{2} \cdot \phi(u)\ du.$$

The integral is bounded by $H\sigma_n^2/2$, yielding the desired approximation.

□

Substituting this approximation in the Tweedie formulas in Proposition A.1 yields the following asymptotic versions of the Tweedie formulas. Note that the variance formula is consistent with the intuition from the Bernstein von-Mises theorem, that the asymptotic variance of the Bayesian posterior is close to σ_n^2, which is the variance of a frequentist estimator that ignores the prior.

Corollary A.1 (Asymptotic Tweedie's Formula)**.** *Consider $\hat{\delta}_0$ with $g(\hat{\delta}_0) > 0$. Then, for all $\hat{\delta}$ in a neighborhood of $\hat{\delta}_0$, as n converges to infinity,*

$$P(\hat{\delta}, n) = \hat{\delta} + \sigma_n^2 \cdot \frac{d}{d\hat{\delta}} \log g(\hat{\delta}) + O(1/n^2),$$

and

$$\mathrm{Var}[\Delta | \hat{\Delta} = \hat{\delta}, n] = \sigma_n^2 + O(1/n^2).$$

These bound hold uniformly in $\hat{\delta}$. In particular,

$$\lim_{n\to\infty} P(\hat{\delta}_0, n) = \hat{\delta}_0.$$

Part 2: Completing the Proof

Proof of Theorem 1. Consider $\hat{\delta} > 0$ with $g(\hat{\delta}) > 0$ and $g(-\hat{\delta}) > 0$. By corollary A.1, $P(\hat{\delta}, n)$ converges to $\hat{\delta} > 0$ and $P(-\hat{\delta}, n)$ converges to $-\hat{\delta} < 0$. By the monotonicity of P, the limit of δ_n^* must be between $-\hat{\delta}$ and $\hat{\delta}$. Because $g(0) > 0$, there exist arbitrarily small such $\hat{\delta}$, so the limit of δ_n^* is zero.

The threshold δ_n^* satisfies $P(\delta_n^*, n) = 0$. Substituting the asymptotic Tweedie formula for P from Corollary A.1, we get

$$\begin{aligned}\delta_n^* &= -\sigma_n^2 \frac{d}{d\hat{\delta}} \log g(\delta_n^*) + O(1/n^2) \\ &= -\sigma_n^2 \cdot \frac{g'(0)}{g(0)} + O\left(\frac{1}{n} \cdot \delta_n^*\right) + O\left(\frac{1}{n^2}\right).\end{aligned}$$

The approximation in the second line follows because $g(0) > 0$ and the second derivative of g is bounded. This proves the desired asymptotic formula for t_n^*.

For the marginal product, if we substitute the aproximation for the marginal density in lemma A.3 and for the variance in corollary A.1 into the marginal product formula (A.3), we obtain

$$f'(n) = \frac{1}{2n} \cdot g(0) \cdot \sigma_n^2 + o\left(\frac{1}{n} \cdot \sigma_n^2\right),$$

implying the desired formula.

□

A.3.2 Proof of Theorem 2

Throughout this section we assume there is a *slowly varying* function $c(\delta)$ such that[34]

$$g(\delta) \sim \alpha c(\delta) \delta^{-(1+\alpha)} \tag{A.5}$$

as $|\delta| \to \infty$. In words, we will be assuming that the p.d.f. $g(\delta)$ is *regularly varying* at ∞ and $-\infty$ with exponent $-(\alpha + 1)$. We assume the existence of a strictly positive constant C such that $c(\delta) > C$ for $|\delta|$ large enough. Finally, we assume that $\mathbb{E}[\Delta] \equiv M < 0$.

Part 1: Integration Formulas and Auxiliary Definitions:

Define

[34] In a slight abuse of terminology we say that a positive function $c(\cdot)$ is slowly varying if $\lim_{|\delta| \to \infty} c(\lambda\delta)/c(\delta) \to 1$ for any $\lambda > 0$. Examples include constant functions, logarithmic functions, and others.

$$I_n(\underline{\delta}, \overline{\delta}, \beta) \equiv \int_{\underline{\delta}}^{\overline{\delta}} \delta^\beta g(\delta) \exp\left(-\frac{1}{2}\left(\frac{\delta - \delta_n^*}{\sigma_n}\right)^2\right) d\delta. \tag{A.6}$$

Both the marginal density and the posterior moments evaluated at the threshold signal δ_n^* can be written in terms of (A.6):

$$\begin{aligned} m(\delta_n^*\,,\, n) &= (\sqrt{2\pi}\sigma_n)^{-1} I_n(-\infty, \infty, 0), \\ \mathbb{E}[\Delta^\beta \mid \widehat{\Delta} = \delta_n^*; n] &= I_n(-\infty, \infty, \beta)/I_n(-\infty, \infty, 0). \end{aligned}$$

The definition of the threshold signal implies that $I_n(-\infty, \infty, 1) = 0$. We establish the asymptotics of the threshold $\delta^*(n)$ and the marginal product in a series of claims.

Claim 1: Divergence of the Threshold t-statistic

Claim 1. $t_n^* \equiv \delta_n^*/\sigma_n \to \infty$.

Proof. We establish the claim using a contradiction argument. Suppose $\delta_n^*/\sigma_n \nrightarrow \infty$. This implies the existence of a subsequence along which $\delta_{n_k}^*/\sigma_{n_k} \to C$, where either i) $-\infty < C < \infty$ or ii) $C = -\infty$. In the first case

$$\left(\frac{\delta - \delta_{n_k}^*}{\sigma_{n_k}}\right)^2 \to C^2, \quad \forall \delta.$$

The integrand in $I_n(-\infty, \infty, 1)$ is dominated by the integrable function $\delta g(\delta)$. The Dominated Convergence Theorem thus implies that

$$I_{n_k}(-\infty, \infty, 1) \to \int_{-\infty}^{\infty} \delta g(\delta) \exp\left(-\frac{C^2}{2}\right) d\delta = M \exp\left(-\frac{C^2}{2}\right) < 0,$$

which contradicts the optimality condition $I_n(-\infty, \infty, 1) = 0$ for all n. Thus, i) cannot hold.

