Confidence intervals for policy evaluation in adaptive experiments

Vitor Hadad[a,1], David A. Hirshberg[a], Ruohan Zhan[b], Stefan Wager[a], and Susan Athey[a,1]

[a]Stanford Graduate School of Business, Stanford University, Stanford, CA 94305; and [b]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305

Contributed by Susan Athey, February 16, 2021 (sent for review July 10, 2020; reviewed by Maximilian Kasy and Jasjeet Sekhon)

Adaptive experimental designs can dramatically improve efficiency in randomized trials. But with adaptively collected data, common estimators based on sample means and inverse propensity-weighted means can be biased or heavy-tailed. This poses statistical challenges, in particular when the experimenter would like to test hypotheses about parameters that were not targeted by the data-collection mechanism. In this paper, we present a class of test statistics that can handle these challenges. Our approach is to adaptively reweight the terms of an augmented inverse propensity-weighting estimator to control the contribution of each term to the estimator's variance. This scheme reduces overall variance and yields an asymptotically normal test statistic. We validate the accuracy of the resulting estimates and their CIs in numerical experiments and show that our methods compare favorably to existing alternatives in terms of mean squared error, coverage, and CI size.

adaptive experimentation | multiarmed bandits | policy evaluation | central limit theorem | frequentist inference]

Adaptive experimental designs can dramatically improve efficiency for particular objectives: maximizing welfare during the experiment (1, 2) or after it (3, 4); quickly identifying the best treatment arm (5, 6); maximizing the power of a particular hypothesis (7–9); and so on. To achieve these efficiency gains, we adaptively choose assignments to resolve uncertainty about some aspects of the data-generating process, at the expense of learning little about others. For example, welfare-maximizing designs tend to focus on differentiating optimal and near-optimal treatments, collecting relatively little data about suboptimal ones.

However, once the experiment is over, we are often interested in using the adaptively collected data to answer a variety of questions, not all of them necessarily targeted by the design. For example, a company experimenting with many types of web ads may use a bandit algorithm to maximize click-through rates during an experiment, but still want to quantify the effectiveness of each ad. At this stage, fundamental tensions between the experiment objective and statistical inference become apparent: Extreme undersampling or nonconvergence of the assignment probabilities make reusing these data challenging.

In this paper, we propose a method for constructing frequentist CIs based on approximate normality, even when challenges of adaptivity are severe, provided that the treatment-assignment probabilities are known and satisfy certain conditions. To get a better sense of the challenges we face, we'll first consider an example in which traditional approaches to statistical testing fail. Suppose we run a two-stage, two-arm trial as follows. For the first $T/2$ time periods, we randomize assignments with probability 50% for each arm. After $T/2$ time periods, we identify the arm with the higher sample mean, and for the next $T/2$ time periods, we allocate treatment to the seemingly better arm 90% of the time. Then, one estimator of the expected value $Q(w)$ for each arm $w \in \{1, 2\}$ is the sample mean at the end of the experiment,

$$\widehat{Q}^{\mathrm{AVG}}(w) = \frac{1}{T_w} \sum_{\substack{t \leq T \\ W_t = w}} Y_t, \quad T_w := \sum_{\substack{t \leq T \\ W_t = w}} 1, \qquad [1]$$

where W_t denotes the arm pulled in the t-th time period and Y_t denotes the observed outcome. Both arms have the same outcome distribution: $Y_t \mid W_t = w \sim \mathcal{N}(0, 1)$ for all values of t and w.

This example is relatively benign, in that adaptivity is minimal. Yet, as Fig. 1, *Left* shows, the estimate of the value of the first arm $\widehat{Q}^{\mathrm{AVG}}(1)$ is biased downward. This is a well-known phenomenon; see, e.g., refs. 10–16. The downward bias occurs because arms in which we observe random upward fluctuations initially will be sampled more, while arms in which we observe random downward fluctuations initially will be sampled less. The upward fluctuations are corrected as estimates of arms that are sampled more regress to their mean, while the downward ones may not be corrected because of the reduction in sampling. Here, we only show estimates for the first arm, so there are no selection-bias effects; the bias is a direct consequence of the adaptive data collection.

One often-discussed fix to this particular bias problem is to use the inverse-probability weighting estimator, $\widehat{Q}^{\mathrm{IPW}}(w) = T^{-1} \sum_{t=1}^{T} \mathbb{I}\{W_t = w\} Y_t / e_t(w)$, where $e_t(w)$ is the probability with which our adaptive experiment drew arm w in step t. This compensates for the outsize influence of early downward fluctuations that reduce the probability of an arm being assigned

Significance

Randomized controlled trials are central to the scientific process, but they can be costly. For example, a clinical trial may assign patients to treatments that are detrimental to them. Adaptive experimental designs, such as multiarmed bandit algorithms, reduce costs by increasing the probability of assigning promising treatments over the course of the experiment. However, because observations collected by these methods are dependent and their distribution is nonstationary, statistical inference can be challenging. We propose a treatment-effect estimator that has an asymptotically unbiased and normal test statistic under straightforward, relatively weak conditions on the adaptive design. This estimator generalizes for a variety of parameters of interest.

Author contributions: V.H., D.A.H., R.Z., S.W., and S.A. designed research, performed research, contributed analytic tools, analyzed simulated data, and wrote the paper.

Reviewers: M.K., University of Oxford; and J.S., Yale University.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: athey@stanford.edu or vitorh@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2014602118/-/DCSupplemental.

Published April 5, 2021.

STATISTICS | ECONOMIC SCIENCES

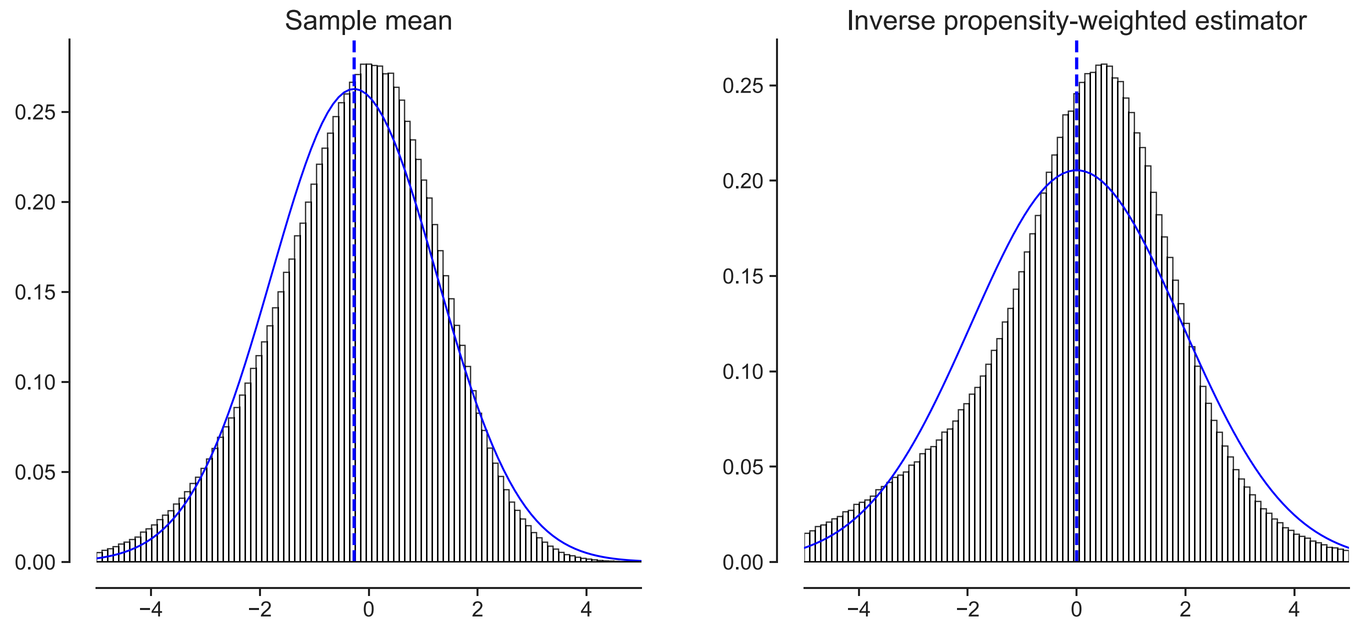

Fig. 1. Distribution of the estimates $\widehat{Q}^{\mathrm{AVG}}(1)$ and $\widehat{Q}^{\mathrm{IPW}}(1)$ described in the introduction. The plots depict the distribution of the estimators for $T = 10^6$, scaled by a factor $\sqrt{T}$ for visualization. The distributions are overlaid with the normal curve that matches the first two moments of the distribution, along with a dashed line that denotes the mean. All numbers are aggregated over 1 million replications.

by up-weighting later observations within that arm when we see them. As seen in Fig. 1, *Right*, inverse-probability weighting fixes the bias problem, but results in a nonnormal asymptotic distribution. In fact, it can exacerbate the problem of inference: When the probability of assignment to the arm of interest tends to zero, the inverse probability weights increase, which, in turn, causes the tails of the distribution to become heavier.

