# Adaptive Inference for Functionals of M-Estimands

**James Leiner**
Carnegie Mellon Univeristy
jleiner@stat.cmu.edu

**Aurélien Bibaut**
Netflix
abibaut@netflix.com

**Nathan Kallus**
Netflix and Cornell Tech
kallus@cornell.edu

**Aaditya Ramdas**
Stanford University
aramdas@stanford.edu

**Koulik Khamaru**
Rutgers University
kk1241@rutgers.edu

## Abstract

Reinforcement learning and contextual bandit algorithms have become increasingly common in sequential decision-making applications. When these methods are deployed in high-stakes domains, there is growing interest not only in learning effective policies, but also in conducting statistical inference for quantities learned under adaptive data collection. However, classical procedures applied naively in these settings can fail: even when estimators are unbiased, their variance becomes path-dependent and as a result may not be asymptotically normal. A growing literature has emerged to ameliorate this problem, but solutions tend to be problem specific and often rely on correct specification of a working model. In this work, we develop a unified framework for constructing asymptotically valid confidence intervals to cover smooth functionals of nonparametric M-estimands under adaptive sampling. Under Neyman orthogonality, we provide two novel methods for performing inference: (1) a self-normalized statistic based on the realized quadratic variation of the influence function and (2) a statistic using a plug-in estimate of the conditional variance based on reweighted influence function increments. Our results allow for flexible nonparametric estimation of nuisance parameters and remain valid under model misspecification. Our theory is supported by a simulation study for a dynamic pricing application which demonstrates that this method can produce asymptotically valid confidence intervals where standard methods fail.

## 1 Introduction

Reinforcement learning algorithms have become widespread due to their state-of-the-art performance in a wide variety of areas including health interventions (Yom-Tov et al., 2017), personalized web recommendations (Afsar et al., 2022), public policy (Kasy and Sautmann, 2021; Caria et al., 2024) and clinical trials (Zhao et al., 2009; Liu et al., 2020). The widespread adoption of these approaches means that many modern datasets no longer obey an $i.i.d.$ assumption, and attempting to apply classical statistical techniques to these data results in biased parameter estimates (Shin et al., 2019) and invalid inferential guarantees.

A classical result (Lai and Wei, 1982) demonstrates that a sufficient condition for ordinary least squares parameters to concentrate in the linear model $y_t = \beta^T x_t + \epsilon_t$ with adaptively chosen $x_t$ is for the covariance matrix $\sum_{t=1}^{T} x_t x_t^T$ to concentrate around a sequence of deterministic matrices $\{A_T\}_{T\geq 1}$. In contextual bandit problems, this condition typically requires that the sampling policy converge to a deterministic stationary policy independent of the data trajectory. In cases where the optimal policy is not unique or there are many close ties, common bandit algorithms (e.g. $\epsilon$-greedy, Thompson sampling, UCB) may not converge (Deshpande et al., 2018; Zhang et al., 2020;

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: List of target estimands studied in prior work that are specific instantiations of this framework. We note that this is not exhaustive and additional targets are discussed in Appendix A.

| **Target estimand** | **Prior Work** | $\theta_{P_e}$ | $\Psi_e(\theta_{P_e})$ |
|---|---|---|---|
| Average treatment effect (Appendix A.1) | Hadad et al. (2021); Bibaut et al. (2021); Kato et al. (2024); Cook et al. (2024) | $\mathbb{E}[Y \mid A = a, C = c]$ | $\mathbb{E}_{P_C}[\theta_{P_e}(1, C) - \theta_{P_e}(0, C)]$ |
| Parametric $M$-estimator (Appendix A.2) | Zhang et al. (2021); Guo and Xu (2025); Leiner et al. (2026) | $\operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_{P_e}[\ell_\theta(C, A, Y)]$ | $v^\top \theta_{P_e}$, fixed $v$ |
| Partial linear regression (Appendix A.3) | Lin et al. (2025) | $\operatorname{argmin}_{\theta} \mathbb{E}_{P_e}\left[\left(Y - \theta^\top X - \eta_{P_e}\right)^2\right]$ | $v^\top \theta_{P_e}$, fixed $v$ |
| Value of an optimal policy (Appendix A.4) | Luedtke and van der Laan (2016) | $\mathbb{E}[Y \mid A = a, C = c]$ | $\sup_{\pi \in \Pi} \mathbb{E}_P\left[\int_{\mathcal{A}} \pi(a \mid C)\theta_{P_e}(a, C)\, \mu(da)\right]$ |