Suppose ii) holds. Since $\mathbb{E}[\Delta] < 0$ and

$$f(n_k) = \int_{-\infty}^{\infty} \delta \Phi\left(\frac{\delta - \delta_{n_k}^*}{\sigma_{n_k}}\right) g(\delta) d\delta,$$

the Dominated Convergence Theorem implies $f(n_k) \to M < 0$. This is a contradiction: one can achieve a higher product by using the implementation strategy that does not implement any innovation regardless of the signal observed. □

Claim 2: Approximation for the integral near δ_n^*

By Claim 1, for any $0 < \epsilon < 1$ there exists n small enough such that

$$\sigma_n < B_n(\epsilon) \equiv \epsilon\delta_n^* < \delta_n^*,$$

Claim 2. For any power $\beta \geq 1$ and any $0 < \epsilon < 1$:

$$\begin{aligned} I_n(\delta_n^* - B_n(\epsilon), \delta_n^* + B_n(\epsilon), \beta) &\sim \sqrt{2\pi}\sigma_n {\delta_n^*}^{\beta} g(\delta_n^*), \\ &\sim \sqrt{2\pi}\sigma_n \alpha c(\delta_n^*) {\delta_n^*}^{\beta-\alpha-1} \end{aligned} \tag{A.7}$$

Proof. Using the change of variables $u \equiv \delta/\delta_n^*$ we can write

$$I_n(\delta_n^* - B_n(\epsilon), \delta_n^* + B_n(\epsilon), \beta)$$

as

$$(\delta_n^*)^{\beta+1} \int_{(1-\epsilon)}^{(1+\epsilon)} u^{\beta} g(u\delta_n^*) \exp\left(-\frac{1}{2}(u-1)^2 {t_n^*}^2\right) du. \tag{A.8}$$

Define:

$$\begin{aligned} I_1 &\equiv \int_{(1-\epsilon)}^{(1+\epsilon)} u^{\beta} (g(u\delta_n^*)/g(\delta_n^*)) \exp\left(-\frac{1}{2}(u-1)^2 {t_n^*}^2\right) du, \\ I_2 &\equiv \int_{(1-\epsilon)}^{(1+\epsilon)} u^{\beta-\alpha-1} \exp\left(-\frac{1}{2}(u-1)^2 {t_n^*}^2\right) du, \\ I_3 &\equiv \int_{(1-\epsilon)}^{(1+\epsilon)} u^{\beta} \exp\left(-\frac{1}{2}(u-1)^2 {t_n^*}^2\right) du. \end{aligned}$$

Laplace's method (Small (2010), Proposition 2, p. 196) implies that:

$$I_2 \sim \sqrt{2\pi}/t_n^* \sim I_3$$

as $t_n^* \to \infty$. Since g is bounded, Theorem A.5, Appendix A of Whitt (2002) implies that for $0 < \epsilon < 1$,

$$g(u\delta_n^*)/g(\delta_n^*) \to u^{-(1+\alpha)}$$

uniformly over $u \in [1-\epsilon, 1+\epsilon]$. Therefore, for any $\zeta > 0$ there exists $n(\zeta)$ small enough below which

$$I_2 - \zeta I_3 \leq I_1 \leq I_2 + \zeta I_3.$$

Since ζ is arbitrary we conclude that

$$I_1 \sim \sqrt{2\pi}/t_n^* = \sqrt{2\pi}\sigma_n/\delta_n^*.$$

Equation (A.8) implies that

$$I_n(\delta_n^* - B_n(\epsilon), \delta_n^* + B_n(\epsilon), \beta) = (\delta_n^*)^{\beta+1} g(\delta_n^*) I_1.$$

Therefore:

$$I_n(\delta_n^* - B_n(\epsilon), \delta_n^* + B_n(\epsilon), \beta) \sim \sqrt{2\pi}\sigma_n {\delta_n^*}^{\beta} g(\delta_n^*) \sim \sqrt{2\pi}\sigma_n \alpha c(\delta_n^*) {\delta_n^*}^{\beta-\alpha-1}.$$

□

Claim 3: Upper bound on δ_n^*

Claim 3.

$$\delta_n^* \leq (1 + o(1))\sqrt{2(\alpha - 1)\log \sigma_n}\sigma_n.$$

Proof. Take $0 < \epsilon < 1$. The optimality condition $I_n(-\infty, \infty, 1) = 0$ implies that

$$I_n(-\infty, 0, 1) + I_n((1-\epsilon)\delta_n^*, (1+\epsilon)\delta_n^*, 1) \leq I_n(-\infty, \infty, 1) = 0. \tag{A.9}$$

The first term in the equation above is bounded from below:

$$I_n(-\infty, 0, 1) > \int_{-\infty}^{0} \delta g(\delta) \exp\left(-\frac{1}{2} {t_n^*}^2\right) d\delta = -D \exp\left(-\frac{1}{2} {t_n^*}^2\right),$$

where $D \equiv \int_{-\infty}^{0} |\delta| g(\delta) d\delta$ is finite and nonzero by assumption. Claim 2 and equation (A.9) imply that

$$(1 + o(1))\sqrt{2\pi}\sigma_n \alpha c(\delta_n^*) {\delta_n^*}^{-\alpha} \leq D \exp\left(-\frac{1}{2} {t_n^*}^2\right),$$

which we can write as

$$(1 + o(1))\sqrt{2\pi}\sigma_n^{1-\alpha} \alpha c(\delta_n^*) {t_n^*}^{-\alpha} \leq D \exp\left(-\frac{1}{2} {t_n^*}^2\right).$$