If adaptivity essentially vanishes over the experiment, then we are justified in using naive estimators that ignore adaptivity. In particular, if assignment probabilities quickly converge to constants, then, in long-running experiments, we can often treat the data as if treatments had always been assigned according to these limiting probabilities; see, e.g., ref. 17 or 9. It can be argued that most adaptive designs would eventually converge in this sense if run forever. However, how quickly they converge, and, therefore, the number of samples we'd need for adaptivity to be ignorable can vary considerably with the data-generating process. For example, we know that so long as some arm is best, an ϵ-greedy K-armed bandit algorithm will eventually assign to the best arm with probability $1 - \epsilon + \epsilon/K$ and the others with probability ϵ/K; however, this happens only after the best arm is identified, which depends strongly on the unknown spacing between the arm values $Q(1), \ldots, Q(K)$. Moreover, in practice, adaptive experiments are often used precisely when there is limited budget for experimentation; therefore, substantial data collection after convergence is rare. As a result, if we try to exploit this convergence by using estimators that are only valid in convergent designs, we get brittle estimators. If we want an estimator that is reliable, we must use one that is valid, regardless of whether the assignment probabilities converge.

In this work, we propose a test statistic that is asymptotically unbiased and normal, even when assignment probabilities converge to zero or do not converge at all.* We believe this approach to be of practical interest because normal CIs are widely used in several fields, including, e.g., medicine and economics. Moreover, though we focus on estimating the value of a prespecified policy, our estimates can also be used as input to procedures for testing adaptive hypotheses, which have as their starting point a vector of normal estimates (e.g., ref. 18).

Other approaches to inference with adaptively collected data are available. One line of research eschews asymptotic normality in favor of developing finite-sample bounds using martingale concentration inequalities (e.g., refs. 19 and 20, and references therein). Ref. 10 considers approaches to debiasing value estimates using ideas from conditional inference (21). And some avoid frequentist arguments altogether, preferring a purely Bayesian approach, although this can produce estimates that have poor frequentist properties (22). *Related Literature* further reviews papers on policy evaluation.

Policy Evaluation with Adaptively Collected Data

We start by establishing some definitions. Each observation in our data is represented by a tuple (W_t, Y_t). The random variables $W_t \in \mathcal{W}$ are called the arms, treatments, or interventions. Arms are categorical. The reward or outcome Y_t represents the individual's response to the treatment. The set of observations up to a certain time $H^T := \{(W_s, Y_s)\}_{s=1}^T$ is called a history. The treatment-assignment probabilities $e_t(w) := \mathbb{P}[W_t = w \mid H^{t-1}]$, also called propensity scores, are time-varying and decided via some known algorithm, as it is the case with many popular bandit algorithms, such as Thompson sampling (23, 24).

We define causal effects using potential outcome notation (25). We denote by $Y_t(w)$ the random variable representing the outcome that would be observed if individual t were assigned to a treatment w. In any given experiment, this individual can be assigned only one treatment, W_t, from a set of options $\mathcal{W}$, so we observe only one realized outcome $Y_t = Y_t(W_t)$. We focus on the "stationary" setting, where individuals, represented by a vector of potential outcomes $(Y_t(w))_{w \in \mathcal{W}}$, are independent and identically distributed. However, the observed outcomes Y_t are, in general, neither independent nor identically

*In *SI Appendix*, section A.5, we revisit the example shown in the introduction and demonstrate how our method leads to an asymptotically normal test statistic for the arm value.

distributed, because the distribution of the treatment assignment W_t depends on the history of outcomes up to time t.

Given this setup, we are concerned with the problem of estimating and testing prespecified hypotheses about the value of an arm, denoted by $Q(w) := \mathbb{E}\left[Y_t(w)\right]$, as well as differences between two such values, denoted by $\Delta(w, w') := \mathbb{E}\left[Y_t(w)\right] - \mathbb{E}\left[Y_t(w')\right]$. We would like to do that even in data-poor situations, in which the data-collection mechanism did not target these estimands.

We will provide consistent and asymptotically normal test statistics for $Q(w)$ and $\Delta(w, w')$. This is done in three steps. First, we start with a class of *scoring rules*, which are transformations of the observed outcomes that can be used for unbiased arm evaluation, but whose sampling distribution can be nonnormal and heavy-tailed due to adaptivity. Second, we average these objects with carefully chosen data-adaptive weights, obtaining an estimator with controlled variance at the cost of some finite-sample small bias. Finally, by dividing these estimators by their SE, we obtain a test statistic that has a centered and standard normal limiting distribution.

Unbiased Scoring Rules. A first step in developing methods for inference with adaptive data is to account for sampling bias. The following construction provides a generic way of doing so. We say that $\widehat{\Gamma}_t(w)$ is an unbiased scoring rule for $Q(w)$ if for all $w \in \mathcal{W}$ and $t = 1, \ldots, T$,

$$\mathbb{E}\left[\widehat{\Gamma}_t(w) \mid H^{t-1}\right] = Q(w). \quad [2]$$

Given this definition, we can readily verify that a simple average of such a scoring rule,

$$\widehat{Q}_T(w) = \frac{1}{T}\sum_{i=1}^{T}\widehat{\Gamma}_t(w), \quad [3]$$

is unbiased for $Q(w)$, even though the $\widehat{\Gamma}_t(w)$ are correlated over time, as the next proposition shows.

Proposition 1. *Let $\{Y_t(w)\}_{w\in\mathcal{W}}$ be an independent and identically distributed sequence of potential outcomes for $t = 1, \ldots, T$, and let H^t denote the observation history up to time t, as described above. Then, any estimator of the form* [**3**] *based on an unbiased scoring rule* [**2**] *satisfies $\mathbb{E}[\widehat{Q}_T(w)] = Q(w)$.*

One can easily verify Proposition 1 by applying the law of iterated expectations and [**2**],

$$\mathbb{E}\left[\widehat{Q}_T(w)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\widehat{\Gamma}_t(w) \mid H^{t-1}\right]\right] = Q(w).$$

The key fact underlying this result is that the normalization factor $1/T$ used in [**3**] is deterministic, and so cannot be correlated with stochastic fluctuations in the $\widehat{\Gamma}_t(w)$. In particular, we note that the basic averaging estimator [**1**] is not of the form [**3**] and, instead, has a random denominator T_w—and is thus not covered by Proposition 1.

Given Proposition 1, we can readily construct several unbiased estimators for $Q(w)$. One straightforward option is to use an inverse propensity score weighted (IPW) estimator:

$$\widehat{Q}^{IPW}_T(w) := \frac{1}{T}\sum_{t=1}^{T}\widehat{\Gamma}^{IPW}_t(w), \quad \widehat{\Gamma}^{IPW}_t(w) := \frac{\mathbb{I}\{W_t = w\}}{e_t(w)}Y_t. \quad [4]$$

This estimator is simple to implement, and one can directly check that the condition [**2**] holds because, by construction, $\mathbb{P}\left[W_t = w \mid H^{t-1}, Y_t(w)\right] = \mathbb{P}\left[W_t = w \mid H^{t-1}\right] = e(w; H^{t-1})$.

The augmented inverse propensity weighted (AIPW) estimator generalizes this by incorporating regression adjustment (26):

$$\begin{aligned}\widehat{Q}^{AIPW}_T(w) &:= \frac{1}{T}\sum_{t=1}^{T}\widehat{\Gamma}^{AIPW}_t(w),\\ \widehat{\Gamma}^{AIPW}_t(w) &:= \frac{\mathbb{I}\{W_t = w\}}{e_t(w)}Y_t + \left(1 - \frac{\mathbb{I}\{W_t = w\}}{e_t(w)}\right)\hat{m}_t(w).\end{aligned} \quad [5]$$

The symbol $\hat{m}_t(w)$ denotes an estimator of the conditional mean function $m(w) = \mathbb{E}[Y_t(w)]$ based on the history H^{t-1}, but it need not be a good one—it could be biased, or even inconsistent. The second term of $\widehat{\Gamma}^{AIPW}_t(w)$ acts as a control variate: Adding it preserves unbiasedness, but can reduce variance, as it has mean zero conditional on H^{t-1} and, if $\hat{m}_t(w)$ is a reasonable estimator of $m(w)$, is negatively correlated with the first term. When $\hat{m}_t(w)$ is identically zero, the AIPW estimator reduces to the IPW estimator.