Khamaru et al., 2025) and classical approaches to inference applied naively will consequently have non-Gaussian limiting distributions.

Motivated by these results, a recent literature has developed to provide inferential guarantees on adaptively collected data. One set of approaches assumes that the sampling policies are known and appropriately reweighs observations to reduce bias or stabilize variance (Hadad et al., 2021; Kato et al., 2024; Bibaut et al., 2021; Zhang et al., 2021; Cook et al., 2024; Guo and Xu, 2025; Leiner et al., 2026). An alternative line of work assumes that the sampling policies are unknown and instead uses an online debiasing approach to ensure valid inference (Deshpande et al., 2018; Ying et al., 2023; Khamaru et al., 2025).

In the above cases, the inferential target is usually either a finite-dimensional parameter in a well-specified parametric model (e.g. least squares) or a smooth functional of the conditional regression function (e.g. average treatment effect). Despite these differences, the proposed estimators share a common structure: they combine importance weighting or martingale arguments with bias correction terms that resemble influence function adjustments. However, these constructions are typically developed in a problem-specific manner, with separate analyzes for each estimand and modeling framework. As a result, there is currently no general framework that unifies these approaches in the way that semiparametric efficiency theory (Bickel and Klaassen, 1993) and double machine learning (Chernozhukov et al., 2018; Rotnitzky et al., 2021) do in the $i.i.d.$ setting.

Moreover, many existing methods impose restrictive assumptions on the underlying model or data-generating process. In particular, several approaches either assume correctly specified parametric models or rely on limited forms of nuisance estimation, precluding the use of flexible nonparametric estimators. Methods that allow model misspecification are typically confined to parametric Z estimators and require strong conditions such as convergence of the sampling policy to a deterministic limit (Guo and Xu, 2025), accurate estimation of the conditional moments of a score function (Leiner et al., 2026), or longitudinal settings where batch sizes can tend to infinity (Zhang et al., 2023). On the other hand, methods that do allow flexible estimation of nuisance parameters (Lin et al., 2025) do not provide coverage guarantees under misspecification. These limitations restrict the applicability of current techniques in modern machine learning settings, where models are often misspecified and nuisance components are estimated using black box regression approaches.

**Summary of Contributions** To address these limitations, we develop a unified framework for inference on smooth functionals of nonparametric M-estimands under adaptive data collection. Our approach builds on recent advances in automatic debiased machine learning (Chernozhukov et al., 2022; van der Laan et al., 2026), which show that in the i.i.d. setting, smooth functionals of M-estimands can be represented via a Riesz representer that links the target functional to the curvature of the underlying loss. For adaptively collected data, we provide a general construction of one-step estimators that use this representation, enabling valid inference with flexible black box parameter and nuisance estimation. Specifically:

1. We provide a **general method** for performing inference on smooth functionals of non-parametric $M$-estimands, including a first order expansion (Theorem 1) and a generic recipe for evaluating the resulting influence function on adaptively collected datasets.

2. We provide an **asymptotic normality** result (Theorem 2) that uses a novel plug-in estimate (Theorem 4) of the (time-varying) conditional variance of the influence function. Crucially, our variance estimator is based on reweighted estimated influence function increments and does not require training additional models or data splitting in contrast to prior work.
3. In prior work, variance was either assumed to stabilize to a deterministic limit or assumed not to converge. We identify an **intermediate regime** where variance converges to a stable but non-deterministic quantity. In this case, we provide an additional central limit theorem (Theorem 3) based on *self-normalized* increments of the influence function that does not require bespoke variance estimation techniques.