Taking logarithms on both sides and dividing by $-(1/2){t_n^*}^2$ implies

$$\frac{2(\alpha - 1)\log \sigma_n}{{t_n^*}^2} - \frac{2\log(c(\delta_n^*))}{{t_n^*}^2} \geq 1 + o(1).$$

By assumption $c(\delta_n^*)$ is bounded from below by a constant $C > 0$. Hence

$$\frac{2(\alpha-1)\log\sigma_n}{t_n^{*2}} \geq 1 + o(1),$$

which implies that:

$$(1+o(1))\sqrt{2(\alpha-1)\log\sigma_n}\sigma_n \geq \delta_n^*.$$

□

Claim 4: Integral around 0 for $1 \leq \beta < \alpha$

For $\gamma \in (0,1)$ define

$$A_n(\gamma) \equiv \left(\frac{\sigma_n^2}{\delta_n^*}\right)^\gamma.$$

Claim 1 implies $A_n(\gamma) \in o(\sigma_n)$ and $A_n(\gamma) < \delta_n^*$ for n small enough. Claim 3 implies $A_n(\gamma) \to \infty$ and $A_n(\gamma) \in o(\sigma_n^2/\delta_n^*)$. In the remaining part of this appendix we will often use A_n instead of $A_n(\gamma)$, for the sake of notational simplicity.

We split the integral I_n into different regions. Most of the value of the integral comes from two regions: $\delta \in [-A_n(\gamma)\ ,\ A_n(\gamma)]$ (where g is large and the exponential is small) and $\delta_n \in [\delta_n^* - B_n(\epsilon)\ ,\ \delta_n^* + B_n(\epsilon)]$ (where g is small and the exponential is large).

Claim 4. For any integer β such that $1 \leq \beta < \alpha$, $\mathbb{E}[\Delta^\beta] \neq 0$; and any $0 < \gamma < 1$,

$$I_n(-A_n(\gamma), A_n(\gamma), \beta) \sim \mathbb{E}[\Delta^\beta]\exp\left\{-\frac{1}{2}t_n^{*2}\right\}.$$

Proof. The difference

$$I_n(-A_n, A_n, \beta) - \mathbb{E}[\Delta^\beta]\exp\left(-\frac{1}{2}t_n^{*2}\right)$$

can be decomposed as

$$\begin{aligned}
&\int_{-A_n}^{A_n} \delta^\beta g(\delta)\cdot\left[\exp\left\{-\frac{1}{2}\left(\frac{\delta-\delta_n^*}{\sigma_n}\right)^2\right\} - \exp\left\{-\frac{1}{2}\left(\frac{\delta_n^*}{\sigma_n}\right)^2\right\}\right]d\delta \\
+\ &\left[\int_{-A_n}^{A_n}\delta^\beta g(\delta)d\delta - \int_{-\infty}^{\infty}\delta^\beta g(\delta)d\delta\right]\cdot\exp\left\{-\frac{1}{2}\left(\frac{\delta_n^*}{\sigma_n}\right)^2\right\} \qquad \text{(A.10)}
\end{aligned}$$

The first term in equation (A.10) is smaller than

$$\mathbb{E}[|\Delta|^\beta]\left[\exp\left\{A_n\cdot\frac{\delta_n^*}{\sigma_n^2}\right\} - 1\right]\cdot\exp\left\{-\frac{1}{2}t_n^{*2}\right\}.$$

By construction $A_n \in o(\sigma_n^2/\delta_n^*)$, implying the term above is $o\left(\exp(-(1/2)t_n^{*2})\right)$. The second term equals

$$-\left[\int_{-\infty}^{-A_n} \delta^\beta g(\delta)d\delta + \int_{A_n}^{\infty} \delta^\beta g(\delta)d\delta\right] \cdot \exp\left\{-\frac{1}{2}\left(\frac{\delta_n^*}{\sigma_n}\right)^2\right\}.$$

Since $\beta < \alpha$, Karamata's integral theorem (Theorem 1a p. 281 in Feller (1967)) implies the second term equals

$$-(1+o(1))\left[\frac{\alpha}{\alpha-\beta}c(A_n)A_n^{\beta-\alpha} + (-1)^\beta \frac{\alpha}{\alpha-\beta}c(-A_n)A_n^{\beta-\alpha}\right] \cdot \exp\left\{-\frac{1}{2}t_n^{*2}\right\}.$$

Since any slowly varying function is such that $|\delta|^{-\eta}c(|\delta|) \to 0$ for all $\eta > 0$ (see equation 2 in Karamata (1962)), then

$$I_n(-A_n, A_n, \beta) - \mathbb{E}[\Delta^\beta]\exp\left(-\frac{1}{2}t_n^{*2}\right) = o\left(\exp\left\{-\frac{1}{2}t_n^{*2}\right\}\right).$$

Since $\mathbb{E}[\Delta^\beta] \neq 0$, the result follows.

□

Claim 5: Integral around 0 for arbitrary β

Claim 5. For any integer $\beta \geq 1$ and any $0 < \gamma < 1$,

$$I_n(-A_n(\gamma)\,,\ A_n(\gamma), \beta) \in O\left(A_n^{\beta-1}\exp\left\{-\frac{1}{2}t_n^{*2}\right\}\right).$$

Proof. $|I_n(-A_n, A_n, \beta)|$ is bounded by

$$\begin{aligned}
&\int_{-A_n}^{A_n} |\delta|^\beta g(\delta)\exp\left\{-\frac{1}{2}\left(\frac{\delta-\delta_n^*}{\sigma_n^2}\right)^2\right\}d\delta \\
\leq\ & A_n^{\beta-1}\int_{-A_n}^{A_n} |\delta| g(\delta)\exp\left\{-\frac{1}{2}\left(\frac{\delta-\delta_n^*}{\sigma_n^2}\right)^2\right\}d\delta \\
=\ & A_n^{\beta-1}(1+o(1))\mathbb{E}[|\Delta|]\exp\left\{-\frac{1}{2}t_n^{*2}\right\},
\end{aligned}$$

where the last line follows from an argument identical to the proof of Claim 4. □