Asymptotically Normal Test Statistics. The estimators discussed in the previous section are unbiased by construction, but, in general, they are not guaranteed to have an asymptotically normal sampling distribution. The reason for this failure of normality was illustrated in the example in the introduction for the IPW estimator. When, purely by chance, an arm has a higher sample mean in the first half of the experiment, it is sampled often in the second half, and the IPW estimator concentrates tightly. When the opposite happens, the arm is sampled less frequently, and the IPW estimator is more spread out. What we see in Fig. 1 is a heavy-tailed distribution corresponding to the mixture of the two behaviors. Qualitatively, what we need for normality is for the variability of the estimator to be deterministic. Formally, what is required is that the sum of conditional variances of each term in the sequence converges in ratio to the unconditional variance of the estimator (see e.g., ref. 27, theorem 3.5). Simple averages of unbiased scoring rules [**1**] fail to satisfy this because, as we'll elaborate below, the conditional variances of the terms in [**5**] depend primarily on the behavior of the inverse assignment probabilities $1/e_t(w)$, which may diverge to infinity or fail to converge.

To address this difficulty, we consider a generalization of the AIPW estimator [**5**] that nonuniformly averages the unbiased scores $\widehat{\Gamma}^{AIPW}$ using a sequence of *evaluation weights* $h_t(w)$. The resulting estimator is the *adaptively weighted AIPW estimator*:

$$\widehat{Q}^{h}_T(w) = \frac{\sum_{t=1}^{T} h_t(w)\widehat{\Gamma}^{AIPW}_t(w)}{\sum_{t=1}^{T} h_t(w)}. \quad [6]$$

Evaluation weights $h_t(w)$ provide an additional degree of flexibility in controlling the variance and tails of the sampling distribution. When chosen appropriately, these weights compensate for erratic trajectories of the assignment probabilities $e_t(w)$, stabilizing the variance of the estimator. With such weights, the adaptively weighted AIPW estimator [**6**], when normalized by an estimate of its SD, has a centered and normal asymptotic distribution. Similar "self-normalization" schemes are often key to martingale central-limit theorems (see e.g., ref. 28).

Throughout, we will use evaluation weights $h_t(w)$ that are a function of the history H^{t-1}; we will call such functions *history-adapted*. Note that if we used weights with sum equal to one, we would have a generalization of the unbiased scoring property [**3**], $\mathbb{E}[h_t(w)\widehat{\Gamma}_t(w) \mid H^{t-1}] = h_t(w)Q(w)$, so the adaptively weighted estimator [**6**] would be unbiased. In *Constructing Adaptive Weights*, we will discuss weight heuristics that do not sum to one, but that empirically seem to reduce variance and

https://doi.org/10.1073/pnas.2014602118

mean-squared error relative to alternatives. In that case, the estimator [6] will have some bias due to the random denominator $\sum_{t=1}^{T} h_t(w)$. However, for the appropriate choices of evaluation weights $h_t(w)$, this bias disappears asymptotically.

The main conditions required by our weighting scheme are stated below. Assumption 1 requires that the effective sample size after adaptive weighting—that is, the ratio $(\sum_{t=1}^{T} \mathbb{E}[\alpha_t | H^{t-1}])^2 / \mathbb{E}[\sum_{t=1}^{T} \alpha_t^2]$ where $\alpha_t := h_t(w) \mathbb{I}\{W_t = w\} / e_t(w)$—goes to infinity. This implies that the estimator converges. Assumption 2 is the more subtle condition that unbiased estimators such as [3] [i.e., estimators with $h_t(w) \equiv 1$] often fail to satisfy. Assumption 3 is a Lyapunov-type regularity condition on the weights controlling higher moments of the distribution.

Assumption 1 (Infinite Sampling). *The weights used in* [6] *satisfy*

$$\left(\sum_{t=1}^{T} h_t(w) \right)^2 \Big/ \mathbb{E}\left[\sum_{t=1}^{T} \frac{h_t^2(w)}{e_t(w)} \right] \xrightarrow[T \to \infty]{p} \infty. \qquad [7]$$

Assumption 2 (Variance Convergence). *The weights used in* [6] *satisfy, for some* $p > 1$,

$$\sum_{t=1}^{T} \frac{h_t^2(w)}{e_t(w)} \Big/ \mathbb{E}\left[\sum_{t=1}^{T} \frac{h_t^2(w)}{e_t(w)} \right] \xrightarrow[T \to \infty]{L_p} 1. \qquad [8]$$

Assumption 3 (Bounded Moments). *The weights used in* [6] *satisfy, for some* $\delta > 0$,

$$\sum_{t=1}^{T} \frac{h^{2+\delta}(w)}{e^{1+\delta}(w)} \Big/ \mathbb{E}\left[\sum_{t=1}^{T} \frac{h_t^2(w)}{e_t(w)} \right]^{1+\delta/2} \xrightarrow[T \to \infty]{p} 0. \qquad [9]$$

Theorem 2. *Suppose that we observe arms* W_t *and rewards* $Y_t = Y_t(W_t)$*, and that the underlying potential outcomes* $(Y_t(w))_{w \in \mathcal{W}}$ *are independent and identically distributed with nonzero variance, and satisfy* $\mathbb{E}|Y_t(w)|^{2+\delta} < \infty$ *for some* $\delta > 0$ *and all* w. *Suppose that the assignment probabilities* $e_t(w)$ *are strictly positive, and let* $\hat{m}_t(w)$ *be any history-adapted estimator of* $Q(w)$ *that is bounded and that converges almost-surely to some constant* $m_\infty(w)$. *Let* $h_t(w)$ *be nonnegative, history-adapted weights satisfying Assumptions* 1, 2, and 3. *Suppose that either* $\hat{m}_t(w)$ *is consistent or* $e_t(w)$ *has a limit* $e_\infty(w) \in [0, 1]$, *i.e., either*

$$\hat{m}_t(w) \xrightarrow[t \to \infty]{a.s.} Q(w) \quad \text{or} \quad e_t(w) \xrightarrow[t \to \infty]{a.s.} e_\infty(w). \qquad [10]$$

Then, for any arm $w \in \mathcal{W}$, *the adaptively weighted estimator* [6] *is consistent for the arm value* $Q(w)$, *and the following studentized statistic is asymptotically normal*:

$$\frac{\widehat{Q}_T^h(w) - Q(w)}{\widehat{V}_T^h(w)^{\frac{1}{2}}} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{where}$$

$$\widehat{V}_T^h(w) := \frac{\sum_{t=1}^{T} h_t^2(w) \left(\widehat{\Gamma}_t(w) - \widehat{Q}_T(w) \right)^2}{\left(\sum_{t=1}^{T} h_t(w) \right)^2}. \qquad [11]$$

As Theorem 2 suggests, the asymptotic behavior of our estimator is largely determined by the behavior of the propensity scores $e_t(w)$ and evaluation weights $h_t(w)$. If the former's behavior is problematic, the latter can correct for that. For instance, the bounded-moments condition [9] implies that when $e_t(w)$ decays very fast, the evaluation weights must also decay at an appropriate rate, so that the variance of the estimator does not explode. However, there are limits to what this approach can correct. For example, aggressive bandit procedures may assign some arms only finitely many times, and, in that case, it is impossible to estimate their values consistently. This scenario is ruled out by our infinite-sampling condition [7], which would not be satisfied.

To build more intuition for the variance-convergence condition [8], pretend for the moment that the evaluation weights $h_t(w)$ sum to one and that $\hat{m}_t(w)$ is consistent. Under these conditions, the variance of each AIPW score $\widehat{\Gamma}_t(w)$ conditional on the past can be shown to be $\mathrm{Var}[Y_1(w)]/e_t(w)$ plus asymptotically negligible terms, so the sum of conditional variances is asymptotically equivalent to $\mathrm{Var}[Y_1(w)] \sum_{t=1}^{T} h_t^2(w)/e_t(w)$. The variance-convergence condition [8] says that this sum of conditional variances converges to its mean, which will be the unconditional variance of the estimator. If these simplifying assumptions really held, this argument would almost suffice to establish asymptotic normality, since variance-convergence conditions like this are nearly sufficient for martingale central-limit theorems (see, e.g., ref. 27, theorem 3.5).

When we use history-dependent weights $h_t(w)$ that do not sum to one, the normalized weights $\tilde{h}_t := h_t(w) / \sum_{s=1}^{T} h_s(w)$ that scale our terms in [6] are not a function of past data alone. However, the argument above provides valuable intuition in our proof, and $\mathrm{Var}[Y_1(w)] \sum_{t=1}^{T} \tilde{h}_t^2 / e_t(w)$ can be thought of as a reasonable proxy for the estimator's variance. Minimizing this variance proxy suggests the use of weights $h_t(w) \propto e_t(w)$. This heuristic, in combination with appropriate constraints, motivates an empirically successful weighting scheme that we'll further discuss in *Constructing Adaptive Weights*.