**Paper Organization** In Section 2 we introduce the contextual bandit problem and define our class of target estimands. In Section 3 we derive the corresponding efficient influence function and show that an appropriately defined one-step estimator is asymptotically normal. In Section 4, we demonstrate the utility of our framework using a dynamic pricing application. We provide concluding thoughts in Section 5. All proofs are deferred to the Appendix.

**Notation** We define $[n] := \{1, ..., n\}$ for a positive integer $n$. We denote $\lfloor x \rfloor$ for some $x \in \mathbb{R}$ to refer to the largest integer less than or equal to $x$. For a probability measure $P$, we write $Pf := \int f(z)\, dP(z)$.

When $\Theta$ and $\Upsilon$ are normed linear spaces and $f_{\theta,\eta} : \mathcal{Z} \to \mathbb{R}$ is a function indexed by $(\theta, \eta) \in \Theta \times \Upsilon$, we denote the directional derivative with respect to the $\theta$-argument by

$$\partial_\theta f_{\theta,\eta}(z)[h] := \frac{d}{d\epsilon} f_{\theta+\epsilon h,\eta}(z)\Big|_{\epsilon=0},$$

for $h \in \Theta$. Similarly, we define the second-order derivative as

$$\partial_\theta^2 f_{\theta,\eta}(z)[h,u] := \frac{\partial^2}{\partial\epsilon\partial\delta} f_{\theta+\epsilon h+\delta u,\eta}(z)\Big|_{\epsilon=\delta=0}.$$

Mixed derivatives are defined in the natural way, e.g.

$$\partial_\eta\partial_\theta f_{\theta,\eta}(z)[h,u] := \frac{\partial^2}{\partial\delta\partial\epsilon} f_{\theta+\epsilon h,\eta+\delta u}(z)\Big|_{\epsilon=\delta=0}.$$

When needed, we assume that $f_{\theta,\eta}$ is Fréchet differentiable in $\theta$.

## 2 Problem Setup

We consider the contextual bandit problem over $T$ rounds. At each round $t = 1, ..., T$, a context $C_t \in \mathcal{C}$ is observed, an action $A_t \in \mathcal{A}$ is selected, and then an outcome $Y_t \in \mathbb{R}$ is observed. We denote the history up to time $t-1$ as $\mathcal{H}_{t-1} := \{C_i, A_i, Y_i\}_{i=1}^{t-1}$ and let $\mathcal{F}_{t-1} := \sigma(\mathcal{H}_{t-1})$ be the corresponding filtration. At each round $t$, the observer selects a policy $\Pi_t$, measurable with respect to $\mathcal{F}_{t-1}$ and then samples $A_t \sim \Pi_t(\cdot \mid C_t)$. In certain contexts, it may be more useful to collapse $C_t$ and $A_t$ into a single object which we will label as $X_t := (C_t, A_t) \in \mathcal{C} \times \mathcal{A}$. Similarly, we will label $Z_t = (C_t, A_t, Y_t) \in \mathcal{Z}$ with $\mathcal{Z} := \mathcal{C} \times \mathcal{A} \times \mathbb{R}$ as a compact notation when referring to the observation tuple at step $t$. We denote $P_t$ as the joint distribution of $Z_t$ conditional on $\mathcal{F}_{t-1}$. As is standard in the literature, we assume that the marginal distribution of contexts and conditional distribution of rewards *given* particular choices of actions and rewards are fixed across time. The dependence across time arises solely through the policy $\Pi_t$, which may depend on past history.