Claim 6: Integral below $-A_n$

Claim 6. Let $0 < \gamma < 1$. For any $\alpha > 1$, and any integer $\beta \geq 1$:

$$I_n(-\infty, -A_n(\gamma), \beta) \in o\left(\delta_n^{*\beta-1} \exp\left\{-\frac{1}{2}t_n^{*2}\right\}\right).$$

Proof. $|I_n(-\infty, -A_n, \beta)|$ is bounded above by the product of $\exp\left\{-(1/2)t_n^{*2}\right\}$ and

$$\int_{-\infty}^{-A_n} |\delta|^\beta g(\delta) \exp\left\{\frac{\delta_n^* \delta}{\sigma_n^2} - \frac{1}{2}\left(\frac{\delta}{\sigma_n}\right)^2\right\} d\delta. \tag{A.11}$$

Since $\delta \leq 0$, equation (A.11) is further bounded by

$$\int_{-\infty}^{-A_n} |\delta| g(\delta) H_\beta(|\delta|) d\delta, \tag{A.12}$$

where $H_\beta(\delta) \equiv \delta^{\beta-1} e^{\frac{-\delta^2}{2\sigma_n^2}}$ is defined for $\delta \geq 0$. In this range, the function $H_\beta(\cdot)$ is maximized at $\delta_n^+ \equiv (\beta-1)^{1/2}\sigma_n$.[35] The integral in (A.12) can then be bounded by

$$H_\beta(\delta_n^+) \int_{-\infty}^{-A_n} |\delta| g(\delta) d\delta \tag{A.13}$$

where $H_\beta(\delta_n^+) = (\beta-1)^{\frac{\beta-1}{2}} e^{\frac{-(\beta-1)}{2}} {\sigma_n}^{\beta-1} = O\left({\sigma_n}^{\beta-1}\right)$ if $\beta > 1$, and $H_1(\delta_n^+) = 1$ when $\beta = 1$. By assumption, $E[|\Delta|] < +\infty$, as a result $\int_{-\infty}^{-A_n} |\delta| g(\delta) d\delta \in o(1)$. Therefore

$$|I_n(-\infty, -A_n, \beta)| \in o\left({\sigma_n}^{\beta-1} \exp\left\{-(1/2)t_n^{*2}\right\}\right).$$

Since $\sigma_n \in o(\delta_n^*)$, we conclude that

$$|I_n(-\infty, -A_n, \beta)| \in o\left(\delta_n^{*\beta-1} \exp\left\{-(1/2)t_n^{*2}\right\}\right).$$

□

Claim 7: Integral between A_n and $\delta_n^* - B_n(\epsilon)$

Claim 7. Take any $\alpha > 1$ and $\beta \geq 1$. For any $\epsilon, \gamma \in (0,1)$ such that $\gamma > 2(1-\epsilon)$, then

[35] The derivative of $H_\beta(\cdot)$ is given by $H'_\beta(\delta) = [(\beta-1) - (\delta^2/\sigma_n^2)]H_\beta(\delta)/\delta$.

$$I_n(A_n(\gamma), \delta_n^* - B_n(\epsilon), \beta) \in o\left(\delta_n^{*\beta-1} \exp\left\{-\frac{1}{2}t_n^{*2}\right\}\right).$$

Proof. $|I_n(A_n, \delta_n^* - B_n(\epsilon), \beta)|$ equals

$$\exp\left\{-\frac{1}{2}t_n^{*2}\right\} \int_{A_n}^{(1-\epsilon)\delta_n^*} \delta g(\delta) H_\beta(\delta) d\delta,$$

where $H_\beta(\delta) \equiv \delta^{\beta-1} \exp\left(\frac{-\delta^2}{2\sigma_n^2} + \frac{\delta_n^* \delta}{\sigma_n^2}\right)$. $H_\beta(\cdot)$ is an increasing function on the interval $[A_n, (1-\epsilon)\delta_n^*]$.[36]Consequently, $|I_n(A_n, \delta_n^* - B_n(\epsilon), \beta)|$ can be bounded by the product of $\exp\left\{-\frac{1}{2}t_n^{*2}\right\}$ and

$$\int_{A_n}^{(1-\epsilon)\delta_n^*} \delta g(\delta) H_\beta((1-\epsilon)\delta_n^*) d\delta \leq H_\beta((1-\epsilon)\delta_n^*) R_n,$$

where $R_n \equiv \int_{A_n}^{\infty} \delta g(\delta) d\delta$. Karamata's integral theorem (Theorem 1a p. 281 in Feller (1967)) implies that

$$R_n \sim \frac{\alpha}{\alpha - 1} A_n^2 g(A_n) \sim \frac{\alpha}{\alpha - 1} A_n^{-(\alpha-1)} c(A_n).$$

Consider $0 < \eta \equiv (\alpha - 1)/2 < \alpha - 1$:

$$\begin{aligned}
\frac{\alpha}{\alpha - 1} A_n^{-(\alpha-1)} c(A_n) &= \frac{\alpha}{\alpha - 1} A_n^{-(\alpha-1)/2} A_n^{-\eta} c(A_n), \\
&= \frac{\alpha}{\alpha - 1} A_n^{-(\alpha-1)/2} o(1), \\
&\quad \text{(since for any } \eta > 0, A_n^{-\eta} c(A_n) \to 0 \text{ as } A_n \to \infty) \\
&= \frac{\alpha}{\alpha - 1} t_n^{*\gamma(\alpha-1)/2} \frac{1}{\sigma_n^{\gamma(\alpha-1)/2}} o(1), \\
&\quad \text{(by definition of } A_n(\gamma)).
\end{aligned}$$

Therefore,

$$R_n \in o\left(t_n^{*\gamma(\alpha-1)/2} \frac{1}{\sigma_n^{\gamma(\alpha-1)/2}}\right).$$

[36] $H'_\beta(\delta) = [-\delta^2 + \delta_n^* \delta + (\beta - 1)\sigma_n^2] H_\beta(\delta)/(\delta \sigma_n^2)$. The sign of the derivative thus depends on the sign of the quadratic function $-\delta^2 + \delta_n^* \delta + (\beta - 1)\sigma_n^2$, which can be written as $-(\delta - \delta_n^-)(\delta - \delta_n^+)$ where

$$\delta_n^\pm = \frac{\delta_n^*}{2}\left(1 \pm \sqrt{1 + 4(\beta - 1)\sigma_n^2/\delta_n^{*2}}\right)$$

For $n$small enough, we have $\delta_n^- \leq 0 \leq A_n$ and $(1-\epsilon)\delta_n^* \leq \delta_n^+ \sim \delta_n^*$.