If the weights $h_t(w)$ are not constructed appropriately, then $h_t^2(w)/e_t(w)$ may behave erratically, and the variance-convergence condition will fail to hold. This can happen, for example, in a bandit experiment, in which there are multiple optimal arms, and uniform weights $h_t(w) \equiv 1$ are used. In this setting, the bandit algorithm may spuriously choose one arm at random early on and assign the vast majority of observations to it, so that no run of the experiment will look like an "average run," and the ratio in [8] will not converge. Or it may switch between arms infinitely often, in which case the ratio will converge only if it switches quickly enough that the random order is "forgotten" in the average. This issue persists even when assignment probabilities are guaranteed to stay above a strictly positive lower bound. In the next section, we will construct weights so that "variance-convergence" [8] is guaranteed to be satisfied not only asymptotically, but for all sample sizes T.

The simplifying assumption that $\hat{m}_t(w)$ is consistent is not necessary: As stated in [10], a variant of the argument above still goes through if the propensity score $e_t(w)$ converges. In a nonadaptive experiment, the AIPW estimator will have optimal asymptotic variance if $\hat{m}_t(w)$ is consistent; if it is not, the excess asymptotic variance is a function of $\lim_{t \to \infty} e_t(w)$ and $\lim_{t \to \infty} \hat{m}_t(w) - Q(w)$.

Constructing Adaptive Weights. A natural question is how to choose evaluation weights $h_t(w)$ for which the adaptively weighted AIPW estimator [6] is asymptotically unbiased and normal with low variance, i.e., for which we get narrow and approximately valid CIs. To this end, we'll start by focusing on the variance-convergence condition [8]. Once we have a recipe for building weights that satisfy it, we'll consider how to satisfy the other conditions of Theorem 2 and how to optimize for power.

The variance-convergence condition [8] requires the sum $\sum_{t=1}^{T} h_t^2/e_t$ to concentrate around its expectation. A direct way to ensure this is to make the sum deterministic. To do this, we choose weights via a recursively defined "stick-breaking" procedure,[†]

$$\frac{h_t^2}{e_t} = \left(1 - \sum_{s=1}^{t-1} \frac{h_s^2}{e_s}\right)\lambda_t, \qquad [12]$$

where λ_t satisfies $0 \le \lambda_t < 1$ for all $1 \le t \le T-1$, and $\lambda_T = 1$. Because $\lambda_T = 1$, the definition above for $t = T$ directly implies that $\sum_{t=1}^{T} h_t^2/e_t = 1$. This ensures that the variance-convergence condition **[8]** is satisfied, so we call these *variance-stabilizing* weights.

We call the function λ_t an *allocation rate* because it qualitatively captures the fraction of our remaining variance that we allocate to the upcoming observation. This is a useful class to consider because the analyst has substantial freedom in constructing weights by choosing different allocation rates λ_t, while ensuring that the resulting evaluation weights automatically satisfy the variance-convergence assumption, and satisfy other assumptions of Theorem 2 with some generality.

Theorem 3. *In the setting of Theorem* 2, *suppose that the treatment propensities satisfy*

$$e_t(w) \ge Ct^{-\alpha}, \qquad [13]$$

for $\alpha \in [0,1)$ *and any positive constant* C. *Then, the variance-stabilizing weights* **[12]** *defined by a history-adapted allocation rate* $\lambda_t(w)$ *are history-adapted and satisfy Assumptions* 1, 2, *and* 3 *if* $\lambda_t(w) < 1$ *for* $t < T$, $\lambda_T(w) = 1$ *and, for a finite positive constant* C',

$$\frac{1}{T-t+1} \le \lambda_t(w) \le C' \frac{e_t(w)}{t^{-\alpha} + T^{1-\alpha} - t^{1-\alpha}}. \qquad [14]$$

The main requirement of Theorem 3 is **[13]**, a limit on the rate at which treatment-assignment propensities e_t decay. In a bandit setting, this constraint requires that suboptimal arms be pulled more often than implied by rate-optimal algorithms (see e.g., ref. 29, chapter 15), but still allows for sublinear regret. Given this constraint, the allocation-rate bounds **[14]** are weak enough to allow us to construct variance-reducing heuristics, like **[18]** below.

Given these simple sufficient conditions for our asymptotic normality result (Theorem 2) when we use variance-stabilizing weights, it remains to choose a specific allocation rate λ_t. This next step is what will allow us to be able to provide valid estimates, even when the share of relevant data vanishes asymptotically. A simple choice of allocation rate is

$$\lambda_t^{\text{const}} := \frac{1}{T-t+1}. \qquad [15]$$

Given this choice, we can solve **[12]** in closed form and get $h_t = \sqrt{e_t/T}$. Weights of this type were proposed by ref. 30 for the purpose of estimating the expected value of nonunique optimal policies that possibly depend on observable covariates. We call this method the *constant-allocation scheme*, because the variance contribution of each observation is constant (since $h_t^2/e_t \equiv 1/T$ for these weights).

The constant-allocation scheme guarantees the variance-convergence condition **[8]** and ensures asymptotic normality of the test statistic **[11]**, but it does not result in a variance-optimal estimator. We propose an alternative scheme in which λ_t adapts to past data and reweights observations to better control the estimator's variance. To get some intuition, recall from the discussion following Theorem 2 that the variance of $\widehat{Q}_T^h(w)$ essentially scales like $\sum_t (h_t^2/e_t)/(\sum_{t=1}^T h_t)^2$. This implies that, in the absence of any constraints on how we choose the weights, we would minimize variance by setting $h_t \propto e_t$; this can be accomplished by using the allocation rate $\lambda_t = e_t / \sum_{s=t}^T e_s$. If we use these weights and set $\hat{m}_t \equiv 0$ in **[6]**, the result is an estimator that differs from the sample average **[1]** only in that it replaces the normalization $1/T_w$ with $1/\sum_{t=1}^T e_t$. Our results do not apply to this choice of allocation rate λ_t because it depends on future treatment-assignment probabilities, and Theorem 3 requires that λ_t depend only on the history H^{t-1}.

However, this form of allocation rate suggests a natural heuristic choice of allocation rate:

$$\lambda_t = \widehat{\mathbb{E}}_{t-1}\left[\frac{e_t(w)}{\sum_{s=t}^T e_s(w)}\right], \qquad [16]$$

where $\widehat{\mathbb{E}}_{t-1}$ denotes an estimate of the future behavior of the propensity scores using information up to the beginning of the current period. It can be estimated via Monte Carlo methods. A high-quality approximation is unnecessary for valid inference. All that is required is that the allocation-rate bounds **[14]** be satisfied, although better approximations likely lead to better statistical efficiency.

In practice, the need to compute these estimates renders the construction **[16]** unwieldy. Furthermore, the way the resulting weights depend on our model of the assignment mechanism is fairly opaque. As an alternative, we consider a simple heuristic that exhibits similar behavior and can be used when assignment probabilities are decided via Bayesian methods, such as Thompson sampling.

To derive our scheme, we consider two scenarios: one in which the assignment probabilities e_t are currently high and will continue being so in the future, as is the case when a bandit algorithm deems w to be an optimal arm; and a second scenario, in which the assignment probabilities will asymptotically decay toward zero as fast as the lower bound **[13]** permits it to. If the first scenario is true, then we could approximate the behavior of **[16]** by setting $e_s = 1$ for all periods. If the second scenario is true, then we could do so by setting $e_s = Cs^{-\alpha}$ for all periods. Letting A be an indicator that we are in the first scenario, we can consider a heuristic approximation to **[16]**:

$$\lambda_t \approx A \frac{1}{\sum_{s=t}^T 1} + (1-A)\frac{t^{-\alpha}}{\sum_{s=t}^T s^{-\alpha}}. \qquad [17]$$

Of course, we don't know which scenario we are in. However, when assigning treatment via Thompson sampling, e_t is the posterior probability at time t that arm w is optimal. This suggests the heuristic of averaging the two possibilities according to this posterior probability. Substituting, in addition, an integral approximation to $\sum_{s=t+1}^T s^{-\alpha}$, we get the following allocation rate:

$$\lambda_t^{\text{two-point}} := e_t \frac{1}{T-t+1} + (1-e_t)\frac{t^{-a}}{t^{-\alpha} + \frac{T^{1-\alpha} - t^{1-\alpha}}{1-\alpha}}. \qquad [18]$$

We call this the *two-point allocation scheme*. Both the constant **[15]** and two-point **[18]** allocation schemes satisfy the allocation-rate bounds **[14]** from Theorem 3; see *SI Appendix*, section C.5.

Estimating Treatment Effects

Our discussion so far has focused on estimating the value $Q(w)$ of a single arm w. In many applications, however, we may seek to

[†]For notational efficiency, whenever it does not lead to confusion, we will drop the dependence on arm and write, e.g., $\widehat{Q}_T$ simply to mean our adaptively weighted estimator [6], h_t for evaluation weights, e_t for assignment probabilities, and so on.

https://doi.org/10.1073/pnas.2014602118

provide inference for a wider variety of estimands, starting with treatment effects of the form $\Delta(w_1, w_2) = \mathbb{E}[Y_t(w_1) - Y_t(w_2)]$. There are two natural ways to approach this problem in our framework. The first involves revisiting our discussion from *Unbiased Scoring Rules*, and directly defining unbiased scoring rules for $\Delta(w_1, w_2)$ that can then be used as the basis for an adaptively weighted estimator. The second is to reuse the value estimates derived above and set $\widehat{\Delta}(w_1, w_2) = \widehat{Q}(w_1) - \widehat{Q}(w_2)$; the challenge then becomes how to provide uncertainty quantification for $\widehat{\Delta}(w_1, w_2)$. We discuss both approaches below.