**Assumption 1.** *The data generating process has the following properties*

*(a) (Exogenous contexts). The contexts $\{C_t\}_{t=1}^T$ are i.i.d. with $C_t \sim P_{\mathcal{C}}$ and $C_t \perp\!\!\!\perp \mathcal{F}_{t-1}$ for all $t \in [T]$.*

*(b) (Fixed conditional reward distribution). There exists a conditional distribution $P_Y(\cdot \mid c, a)$ such that for all $t \in [T]$,*

$$Y_t \mid C_t, A_t \sim P_Y(\cdot \mid C_t, A_t),$$

*and $Y_t \perp\!\!\!\perp \mathcal{F}_{t-1} \mid C_t, A_t$.*

Our goal will be to conduct inference on functionals of a non-parametric $M$-estimand. We mirror the framework taken in van der Laan et al. (2026) but with the target and inference procedures adapted to the sequential, adaptive setting. Let $\Pi_e(\cdot \mid c)$ be an evaluation distribution that we assume to be *fixed* and independent of history. We denote $P_e$ as the corresponding law where $C \sim P_C$, $A \sim \Pi_e(\cdot \mid C)$, and $Y \mid C, A \sim P_Y(\cdot \mid C, A)$. We note that at each $t$, $P_e$ shares the same marginal distribution of contexts and outcome mechanism as $P_t$, differing only in the distribution of actions.

**Assumption 2.** *There exists a common dominating measure $\mu$ such that for all $t$, $\Pi_t(\cdot \mid c), \Pi_e(\cdot \mid c) \ll \mu(\cdot \mid c)$ with corresponding densities $\pi_t(\cdot \mid c)$ and $\pi_e(\cdot \mid c)$. We assume these densities are known and $\Pi_e(\cdot \mid c) \ll \Pi_t(\cdot \mid c)$ almost surely.*

Define a loss function $\ell : \Theta \times \Upsilon \times \mathcal{Z} \to \mathbb{R}$ written as $\ell_{\theta,\eta}(z)$, where $(\Theta, \|\cdot\|_\Theta)$ and $(\Upsilon, \|\cdot\|_\Upsilon)$ are normed linear spaces. Define the $M$-estimand,

$$\theta_{P_e} := \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{P_e}\left[\ell_{\theta,\eta_{P_e}}(Z)\right], \tag{1}$$

where $\eta_{P_e}$ is a nuisance function. Our parameter of interest will be a smooth functional of the $M$-estimand defined as

$$\Psi_e(\theta_{P_e}) := \mathbb{E}_{P_e}\left[m_{\theta_{P_e}}(Z)\right], \tag{2}$$

for some measurable function $m : \Theta \times \mathcal{Z} \to \mathbb{R}$. We note that this framework is general enough to cover most target estimands studied in the adaptive inference literature, including: projection parameters for linear models and parametric $M$-estimators, partial linear regression parameters, and average treatment effect. See Table 1 and Appendix A for an extended discussion.

## 3 Main Results

We first derive the influence function at $P_e$ and establish a local first-order expansion of $\Psi_e$.

**Assumption 3** (Regularity of the loss at evaluation law)**.** *We assume that $\theta_{P_e}$ as defined in Equation (1) exists and is unique. Moreover, the following conditions hold:*

- *(a) (Differentiability) The loss $\ell_{\theta,\eta}(z)$ is twice Fréchet differentiable in $\theta$ in a neighborhood of $\theta_{P_e}$ and the functional $m_\theta(z)$ is once Fréchet differentiable in $\theta$ in a neighborhood of $\theta_{P_e}$.*
- *(b) (First order condition). $\mathbb{E}_{P_e}\left[\partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(Z)[h]\right] = 0$ for all $h \in \Theta$.*
- *(c) (Neyman orthogonality) $\mathbb{E}_{P_e}\left[\partial_\eta \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(Z)[h,u]\right] = 0$ for all $h \in \Theta$ and $u \in \Upsilon$.*
- *(d) (Positive definite Hessian) Let $H_{P_e}(u,h) := \mathbb{E}_{P_e}\left[\partial_\theta^2 \ell_{\theta_{P_e},\eta_{P_e}}(Z)[h,u]\right]$ for $h, u \in \Theta$. We assume there exists a constant $c > 0$ such that $c\|h\|_\Theta^2 \le H_{P_e}(h,h)$ for all $h \in \Theta$.*