The definition of $H_\beta(\cdot)$ further implies that

$$\begin{aligned} H_\beta((1-\epsilon)\delta_n^*) &= (1-\epsilon)^{\beta-1}\delta_n^{*\beta-1}\exp\left(-\frac{(1-\epsilon)^2\delta_n^{*2}}{2\sigma_n^2}+\frac{\delta_n^*(1-\epsilon)\delta_n^*}{\sigma_n^2}\right), \\ &= (1-\epsilon)^{\beta-1}\delta_n^{*\beta-1}\exp\left(t_n^{*2}\left((1-\epsilon)-\frac{1}{2}(1-\epsilon)^2\right)\right), \\ &= (1-\epsilon)^{\beta-1}\delta_n^{*\beta-1}\exp\left((1-\epsilon){t_n^*}^2/2\right). \end{aligned}$$

Claim 3 showed that ${t_n^*}^2/2 \le (1+o(1))(\alpha-1)\log\sigma_n$. Consequently:

$$\begin{aligned} H_\beta((1-\epsilon)\delta_n^*) &\le (1-\epsilon)^{\beta-1}\delta_n^{*\beta-1}\exp\left((1-\epsilon)(1+o(1))(\alpha-1)\log\sigma_n\right), \\ &= (1-\epsilon)^{\beta-1}\delta_n^{*\beta-1}\sigma_n^{(1-\epsilon)(\alpha-1)(1+o(1))}. \end{aligned}$$

Therefore

$$H_\beta((1-\epsilon)\delta_n^*)R_n = H_\beta((1-\epsilon)\delta_n^*)o\left(t_n^{*\gamma(\alpha-1)/2}\frac{1}{\sigma_n^{\gamma(\alpha-1)/2}}\right),$$

and the bound on $H_\beta((1-\epsilon)\delta_n^*)$ implies

$$H_\beta((1-\epsilon)\delta_n^*)R_n \le \delta_n^{*\beta-1}o\left(t_n^{*\gamma(\alpha-1)/2}\frac{1}{\sigma_n^{(\gamma(\alpha-1)/2)-(1-\epsilon)(\alpha-1)(1+o(1))}}\right).$$

Using again the upper bound for t_n^* in Claim 3 gives.

$$H_\beta((1-\epsilon)\delta_n^*)R_n \le \delta_n^{*\beta-1}o\left(\frac{\log\sigma_n{}^{\gamma(\alpha-1)/2}}{\sigma_n^{((\alpha-1)/2)[\gamma-2(1-\epsilon)(1+o(1))]}}\right).$$

Since $\gamma-2(1-\epsilon)>0$, then $\gamma-2(1-\epsilon)(1+o(1))>0$ for n small enough. We conclude that $H_\beta((1-\epsilon)\delta_n^*)R_n \in o(\delta_n^{*\beta-1})$ and therefore

$$|I_n(A_n,\delta_n^*-B_n(\epsilon),\beta)| \in o\left(\delta_n^{*\beta-1}\exp\left\{-\frac{1}{2}t_n^{*2}\right\}\right).$$

□

Claim 8: Integral to the right of $\delta_n^*+B_n(\epsilon)$ is small

Claim 8. For any $\beta<\alpha+1$ and $0<\epsilon<1$:

$$I_n(\delta_n^*+B_n(\epsilon),\infty,\beta) \in o\left(\sigma_n\delta_n^{*\beta}g(\delta_n^*)\right).$$

Proof. Define:

$$\begin{aligned}
I_1 &\equiv t_n^* \int_{1+\epsilon}^{\infty} u^{\beta} (g(u\delta_n^*)/g(\delta_n^*)) \exp\left(-\frac{1}{2}(u-1)^2 t_n^{*2}\right) du, \\
I_2 &\equiv t_n^* \int_{1+\epsilon}^{\infty} u^{\beta-(\alpha+1)} \exp\left(-\frac{1}{2}(u-1)^2 t_n^{*2}\right) du, \\
I_3 &\equiv t_n^* \int_{1+\epsilon}^{\infty} u^{\beta} \exp\left(-\frac{1}{2}(u-1)^2 t_n^{*2}\right) du, \\
I_4 &\equiv t_n^* \int_{1-\epsilon}^{\infty} u^{\beta} \exp\left(-\frac{1}{2}(u-1)^2 t_n^{*2}\right) du.
\end{aligned}$$

1. Since $\beta < \alpha + 1$

$$\begin{aligned}
I_2 &\leq \sqrt{2\pi}\,(1+\epsilon)^{\beta-(\alpha+1)}\ (1-\Phi(\epsilon t_n^*)), \\
&\quad \text{(by definition of the standard normal c.d.f.)} \\
&= \sqrt{2\pi}\,(1+\epsilon)^{\beta-(\alpha+1)}\ \Phi(-\epsilon t_n^*), \\
&= O\left(\exp\left(-\frac{1}{2}\epsilon t_n^{*2}\right)\right). \\
&\quad \text{(by equation 26.2.12 in Abramowitz and Stegun (1964))}
\end{aligned}$$

2. Laplace's method implies that $I_4 \sim \sqrt{2\pi}$.

3. By assumption, g is bounded. Hence, Theorem A.5, Appendix A of Whitt (2002) implies that for $\epsilon > 0$:
$$(g(u\delta_n^*)/g(\delta_n^*) \to u^{-(1+\alpha)}$$
uniformly over $u \in [1-\epsilon, \infty)$.