In the first approach, we use the difference in AIPW scores as the unbiased scoring rule for $\Delta(w_1, w_2)$.

$$\begin{aligned} &\widehat{\Gamma}_t(w_1, w_2) = \widehat{\Gamma}_t^{AIPW}(w_1) - \widehat{\Gamma}_t^{AIPW}(w_2), \\ &\mathbb{E}\left[\widehat{\Gamma}_t(w_1, w_2) \,\middle|\, H^{t-1}\right] = \Delta(w_1, w_2). \end{aligned} \qquad [19]$$

One can then construct asymptotically normal estimates of $\Delta(w_1, w_2)$ by adaptively weighting the scores $\widehat{\Gamma}_t(w_1, w_2)$ as in [**6**]. In *SI Appendix*, section B, our main formal result allows for adaptively weighted estimation of general targets, such that both Theorem 2 and adaptively weighted estimation with scores [**19**] are special cases of this result.[‡]

The second approach is conceptually straightforward; however, we still need to check that the estimator $\widehat{\Delta}(w_1, w_2) = \widehat{Q}(w_1) - \widehat{Q}(w_2)$ can be used for asymptotically normal inference about $\Delta(w_1, w_2)$. Theorem 4 provides such a result, under a modified version of the conditions of Theorem 2, along with an extra assumption [**21**].

Theorem 4. *In the setting of Theorem* 2, *let* $w_1, w_2 \in \mathcal{W}$ *denote a pair of arms*, *and suppose that Assumptions* 1, 2, *and* 3 *are satisfied for both arms*. *In addition*, *suppose that the variance estimates defined in* [**11**] *satisfy*

$$\widehat{V}_T^h(w_1) \,/\, \widehat{V}_T^h(w_2) \xrightarrow[T\to\infty]{p} r \in [0, \infty], \qquad [21]$$

and that for at least one $j \in \{1, 2\}$, *either*

$$\hat{m}_t(w_j) \xrightarrow[t\to\infty]{a.s.} Q(w_j) \quad \text{or} \quad e_t(w_j) \xrightarrow[t\to\infty]{a.s.} 0. \qquad [22]$$

Then, *the vector of studentized statistics* [**11**] *for* w_1 *and* w_2 *is asymptotically jointly normal with identity covariance matrix*. *Moreover*, $\widehat{\Delta}_T(w_1, w_2)$ *below is a consistent estimator of* $\Delta(w_1, w_2) = \mathbb{E}[Y_t(w_1) - Y_t(w_2)]$,

$$\widehat{\Delta}_T(w_1, w_2) := \frac{\sum_{t=1}^T h_t(w_1)\widehat{\Gamma}_t(w_1)}{\sum_{t=1}^T h_t(w_1)} - \frac{\sum_{t=1}^T h_t(w_2)\widehat{\Gamma}_t(w_2)}{\sum_{t=1}^T h_t(w_2)}, \qquad [23]$$

and the following studentized statistic is asymptotically standard normal.

$$\frac{\widehat{\Delta}_T(w_1, w_2) - \Delta(w_1, w_2)}{\left(\widehat{V}_T^h(w_1) + \widehat{V}_T^h(w_2)\right)^{1/2}} \xrightarrow[T\to\infty]{d} \mathcal{N}(0, 1). \qquad [24]$$

Both approaches to inference about $\Delta(w_1, w_2)$ are of interest, and may be relevant in different settings. In our experiments, we focus on the estimator $\widehat{\Delta}(w_1, w_2) = \widehat{Q}(w_1) - \widehat{Q}(w_2)$ studied in Theorem 4, as we found it to have higher power—presumably because allowing separate weights $h_t(w)$ for different arms gives us more control over the variance. However, adaptively weighted estimators following [19] that directly target the difference $\Delta(w_1, w_2)$ may also be of interest in some applications. In particular, they render an assumption like [**21**] unnecessary and, following the line of argumentation in ref. 32, they may be more robust to nonstationarity of the distribution of the potential outcomes $Y_t(w)$.

Numerical Experiments

We compare methods for estimating of arm values $Q(w)$ and their differences $\Delta(w, w')$, as well as constructing CIs around these estimates, under different data-generating processes.

We consider four point estimators of arm values $Q(w)$: the sample mean [**1**], the AIPW estimator [**5**], and the adaptively weighted AIPW estimator [**6**] with constant [**15**] and two-point [**18**] allocation rates. Around each of these estimators, we construct CIs $\widehat{Q}_T \pm z_{\alpha/2}\widehat{V}_T^{1/2}$ based on the assumption of approximate normality. For the AIPW-based estimators,[§] we use the sample mean of arm rewards up to period $t-1$ as the plug-in estimator $\hat{m}_t(w)$ and the estimate of the variance given in [**11**]. For the sample mean, we use the usual variance estimate $\widehat{V}^{AVG}(w) := T_w^{-2}\sum_{t:\, W_t=w}^T (Y_t - \widehat{Q}_T^{AVG}(w))^2$. Approximate normality is not theoretically justified for the unweighted AIPW estimator or for the sample mean. We also consider nonasymptotic CIs for the sample mean, based on the method of time-uniform confidence sequences described in ref. 19. See *SI Appendix*, section D.2 for details.

In addition, we consider four analogous estimators for the treatment effect $\Delta(w, w')$: the difference in sample means, the difference in AIPW estimators [**5**], and the adaptively weighted AIPW estimator [**23**] with constant [**15**] and two-point [**18**] allocation rates. For the AIPW-based estimators, we use $\hat{m}_t(w)$ as above for the plug-in and the estimate of the variance given in [**24**]. For the sample mean, we use the usual variance estimate $\widehat{V}^{AVG}(w) + \widehat{V}^{AVG}(w')$. For the method based on ref. 19, we construct CIs for the treatment effect by using the union bound to combine intervals around each sample mean.

We have three simulation settings, each with $K = 3$ arms, yielding rewards that are observed with additive uniform$[-1, 1]$ noise. The settings differ in the size of the gap between the arm values. In the *no-signal* case, we set arm values to $Q(w) = 1$ for all $w \in \{1, 2, 3\}$; in the *low-signal* case, we set $Q(w) = 0.9 + 0.1w$; and in the *high-signal* case, we set $Q(w) = 0.5 + 0.5w$. During the experiment, treatment is assigned by a modified Thompson sampling method (see, e.g., ref. 24): First, tentative assignment probabilities are computed via Thompson sampling with normal likelihood and normal prior; they are then adjusted to impose the lower bound $e_t(w) \geq (1/K)t^{-0.7}$. See *SI Appendix*, section D.2 for details.

As a short mnemonic, in what follows, we call arms 1 and 3 the "bad" arm and "good" arm, respectively. As these labels are fixed, tests involving the value of the good arm are tests of a

[‡] Our result allows for considerably more generality than either of the cases discussed above and applies whenever our target admits a doubly robust estimator in the sense of ref. 31, whose Riesz representer is a function of the treatment-assignment mechanism. For example, consider an adaptive clinical trial setting in which patients were given random doses of a continuous treatment W_t drawn from a time-varying dosing policy $f_t(w)$, i.e., W_t is a random variable with density $f_t(w)$, and write $m(w) = \mathbb{E}[Y_t(w)]$. Now, suppose that, given a specific treatment-assignment policy with density $f(w)$, we are interested in estimating $\psi(m) = \int m'(w)f(w)dw$, i.e., how patients' outcomes would change if they received doses in slightly larger amounts than those suggested by the baseline policy $f(w)$. An unbiased scoring rule for this estimand is

$$\widehat{\Gamma}_t = \gamma_t Y_t + \left(\int \hat{m}'(w)f(w)dw - \gamma_t\hat{m}(W_t)\right) \quad \text{where} \quad \gamma_t = -\frac{f'(W_t)}{f_t(W_t)}, \qquad [20]$$

and our results apply to inference about $\psi(m)$ by adaptively weighted aggregation of these $\widehat{\Gamma}_t$. See *SI Appendix*, section B for further discussion.

[§] Recall that the AIPW estimator [**5**] is an instance of the adaptively weighted AIPW [**6**] with uniform weights $h_t \equiv 1$, so we may use the same formula for the variance [**11**].