Condition (a) guarantees sufficient smoothness of the loss and functional so that a first order expansion can be taken. Condition (b) is a standard first order condition defining the target estimand. Condition (c) imposes Neyman orthogonality, which ensures that first-order perturbations in the nuisance parameter do not affect the target. We note that this is not a substantial restriction, as most common loss functions can be orthogonalized without changing the underlying risk. Condition (d) requires the Hessian be positive definite, ensuring local identification, and that the curvature of the loss induces a well-behaved inner product structure. Finally, we assume an envelope condition (Assumption 4) to bound higher order remainder terms in our first order expansion but defer its statement to Appendix D.1.

Under Assumptions 3 and 4, the bilinear form $H_{P_e}$ defines an inner product on $\Theta$ and the map $\mathbb{E}_{P_e}\left[\partial_\theta m_{\theta_{P_e}}(Z)[h]\right]$ is a bounded linear functional with respect to the induced norm.Write $\|h\|_{H_{P_e}}^2 := H_{P_e}(h,h)$ and let $\bar{\Theta}$ be the completion of $\Theta$ in this norm. Lemma 1 justifies extending the original norm and the derivative maps in their direction arguments to $\bar{\Theta}$. By the Riesz representation theorem, there is a unique element $\alpha_{P_e} \in \bar{\Theta}$ that solves the Hessian equation

$$E_{P_e}\left[\partial_\theta m_{\theta_{P_e}}(Z)[h]\right] = H_{P_e}(\alpha_{P_e}, h) \text{ for all } h \in \Theta.$$

In particular,

$$\alpha_{P_e} = \arg\min_{\alpha \in \bar{\Theta}} \left\{ \frac{1}{2}\mathbb{E}_{P_e}\left[\partial_\theta^2 \ell_{\theta_{P_e},\eta_{P_e}}(Z)[\alpha,\alpha]\right] - E_{P_e}\left[\partial_\theta m_{\theta_{P_e}}(Z)[\alpha]\right] \right\}.$$

An extended discussion of this is deferred to Section D. This representation allows us to express first-order changes in $\Psi$ in terms of the curvature of the loss, and forms the basis for the expansion below.

**Theorem 1** (One-step expansion at the evaluation law)**.** *Suppose Assumptions 3-4 hold. Let* $(\bar{\theta}, \bar{\eta}, \bar{\alpha}) \in \Theta \times \Upsilon \times \Theta$ *be in a neighborhood of* $(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})$*. Let* $\delta_{\text{lin}} := \mathbf{1}\{m_\theta \text{ is not linear in } \theta \text{ at } \theta_{P_e}\}$ *and* $\delta_{\text{quad}} := \mathbf{1}\{\ell_{\theta,\eta_{P_e}}(\cdot) \text{ is not quadratic at } \theta_{P_e}\}$*. Then,*

$$\begin{aligned}\Psi_e(\bar{\theta}) - \Psi_e(\theta_{P_e}) = \mathbb{E}_{P_e}\left[\partial_\theta \ell_{\bar{\theta},\bar{\eta}}[\bar{\alpha}]\right] - H_{P_e}[\bar{\alpha} - \alpha_{P_e},\, \bar{\theta} - \theta_{P_e}] \\ + (\delta_{\text{lin}} + \delta_{\text{quad}}) O\big(\|\bar{\alpha}\|_\Theta \|\bar{\theta} - \theta_{P_e}\|_\Theta^2\big) \\ + O\left(\|\bar{\alpha}\|_\Theta \left(\|\bar{\theta} - \theta_{P_e}\|_\Theta \|\bar{\eta} - \eta_{P_e}\|_\Upsilon + \|\bar{\eta} - \eta_{P_e}\|_\Upsilon^2\right)\right).\end{aligned}$$

Motivated by this expansion, we denote

$$\varphi_{P_e}(Z) := m_{\theta_{P_e}}(Z) - \Psi(P_e) - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(Z)[\alpha_{P_e}]. \tag{3}$$

By construction, $\mathbb{E}_{P_e}[\varphi_{P_e}(Z)] = 0$. We note that this function coincides with the efficient influence function of $\Psi$ at $P_e$ in the nonparametric model.