Therefore, for any γ there exists small enough $n(\gamma)$ below which

$$I_1 \leq I_2 + \gamma I_3 \leq I_2 + \gamma I_4 = I_2 + \gamma(1+o(1))\sqrt{2\pi}$$

Using the change of variables $u = \delta/\delta_n^*$ and the inequality above

$$\begin{aligned}
0 \leq I_n((1+\epsilon)\delta_n^*, \infty, \beta)/\sigma_n \delta_n^{*\beta} g(\delta_n^*) &= I_1 \\
&\leq (I_2 + \gamma(1+o(1))\sqrt{2\pi}) \\
&\leq O\left(\exp\left(-\frac{1}{2}\epsilon t_n^{*2}\right)\right) \\
&+ \gamma(1+o(1))\sqrt{2\pi})
\end{aligned}$$

Since this holds for any $\gamma > 0$, we conclude that

$$I_n((1+\epsilon)\delta_n^*, \infty, \beta) \in o\left(\sigma_n {\delta_n^*}^{\beta} g(\delta_n^*)\right).$$

□

Part 2: Asymptotics of the Threshold

Lemma A.4. *Under the assumptions of Theorem 2*

$$t_n^* \equiv \delta_n^*/\sigma_n \sim \sqrt{2(\alpha-1)\log\sigma_n}.$$

Proof. Since $\alpha > 1$, the optimality condition $I_n(-\infty, \infty, 1) = 0$ and Claims 1, 2, 4, 6, 8, and 7 imply that

$$(1+o(1)) \cdot (-m) \cdot \exp\left\{-\frac{1}{2}{t_n^*}^2\right\} = \sqrt{2\pi}\sigma_n \delta_n^* g(\delta_n^*) = (1+o(1))\sqrt{2\pi}\sigma_n^{1-\alpha}\alpha c(\delta_n^*){t_n^*}^{-\alpha}.$$

Taking logs on both sides implies that

$$o(1) + \log(-m) - \frac{1}{2}{t_n^*}^2 = \log(\sqrt{2\pi}\alpha) + \log(c(\delta_n^*)) + (1-\alpha)\log\sigma_n - \alpha t_n^*,$$

which implies that for every $\eta > 0$

$$(1+o(1))\frac{1}{2}{t_n^*}^2 = (\alpha - 1 - \eta)\log\sigma_n - \log(c(\delta_n^*)/{\delta_n^*}^{\eta}).$$

Since $c(\delta_n^*)/{\delta_n^*}^{\eta} \to 0$ for every $\eta > 0$, for any small enough n.

$$\frac{1}{2}{t_n^*}^2 \geq (1+o(1)) \cdot (\alpha - 1 - \eta)\log\sigma_n.$$

We conclude that for any $\eta > 0$

$$\liminf_{n\to 0} \frac{{t_n^*}^2}{2(\alpha-1)\log\sigma_n} \geq 1 - \frac{\eta}{\alpha - 1}.$$

Claim 3 then implies

$${t_n^*}^2 \sim 2(\alpha-1)\log\sigma_n.$$

□

Part 3: Asymptotics of the Marginal Product

Lemma A.5. *Under the assumptions of Theorem 2*

$$f'(n) \sim \frac{1}{2n} \cdot g(\delta_n^*) \cdot \delta_n^{*2} \sim \frac{1}{2n} \cdot \alpha c(\delta_n^*) \cdot (\delta_n^*)^{-(\alpha-1)}$$

Proof. The proof has 3 steps.

STEP 1: Lemma A.4 implies that:

$$\exp\left\{-\frac{1}{2}t_n^{*2}\right\} \in O\left(\sigma_n \delta_n^* g(\delta_n^*)\right).$$

STEP 2: Claims 2, 5, 6, 7, 8 and the fact that $A_n(\gamma) \in o(\delta_n^*)$ for any $0 < \gamma < 1$ imply that

$$I_2(-\infty, \infty, 2) = o(1)\delta_n^* \exp\left\{-\frac{1}{2}t_n^{*2}\right\} + (1 + o(1))\sqrt{2\pi}\sigma_n \delta_n^{*2} g(\delta_n^*)$$

Step 1 and $\delta_n^* \to \infty$ imply that:

$$I_2(-\infty, \infty, 2) = (1 + o(1))\sqrt{2\pi}\sigma_n \delta_n^{*2} g(\delta_n^*)$$

STEP 3: The envelope theorem formula implies that:

$$f'(n) = \frac{1}{2n} \frac{1}{\sqrt{2\pi}\sigma_n} I_n(-\infty, \infty, 2).$$

Steps 2 and 3 imply that for any $\alpha > 1$:

$$f'(n) \sim \frac{1}{2n}\delta_n^{*2} g(\delta_n^*) \sim \frac{1}{2n}\alpha c(\delta_n^*)\delta_n^{*-(\alpha-1)}.$$

□

Part 4: Completing the Proof

Below we establish the four parts of Theorem 2.

1. Lemma A.4 showed that $t_n^{*2} \sim 2(\alpha - 1)\log \sigma_n$. The continuity of the square root function and the definition of the asymptotic equivalence relation($\sim$) imply that

$$t_n^* \sim \sqrt{2(\alpha - 1)\log \sigma_n},$$

where $\sigma_n \equiv \sigma/\sqrt{n}$.