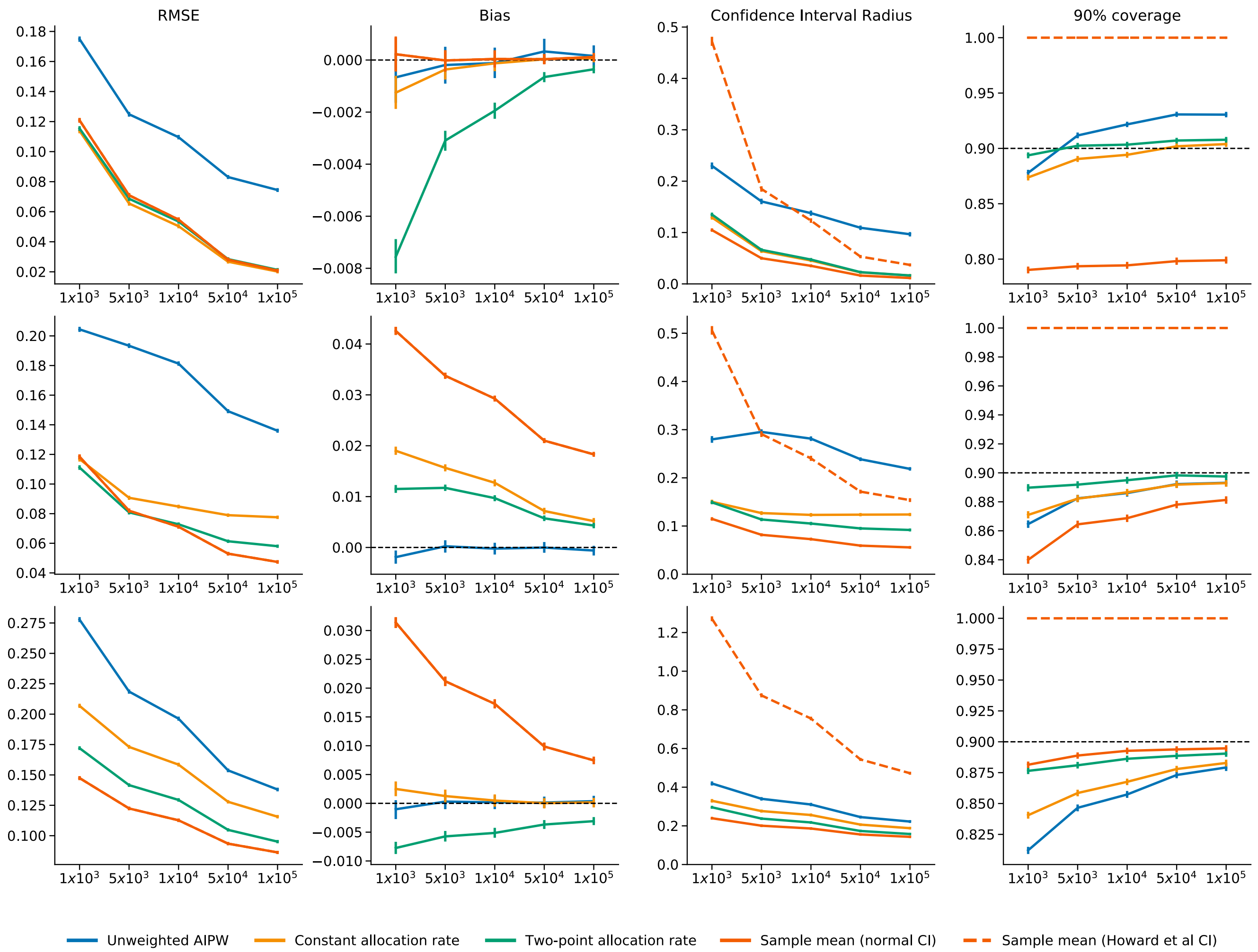

Fig. 2. Evolution of estimates of $Q(3) - Q(1)$ across simulation settings for different experiment lengths. Error bars are 95% CIs around averages across 10^5 replications.

prespecified hypothesis. Fig. 2 shows the evolution of estimates of the difference $\Delta(3,1) = \mathbb{E}[Y(3) - Y(1)]$ over time; Fig. 3 shows the asymptotic distribution of studentized test statistics of the form $(\widehat{\Delta}_T - \Delta)/\widehat{V}_T^{1/2}$ for each estimator at the end of a long ($T = 10^5$) experiment; and Fig. 4 shows arm-value statistics. Additional results are shown in *SI Appendix*, section A.[¶]

Figs. 2 and 4 show that, although the AIPW estimator with uniform weights (labeled as "unweighted AIPW") is unbiased, it performs very poorly in terms of root-mean-square error (RMSE) and CI width. In the low- and high-signal case, its problem is that it does not take into account the fact that the bad arm is undersampled, so its variance is high; in the no-signal case, it yields studentized statistics that are far from normal, as we see in Fig. 3.

Figs. 2 and 4 show that our adaptively weighted AIPW estimators perform relatively well, and normal CIs around them have roughly correct coverage. We see that these estimators do have approximately normal studentized statistics in Fig. 3. Note that even in our longest experiments, in high-signal settings, the bad arm receives only around 50 observations, which suggests that normal approximation does not require an impractical number of observations.[‖] Of these two methods, *two-point allocation* better controls the variance of bad-arm estimates by more aggressively downweighting "unlikely" observations associated with large inverse propensity weights; this results in smaller RMSE and tighter CIs.

As mentioned in the introduction, the sample mean is downwardly biased for arms that are undersampled. Fig. 4 shows that this bias can be nonmonotonic in signal strength. In the *high-signal* case, the probability of pulling the bad arm decays so fast that very few observations are assigned to it. Most of these come from the beginning of the experiment, when the algorithm is still exploring, and sampling is less adaptive, resulting in smaller bias. In the *no-signal case*, the bias is small because the good and bad arms have the same value. In some simulations, one arm or another is discarded, and its estimate is biased downward; in others, it is collected heavily, and its estimate is nearly unbiased. Averaging over these scenarios results in the low bias we observe. Between these extremes, in the *low-signal* case, the bad arm is usually collected for some period and then discarded, so its bias is larger in magnitude. For this estimator, naive CIs based on the normal approximation are invalid, with severe

[¶]Reproduction code can be found at https://github.com/gsbDBI/adaptive-confidence-intervals.

[‖]For intuition about the number of observations we see in the bad arm in the high-signal case, consider that the assignment probabilities of suboptimal arms quickly hit their lower bound $e_t = (1/3)t^{-0.7}$; $(1/3)\sum_{t=1}^{10^5} t^{-0.7} \approx 35$.

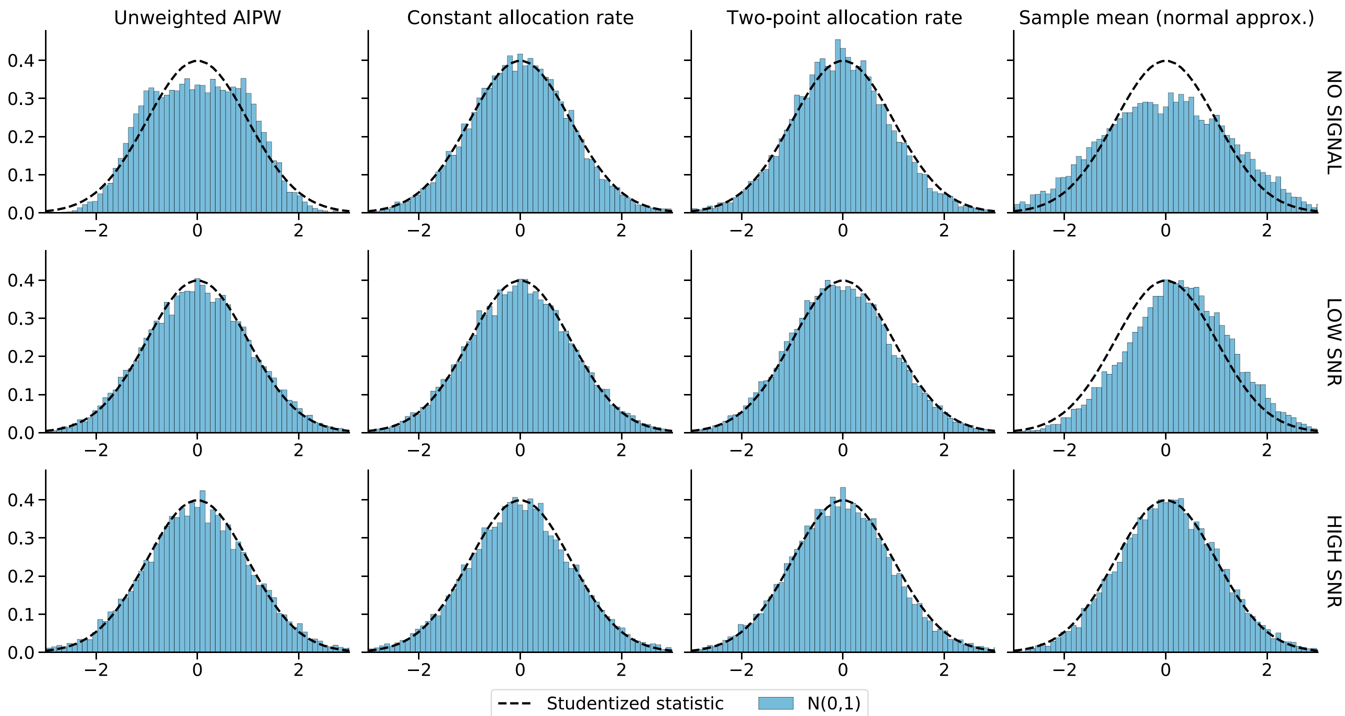

Fig. 3. Histogram of studentized statistics of the form $(\widehat{\Delta}_T - \Delta)/\widehat{V}_T^{1/2}$ for the difference in arm values $\Delta(3,1) = \mathbb{E}[Y_t(3) - Y_t(1)]$ at $T = 10^5$. Numbers are aggregated across over 10^5 replications.

undercoverage when there's little or no signal. On the other hand, the nonasymptotic confidence sequences of ref. 19 are conservative, but often wide.