**Change of measure from adaptive data to evaluation policy** We move now to writing a first order expansion for an *empirically realizable* functional, obtained by replacing population expectation under $P_e$ with an empirical operator. Consider estimating $\Psi_e(\theta_{P_e})$ by taking a weighted average over the sample, for instance, $\frac{1}{T}\sum_{t=1}^T w_t m(\bar{\theta})$ for some $\bar{\theta} \in \Theta$. Then,

$$\left(\frac{1}{T}\sum_{t=1}^T w_t m_{\bar{\theta}}(Z_t)\right) - \Psi_e(\theta_{P_e}) = \underbrace{\left(\frac{1}{T}\sum_{t=1}^T w_t m_{\bar{\theta}}(Z_t)\right) - \Psi_e(\bar{\theta})}_{\text{Empirical process term}} + \underbrace{\left(\Psi_e(\bar{\theta}) - \Psi_e(\theta_{P_e})\right)}_{\text{Plug-in error}}.$$

The second term can be analyzed by applying Theorem 1. For the first term, we note that it introduces first order bias unless $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T w_t m(\bar{\theta})(Z_t)\right] = \mathbb{E}_{P_e}[m(\bar{\theta})(Z_t)]$. To eliminate this bias, we can choose $w_t := \frac{\pi_e(A_t|C_t)}{\pi_t(A_t|C_t)}$ as the inverse propensity weights because for any integrable function $f$,

$$\begin{aligned}\mathbb{E}\left[w_t f(C_t, A_t, Y_t)|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[w_t f(C_t, A_t, Y_t)|\mathcal{F}_{t-1}, C_t\right] \mid \mathcal{F}_{t-1}\right] \\ &= \int f(c,a,y) P_Y(dy \mid c,a) \pi_e(a \mid c) \mu(da \mid c) P_C(dc) \\ &= \mathbb{E}_{P_e}\left[f(C,A,Y)\right].\end{aligned}$$

This allows us to replace any expectation involving $f$ defined relative to the evaluation policy, such as $\mathbb{E}_{P_e}[f(Z)]$, with the empirically realizable form $\frac{1}{T}\sum_{t=1}^T w_t \mathbb{E}\left[f(Z_t) \mid \mathcal{F}_{t-1}\right]$.

### 3.1 One-step estimator and asymptotic normality results

Our asymptotic results rely on the fact that a suitably rescaled influence function admits a martingale structure. Let $\varphi_t := \varphi_{P_e}(Z_t)$ and note that $\mathbb{E}[w_t\varphi_t \mid \mathcal{F}_{t-1}] = P_e\varphi_{P_e} = 0$. This implies that $\{w_t\varphi_t\}_{t=1}^T$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_{t-1}\}_{t=1}^T$. Define

$$\sigma_t^2 := \text{Var}(w_t\varphi_t \mid \mathcal{F}_{t-1}) = \mathbb{E}[w_t^2\varphi_t^2 \mid \mathcal{F}_{t-1}]. \tag{4}$$