2. Lemma A.5 showed that

$$f'(n) \sim \frac{1}{2n}\delta_n^{*2} g(\delta_n^*) \sim \frac{1}{2n}\alpha c(\delta_n^*)\delta_n^{*-(\alpha-1)}.$$

Since $t_n^* \equiv \delta_n^*/\sigma_n$, then

$$\begin{aligned} f'(n) &\sim \frac{1}{2n}\alpha c(\delta_n^*) t_n^{*-(\alpha-1)}\sigma_n^{-(\alpha-1)}, \\ &= \frac{1}{2n}\alpha c(\delta_n^*)(\sigma t_n^*)^{-(\alpha-1)}\sqrt{n}^{(\alpha-1)}, \\ &= \frac{1}{2}\alpha c(\delta_n^*)(\sigma t_n^*)^{-(\alpha-1)} n^{(\alpha-3)/2}. \end{aligned}$$

3. Note that

$$\begin{aligned} f'(n) &> (1+o(1))\frac{1}{2}\alpha C(\sigma t_n^*)^{-(\alpha-1)} n^{(\alpha-3)/2}, \\ &\quad \text{(since we have assumed that } c(\delta_n^*) > C\text{)}, \\ &= O(1)\,(\log \sigma_n)^{-(\alpha-1)/2} n^{(\alpha-3)/2}, \\ &\quad \text{(by Part 1 of Theorem 2)}, \\ &= O(1)\left(\left(log\left(\frac{1}{n}\right)\right)^{-1}\left(\frac{1}{n}\right)^{(3-\alpha)/(\alpha-1)}\right)^{(\alpha-1)/2}. \end{aligned}$$

The result follows, as for $1 < \alpha < 3$

$$\lim_{n\to 0}\left(log\left(\frac{1}{n}\right)\right)^{-1}\left(\frac{1}{n}\right)^{(3-\alpha)/(\alpha-1)} \to \infty.$$

4. For any $\eta > 0$

$$\begin{aligned} f'(n) &= O(1)c(\delta_n^*)(t_n^*)^{-(\alpha-1)} n^{(\alpha-3)/2}, \\ &= O(1)(c(\delta_n^*)/\delta_n^{*\eta})(\delta_n^*)^{\eta}(t_n^*)^{-(\alpha-1)} n^{(\alpha-3)/2}, \\ &= O(1)(c(\delta_n^*)/\delta_n^{*\eta})(t_n^*)^{-(\alpha-1-\eta)} n^{(\alpha-3-\eta)/2}. \end{aligned}$$

Since $\alpha > 3$, there exists $\eta > 0$ such that $\alpha - 1 > \eta$ and $\alpha - 3 > \eta$. For any such η :

$$\lim_{n\to 0}(c(\delta_n^*)/\delta_n^{*\eta})(t_n^*)^{-(\alpha-1-\eta)} n^{(\alpha-3-\eta)/2} = 0,$$

as for any slowly varying function $c(\delta_n^*)/\delta_n^{*\eta} \to 0$ as $\delta_n^* \to \infty$ (Lemma 2 in Feller (1967) p. 277).

A.4 Additional Proofs

A.4.1 Proof of Remark 1

Suppose that for N large enough, the optimal experimentation strategy is not lean. Then, it is possible to construct a sequence $N_k \to \infty$ such that, in the optimal experimentation strategy, an idea i is not experimented on at all, and and idea j is experimented on the most. The idea that is experimented on the most has at least N_k/I users. So the gain from taking half the users in idea j and placing them on idea i is at least

$$f_i(N_k/2I) + f_j(N_k/2I) - f_j(N_k). \tag{A.14}$$

In the proof of Lemma A.2 the Appendix we showed

$$f_i(n) = \int_{-\infty}^{\infty} \delta \Phi\left(\frac{\delta - \delta_n^*}{\sigma_n}\right) g_i(\delta) d\delta.$$

Part 1 of Theorem 1 showed that $\delta_n^*/\sigma_n \to 0$ as $n \to \infty$. Consequently, $f_i(n) \to f_i(\infty) \equiv \mathbb{E}[\Delta_i^+] > 0$. We conclude that along the sequence N_k the bound on A.14 converges to $f_i(\infty)$, reaching a contradiction.

A.4.2 Proof of Corollary 1

Part 1: First we show that if $\alpha < 3$ it is optimal to run experiments on all ideas (go lean). The result follows directly from the first order conditions of the firm's problem.

The firm's optimal experimentation strategy solves

$$\max_{n_1,\ldots,n_I} \sum_{i=1}^{I} f_i(n_i), \text{ s.t. } \sum_{i=1}^{I} n_i \leq N, \quad n_i \geq 0 \, \forall \, i.$$

The Karush-Kuhn-Tucker (KKT) conditions imply that the optimal experimentation strategy $\mathbf{n} = (n_1^*, \ldots, n_I^*)$ must satisfy:

$$\begin{aligned} f_i'(n_i^*) - \lambda + \mu_i &= 0, \quad \forall i, \\ \lambda(\sum_{i=1}^{I} n_i^* - I) &= 0, \quad \lambda \geq 0. \\ n_i^* \mu_i &= 0, \quad \mu_i \geq 0, \quad \forall i \end{aligned}$$

Suppose that $\alpha < 3$ and that $\mathbf{n}$ is not lean. Any such experimentation strategy must leave at least some idea i left untested. Without loss of generality suppose that it is the first one.

This means that the Lagrange Multipliers, which are finite, must satisfy:

$$f'_1(0) - \lambda + \mu_1 = 0.$$

This is a contradiction, as Theorem 2 have shown that whenever $\alpha < 3$, $f'_1(0)$ is infinity. We conclude that when $\alpha < 3$ the optimal experimentation strategy must be lean.

Part 2: We now show that if $\alpha > 3$, it is optimal to concentrate all the experimental resources on only one idea. Suppose this is not the case. Then an optimal experimentation strategy must have at least two ideas i, j such that $n_i > 0$ and $n_j > 0$. Without loss of generality, let $i = 1$ and $j = 2$. We slightly abuse notation and write n_1, n_2 instead of $n_1(N)$ and $n_2(N)$, unless confusion may arise. We derive a contradiction in two steps.