These simulations suggest that, in similar applications, the adaptively weighted AIPW estimator with two-point allocation [**18**] and the sample mean with confidence sequences based on ref. 19 should be preferred. These two methods have complementary advantages. In terms of mean-squared error, the sample mean often performs better, in particular, in the presence of stronger signal. As for inference, normal intervals around the adaptively weighted estimator with two-point allocation have asymptotically nominal coverage, while confidence sequences are often conservative and wider than those based on normal approximations; however, the former is valid only at a prespecified horizon, while the latter is valid for all time periods and allows for arbitrary stopping times. Finally, in terms of assumptions, the adaptively weighted estimator requires knowledge of the propensity scores, and its justification requires that the propensity scores decay at a slow enough rate; the nonasymptotic confidence sequences for the sample mean require no restrictions on the assignment process and can be used even with deterministic methods such as Upper Confidence Bound (UCB) algorithms (2),[**] but require knowledge about other aspects of the distribution of potential outcomes, such as their support or an upper bound on their variance (ref. 34, section 3.2).

Related Literature

Much of the literature on policy evaluation with adaptively collected data focuses on learning or estimating the value of an optimal policy. The classical literature (e.g., ref. 35) focuses on strategies for allocating treatment in clinical trials to optimize various criteria, such as determining whether a treatment is helpful or harmful relative to control. Ref. 9 generalizes this substantially, addressing the problem of optimally allocating treatment to estimate or testing a hypothesis about a finite-dimensional parameter of the distribution of the data. In optimal-allocation problems, the undersampling issue we address by adaptive weighting does not arise, as undersampling treatments relevant to the estimand or hypothesis of interest is suboptimal.

Ref. 36 considers the problem of policy evaluation when treatment is sequentially randomized, but otherwise unrestricted. The estimator they propose in their section 10.3, when specialized to the problem of estimating an arm value, reduces to the AIPW estimator [**6**]. They establish asymptotic normality of their estimator under assumptions, implying that a nonnegligible proportion of participants is assigned the treatment of interest throughout the study. Ref. 30 proposes a stabilized variant of this estimator, which, similarly specialized, reduces to the adaptively weighted estimator with constant allocation rates [**15**]. The applicability of this approach to bandit problems is mentioned in ref. 37. Ref. 32 considers a similar refinement of an analogous weighted-average estimator ([**25**]) for batched adaptive designs.

Drawing on the tradition of debiasing in high-dimensional statistics, ref. 15 proposes a method that can be used to estimate policy values in noncontextual and linear-contextual bandits. Their approach, W-decorrelation, yields consistent and asymptotically normal estimates of linear-regression coefficients when the covariates have strong serial correlation. For multiarmed bandits, where the arm indicators are used as the covariates, their estimates of arm values take the form

$$\widehat{Q}_T^{WD}(w) := \bar{Y}_{w,T} + \sum_{t=1}^{T} a_{t,w}(Y_t - \bar{Y}_{w,T}) \quad \text{where}$$

$$a_{t,w} := \frac{1}{1+\lambda}\left(\frac{\lambda}{1+\lambda}\right)^{N_{W_t,t-1}} \mathbb{I}\{W_t = w\},$$

where $N_{w,t}$ is the number of times arm w was selected up to period t, $\bar{Y}_{w,T}$ is the sample average of its rewards at T, and

[**] For deterministic algorithms, such as UCB, the methods in ref. 33 suggest inverse propensity weighting schemes, whose weights are of the form $1/\max\{\hat{e}_t(w), \gamma\}$, where $\hat{e}_t(w)$ are estimates of assignment probabilities based, e.g., on the empirical distribution of past assignments, and $\gamma > 0$ is a lower bound. However, this heuristic is not guaranteed to produce asymptotically normal estimates.

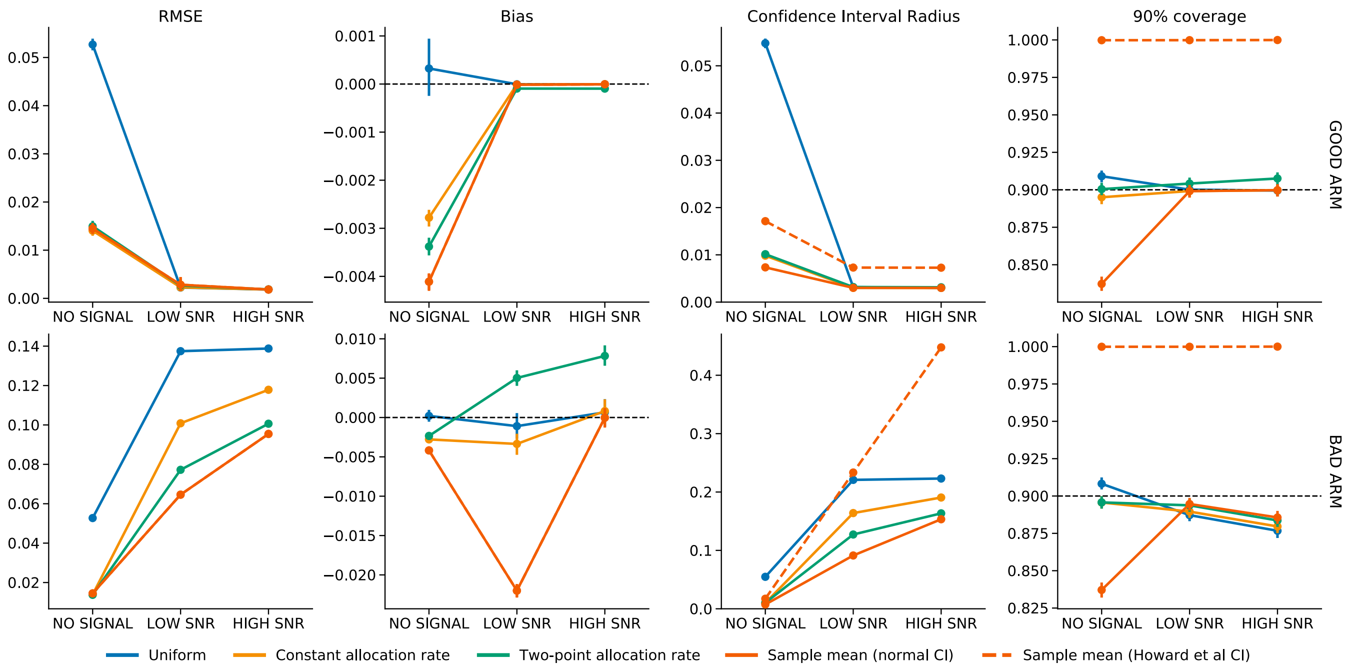

Fig. 4. Estimates of the bad-arm value $Q(1)$ and good-arm value $Q(3)$ at $T = 10^5$. Error bars are 95% CIs around averages across 10^5 replications.

λ is a tuning parameter. We include a numerical evaluation of W-decorrelation in *SI Appendix*, section A, finding it to produce arm-value estimates with high variance.

Discussion

Adaptive experiments such as multiarmed bandits are often a more efficient way of collecting data than traditional randomized controlled trials. However, they bring about several new challenges for inference. Is it possible to use bandit-collected data to estimate parameters that were not targeted by the experiment? Will the resulting estimates have asymptotically normal distributions, allowing for our usual frequentist CIs? This paper provided sufficient conditions for these questions to be answered in the affirmative and proposed an estimator that satisfies these assumptions by construction. Our approach relies on constructing averaging estimators, where the weights are carefully adapted so that the resulting asymptotic distribution is normal with low variance. In empirical applications, we have shown that our method outperforms existing alternatives, in terms of both mean squared error and coverage.

We believe this work represents an important step toward a broader research agenda for policy learning and evaluation in adaptive experiments. Natural questions left open include the following.