If $T^{-1}\sum_{t=1}^T \sigma_t^2 \xrightarrow{p} \sigma^2$ for a deterministic constant $\sigma^2 > 0$, one may apply a martingale central limit theorem (Billingsley, 1961; Dvoretzky, 1972; Hall and Heyde, 1980) to show that under regularity conditions $T^{-1/2}\sum_{t=1}^T w_t\varphi_t \xrightarrow{d} N(0, \sigma^2)$. However, in adaptive settings, the conditional variance often will not converge to a deterministic constant, rendering naive variance estimation inappropriate. Alternatively, one can define the martingale difference sequence as $\zeta_t := \sigma_t^{-1} w_t\varphi_t$ and construct a *sequence* of plug-in estimators $\{\hat{\sigma}_t\}_{t=1}^T$ targeting the conditional variance. This strategy of stabilizing observations using conditional variance estimates was initially used in the non-adaptive setting by Luedtke and van der Laan (2016) and later work has applied it to martingale difference sequences

(Hadad et al., 2021; Kato, 2020; Bibaut et al., 2021; Leiner et al., 2026). Of course, constructing such estimators is nontrivial; we postpone this to Section 3.2.

Supposing that a suitable estimate of $\hat{\sigma}_t$ exists, let $(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t)$ denote estimators of $(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})$. Letting $w_t = \frac{\pi_e(A_t|C_t)}{\pi_t(A_t|C_t)}$ and $\hat{A}_T = \sum_{t=1}^T \hat{\sigma}_t^{-1} w_t$, we define our one-step estimator as

$$\hat{\Psi}_T^{\mathrm{os}} := \hat{A}_T^{-1} \sum_{t=1}^T \hat{\sigma}_t^{-1} w_t \left( m_{\hat{\theta}_t}(Z_t) - \dot{\ell}_{\hat{\theta}_t, \hat{\eta}_t}(Z_t)[\hat{\alpha}_t] \right). \tag{5}$$

One-step estimation is a standard semiparametric technique for debiasing naive plug-in estimators. The correction term involving $-\dot{\ell}_{\hat{\theta}_t, \hat{\eta}_t}(Z_t)[\hat{\alpha}_t]$ arises from the first-order expansion in Theorem 1 and removes the leading bias induced by nuisance estimation. After reweighting and stabilization, the estimator behaves asymptotically like an average of mean-zero influence function increments; we formalize this asymptotic normality result in Theorem 2.

**Theorem 2** (CLT for one-step estimator)**.** *Let $(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t, \hat{\sigma}_t)$ denote $\mathcal{F}_{t-1}$ measurable estimators of $(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e}, \sigma_t)$ and define $\hat{\Psi}_T^{\mathrm{os}}$ as in Equation (5). Recall that $w_t = \frac{\pi_e(A_t|C_t)}{\pi_t(A_t|C_t)}$ and $\varphi_t := \varphi_{P_e}(Z_t)$ as defined in Equation (3). Define $\hat{A}_T := \sum_{t=1}^T \hat{\sigma}_t^{-1} w_t$. Assume*

- *(2a)* $\frac{1}{T}\sum_{t=1}^T \|\hat{\theta}_t - \theta_{P^e}\|_\Theta^4 = o_p(T^{-1})$, $\frac{1}{T}\sum_{t=1}^T \|\hat{\eta}_t - \eta_{P^e}\|_\Upsilon^4 = o_p(T^{-1})$, *and* $\frac{1}{T}\sum_{t=1}^T \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^4 = o_p(T^{-1})$;
- *(2b) For any* $\epsilon > 0$, $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[ w_t^2 \varphi_t^2 \mathbf{1}\{|w_t \varphi_t| > \varepsilon\sqrt{T}\} \mid \mathcal{F}_{t-1} \right] \xrightarrow{p} 0$;
- *(2c)* $\sup_{t\in[T]} \sup_{c,a \in \mathcal{C}\times\mathcal{A}} \frac{\pi_e(a|c)}{\pi_t(a|c)} = O(T^{1/3})$ *almost surely;*
- *(2d)* $\frac{1}{T}\sum_{t=1}^T \sigma_t^2/\hat{\sigma_t}^2 - 1 = o_p(1)$ *and* $\min_{t\in[T]} \hat{\sigma}_t^2 > c$ *for some* $c > 0$ *almost surely.*