STEP 1: We use Theorem 2 to show that if an optimal experimentation strategy has $n_1 > 0, n_2 > 0$, then

$$\lim_{N\to 0} \frac{n_1}{n_2} = \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{\alpha-1}{\alpha-3}}.$$

We will use this result to argue that if N is small enough and $n_1 > 0, n_2 > 0$, it is optimal to assign more participants to the idea with the largest experimental noise.

Proof. If it is indeed optimal to A/B test both ideas, then

$$\frac{f_1'(n_1)}{f_2'(n_2)} = 1$$

(by the first-order conditions). Part 2 of Theorem 2 gives

$$(1 + o(1))\frac{c(\delta^*(n_1))}{c(\delta^*(n_2))}\left(\frac{\sigma_1 t^*(n_1)}{\sigma_2 t^*(n_2)}\right)^{-(\alpha-1)}\left(\frac{n_1}{n_2}\right)^{(\alpha-3)/2} = 1$$

(since, by assumption, $\alpha_1 = \alpha_2$). We have also assumed that $c(\delta)$ converges to some constant c as $\delta \to \infty$. Consequently, Part 1 of Theorem 2 implies

$$\lim_{N\to 0}\left(\frac{\sigma_2^2 \ln\left(\frac{\sigma_2}{\sqrt{n_2}}\right)}{\sigma_1^2 \ln\left(\frac{\sigma_1}{\sqrt{n_1}}\right)}\right)^{(\alpha-1)/2}\left(\frac{n_1}{n_2}\right)^{(\alpha-3)/2} = 1. \tag{A.15}$$

The sequence n_1/n_2 must be bounded (otherwise, the left hand side of the equation above will diverge to infinty). An anlogous argument implies that n_2/n_1 is also bounded.

Take any convergent subsequence of n_1/n_2. Such a subsequence exists by virtue of the Bolzano-Weirstrass Theorem. Let η denote its limit. Since n_2/n_1 is also bounded, then

$\eta > 0$. Algebraic manipulation of equation A.15 implies

$$\left(\frac{\sigma_2^2}{\sigma_1^2}\right)^{(\alpha-1)} \eta^{(\alpha-3)} = 1.$$

Hence

$$\eta = \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{\alpha-1}{\alpha-3}}.$$

The subsequence of n_1/n_2 was taken arbitrarily. Therefore any convergent subsequence must have the same limit. We conclude that

$$\lim_{N\to 0} \frac{n_1}{n_2} = \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{\alpha-1}{\alpha-3}}.$$

□

STEP 2: Assume $\sigma_1 \geq \sigma_2$. We will show that

$$\lim_{N\to 0} \frac{f_1(n_1(N) + n_2(N))}{f_1(n_1(N)) + f_2(n_2(N))} > 1. \tag{A.16}$$

This means that if N is small enough, the experimentation strategy in which $n_1^* = n_1 + n_2$, $n_2^* = 0$ leads to higher output than the strategy in which $n_1 > 0,\ n_2 > 0$. This will contradict the optimality of the strategy in which idea 1 and 2 are A/B tested.

Proof. Since the limit of both the numerator and the denominator in equation A.16 is zero, we can use L'Hôpital's rule and focus on

$$\lim_{N\to 0} \frac{df_1(n_1(N) + n_2(N))/dN}{d(f_1(n_1(N)) + f_2(n_2(N)))/dN}.$$

Both $n_1$and n_2 are differentiable functions of N. Consequently, it is sufficient to show that

$$\lim_{N\to 0} \frac{f_1'(n_1(N) + n_2(N))(n_1'(N) + n_2'(N))}{f_1'(n_1(N))n_1'(N) + f_2'(n_1(N))n_2'(N)} > 1.$$

This inequalty holds if and only if

$$\lim_{N\to 0} \frac{f_1'(n_1(N) + n_2(N))}{f_1'(n_1(N))} > 1,$$

since$f_1'(n_1(N)) = f_2'(n_2(N))$ is a necessary condition for optimality.

In a slight abuse of notation, Theorem 2 implies

$$\begin{aligned}
\frac{f_1'(n_1+n_2)}{f_1'(n_1)} &= (1+o(1))\left(\frac{t^*(n_1+n_2)}{t^*(n_1)}\right)^{-(\alpha-1)}\left(1+\frac{n_1}{n_2}\right)^{(\alpha-3)/2}, \\
&= (1+o(1))\left(\frac{\ln\left(\frac{\sigma_1}{\sqrt{n_1+n_2}}\right)}{\ln\left(\frac{\sigma_1}{\sqrt{n_1}}\right)}\right)^{-(\alpha-1)/2}\left(1+\frac{n_1}{n_2}\right)^{(\alpha-3)/2}, \\
&= (1+o(1))\left(1+\frac{\ln\left(\frac{1}{\sqrt{1+n_2/n_1}}\right)}{\ln\left(\frac{\sigma_1}{\sqrt{n_1}}\right)}\right)^{-(\alpha-1)/2}\left(1+\frac{n_1}{n_2}\right)^{(\alpha-3)/2}.
\end{aligned}$$

And from Step 1

$$\begin{aligned}
\lim_{N\to 0}\frac{f_1'(n_1(N)+n_2(N))}{f_1'(n_1(N))} &= \left(1+\left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{\alpha-1}{\alpha-3}}\right)^{(\alpha-3)/2}, \\
&\geq 2^{(\alpha-3)/2}, \\
&\quad (\text{as } \sigma_1^2 \geq \sigma_2^2) \\
&> 1.
\end{aligned}$$

Where the last line is implied by the condition $\alpha > 3$. Thus, we have shown that moving all the experimental resources to the idea with a larger experimental noise leads to a higher output level. This contradicts the optimality of any experimentation strategy in which $n_1 > 0$ and $n_2 > 0$. Therefore, when N is small the optimal experimentation strategy when $\alpha > 3$ must be big (with all the experimental resources concentrated on only one idea). □