What other estimators can be used for normal inference with adaptively collected data? In this paper, we have focused on estimators derived via the adaptively weighted AIPW construction [6]. However, this is not the only way to obtain normal CIs For example, in the setting of Theorem 2, one could also consider the weighted-average estimator

$$\widehat{Q}_T^{h\text{-avg}}(w) = \sum_{t=1}^{T} h_t(w) \frac{\mathbb{I}\{W_t = w\}}{e_t(w)} Y_t \Bigg/ \sum_{t=1}^{T} h_t(w) \frac{\mathbb{I}\{W_t = w\}}{e_t(w)}. \qquad [25]$$

Asymptotic normality of this estimator essentially follows from Theorem 2 (*SI Appendix*, section C.4). In our numerical experiments, we found its performance to be essentially indistinguishable from that of the adaptively weighted estimator defined as in [6]. We've focused on [6] because it readily allows formal study in a more general setting; however, the simpler estimator [25] is appealing in the special case of evaluating a single arm. For a more general discussion of the relationship between augmented estimators like $\widehat{Q}^h$ and variants like $\widehat{Q}^{h\text{-avg}}$ in adaptive experiments, see ref. 36.

What should an optimality theory look like? Our result in Theorem 2 provides one recipe for building CIs using an adaptive data-collection algorithm like Thompson sampling, for which we know the treatment-assignment probabilities. Here, however, we have no optimality guarantees on the width of these CIs. It would be of interest to characterize, e.g., the minimum worst-case expected width of normal CIs that can be built by using such data.

How do our results generalize to more complex sampling designs? In many application areas, there's a need for methods for policy evaluation and inference that work with more general designs, such as contextual bandits, and in settings with nonstationarity or random stopping.

Data Availability. Algorithms and computer codes have been deposited in GitHub (https://github.com/gsbDBI/adaptive-confidence-intervals/).

ACKNOWLEDGMENTS This work was supported by the Sloan Foundation; Office of Naval Research Grant N00014-17-1-2131; NSF Grant DMS-1916163; Schmidt Futures; Golub Capital Social Impact Laboratory; and the Stanford Institute for Human-Centered Artificial Intelligence. R.Z. was supported by a Total Innovation fellowship and the PayPal Innovation fellowship. In addition, we thank Steve Howard, Sylvia Klosin, Sanath Kumar Krishnamurthy, and Aaditya Ramdas for helpful conversations.

1. T. L. Lai, H. Robbins, Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6**, 4–22 (1985).
2. P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* **32**, 48–77 (2003).
3. S. Bubeck, R. Munos, G. Stoltz, Pure exploration in finitely-armed and continuous-armed bandits. *Theor. Comput. Sci.* **412**, 1832–1852 (2011).
4. M. Kasy, A. Sautmann, Adaptive treatment assignment in experiments for policy choice. *Econometrica* **89**, 113–132 (2021).

5. S. Mannor, J. N. Tsitsiklis, The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.* **5**, 623–648 (2004).
6. D. Russo, Simple Bayesian algorithms for best-arm identification. *Oper. Res.* **68**, 1625–1647 (2020).
7. H. Robbins, Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **58**, 527–535 (1952).
8. F. Hu, W. F. Rosenberger, *The Theory of Response-Adaptive Randomization in Clinical Trials* (Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, 2006), Vol. 525.
9. M. J. van der Laan, "The construction and analysis of adaptive group sequential designs" (U.C. Berkeley Division of Biostatistics Working Paper Series 232, University of California Berkeley, Berkeley, CA, 2008; https://biostats.bepress.com/ucbbiostat/paper232/).
10. X. Nie, X. Tian, J. Taylor, J. Zou, A. Storkey, F. Perez-Cruz, "Why adaptively collected data have negative bias and how to correct for it" in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, A. Storkey and F. Perez-Cruz, Eds. (Proceedings of Machine Learning Research, Microtome Publishing, Brookline, MA, 2018), **Vol. 84**, pp. 1261–1269.
11. M. Xu, T. Qin, T. Y. Liu, "Estimation bias in multi-armed bandit algorithms for search advertising" in *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger, Eds. (Advances in Neural Information Processing Systems, Curran Associates, Inc., Red Hook, NY, 2013), Vol. 26, pp. 2400–2408.
12. J. Bowden, L. Trippa, Unbiased estimation for response adaptive clinical trials. *Stat. Methods Med. Res.* **26**, 2376–2388 (2017).
13. J. Shin, A. Ramdas, A. Rinaldo, On the bias, risk and consistency of sample means in multi-armed bandits in. arXiv [Preprint] (2019). https://arxiv.org/abs/1902.00746 (Accessed 1 January 2021).
14. J. Shin, A. Ramdas, A. Rinaldo, "On conditional versus marginal bias in multi-armed bandits" in *Proceedings of the 37th International Conference on Machine Learning*, H. Daumé III, A. Singh, Eds. (Proceedings of Machine Learning Research, Microtome Publishing, Brookline, MA, 2020), **Vol. 119**, pp. 8852–8861.
15. Y. Deshpande, L. Mackey, V. Syrgkanis, M. Taddy, "Accurate inference for adaptive linear models" in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy, A. Krause, Eds. (Proceedings of Machine Learning Research, Microtome Publishing, Brookline, MA, 2018), **Vol. 80**, pp. 1194–1203.
16. Y. Deshpande, A. Javanmard, M. Mehrabi, Online debiasing for adaptively collected high-dimensional data. arXiv [Preprint] (2019). https://arxiv.org/abs/1911.01040 (Accessed 1 January 2021).
17. V. F. Melfi, C. Page, Estimation after adaptive allocation. *J. Stat. Plann. Inference* **87**, 353–363 (2000).
18. I. Andrews, T. Kitagawa, A. McCloskey, "Inference on winners" (NBER Working Paper 25456, National Bureau of Economic Research, Cambridge, MA, 2019; https://www.nber.org/papers/w25456).
19. S. R. Howard, A. Ramdas, J. McAuliffe, J. Sekhon, Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Stat.*, in press.
20. H. Robbins, Statistical methods related to the law of the iterated logarithm. *Ann. Math. Stat.* **41**, 1397–1409 (1970).
21. W. Fithian, D. Sun, J. Taylor, Optimal inference after model selection. arXiv [Preprint] (2014). https://arxiv.org/abs/1410.2597 (Accessed 1 January 2021).
22. A. P. Dawid, Selection paradoxes of Bayesian inference. *Lecture Notes Monogr. Ser.* **24**, 211–220 (1994).
23. W. R. Thompson, On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285–294 (1933).
24. D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, A tutorial on Thompson sampling. *Found. Trends Mach. Learn.* **11**, 1–96 (2018).
25. G. W. Imbens, D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, UK, 2015).
26. J. M. Robins, A. Rotnitzky, L. P. Zhao, Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **89**, 846–866 (1994).
27. P. Hall, C. C. Heyde, *Martingale Limit Theory and its Application* (Academic Press, New York, NY, 1980).
28. V. H. de la Peña, T. L. Lai, Q. M. Shao, *Self-Normalized Processes: Limit Theory and Statistical Applications* (Springer-Verlag, Berlin, Germany, 2008).
29. T. Lattimore, C. Szepesvári, *Bandit Algorithms* (Cambridge University Press, Cambridge, UK, 2020).
30. A. R. Luedtke, M. J. van der Laan, Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Stat.* **44**, 713 (2016).
31. V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, J. M. Robins, Locally robust semiparametric estimation. arXiv [Preprint] (2016). https://arxiv.org/abs/1608.00033 (Accessed 1 January 2021).
32. K. Zhang, L. Janson, S. Murphy, "Inference for batched bandits" in *NeurIPS 2020: 34th Conference on Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Advances in Neural Information Processing Systems, Curran Associates, Inc., Red Hook, NY 2020), **Vol. 33**, pp. 9818–9829.
33. M. Dimakopoulou, Z. Zhou, S. Athey, G. Imbens, Estimation considerations in contextual bandits. arXiv [Preprint] (2017). https://arxiv.org/abs/1711.07077 (Accessed 1 January 2021).
34. S. R. Howard, A. Ramdas, J. McAuliffe, J. Sekhon, Time-uniform Chernoff bounds via nonnegative supermartingales. *Probab. Surv.* **17**, 257–317 (2020).
35. W. F. Rosenberger, J. M. Lachin, *Randomization in Clinical Trials: Theory and Practice* (John Wiley & Sons, New York, NY, 2015).
36. M. J. van der Laan, S. D. Lendle, "Online targeted learning" (U.C. Berkeley Division of Biostatistics Working Paper Series 330, University of California Berkeley, Berkeley, CA, 2014; https://biostats.bepress.com/ucbbiostat/paper330/).
37. A. R. Luedtke, M. J. van der Laan, Parametric-rate inference for one-sided differentiable parameters. *J. Am. Stat. Assoc.* **113**, 780–788 (2018).