*Then under Assumptions 1–4,*

$$\frac{\hat{A}_T}{\sqrt{T}}(\hat{\Psi}_T^{\mathrm{os}} - \Psi_e(\theta_{P_e})) \xrightarrow{d} N(0,1). \tag{6}$$

**Remark 1.** *In the case where the loss is quadratic and the functional is linear with respect to $\theta$ (i.e. $\delta_{lin} = \delta_{quad} = 0$ as defined in Theorem 1) and there is no nuisance parameter (e.g. $\Upsilon = \{0\}$), condition (2a) can be simplified to a requirement that $\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{P_e}\left[ \partial_\theta^2 \ell_{\theta_{P_e}, \eta_{P_e}}(Z)[\hat{\theta}_t - \theta_{P_e}, \hat{\alpha}_t - \alpha_{P_e}] \right] = o_p(T^{-1/2})$.*

Requiring that all estimators be $\mathcal{F}_{t-1}$ measurable is natural in the sequential setting and ensures that the empirical process terms arising from parameter and nuisance estimation errors form martingale difference sequences. In the absence of this condition, one would instead need uniform control over the underlying function class, which would require stronger assumptions, such as bracketing entropy bounds. However, such conditions are often difficult to verify for flexible, nonparametric estimators, which are explicitly allowed in our framework.

Condition (2a) is stated in terms of time-averaged higher-order moments of the estimation errors. This reflects the fact that, in the sequential setting, the analysis requires control of products of time-indexed estimation errors. The stated moment conditions provide a convenient sufficient condition for controlling such products via repeated applications of the Cauchy–Schwarz inequality. Intuitively, the condition should be interpreted as requiring that large estimation errors arising early in the sequential process do not dominate the time average. In practice, we recommend using an initial burn-in period of length $T_0 = \lfloor aT \rfloor$ for some fixed $a \in (0,1)$ to train pilot estimates and excluding these observations when computing $\hat{\Psi}_T^{\mathrm{os}}$. A simple implementation is then to freeze the nuisance estimators after the burn-in period by setting $(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t) := (\hat{\theta}_{T_0+1}, \hat{\eta}_{T_0+1}, \hat{\alpha}_{T_0+1})$ for all $t > T_0$ and then computing $\hat{\Psi}_T^{\mathrm{os}}$ using only observations from times $T_0 + 1$ through $T$. In this special case, condition (2a) reduces to the more familiar requirement that each estimator converges at rate $o_p(T^{-1/4})$. More generally, nuisance estimators may be updated periodically after the burn-in period provided the resulting estimators continue to satisfy (2a). While updating the nuisance estimators at every time step is theoretically valid under these conditions, doing so may be computationally

burdensome in practice. We therefore recommend batch updates as a practical compromise between statistical and computational efficiency.

Condition (2b) is a standard Lindeberg-type condition ensuring that no single observation dominates the asymptotic variance. In the present setting, this amounts to controlling the tails of the increments $\varphi_t$. This condition can be verified under mild moment assumptions, for example by requiring either bounded increments of the efficient influence function or a bounded fourth moment. Condition (2c) requires that the exploration rate does not diverge from the evaluation policy too quickly.

An intermediate regime arises when the conditional variance converges to a random variable that depends on the realized trajectory but not on the horizon $T$. In this setting, we are able to avoid per-round variance estimation entirely and can instead use the realized quadratic variation of the influence function in order to create an asymptotically normal test statistic.

**Theorem 3** (CLT for self-normalized one-step estimator)**.** *Let $(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t)$ denote $\mathcal{F}_{t-1}$ measurable estimators of $(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})$ Choose $m_T \to \infty$ with $m_T = o(T)$ and set $(\bar{\theta}, \bar{\eta}, \bar{\alpha}) := (\hat{\theta}_{m_T}, \hat{\eta}_{m_T}, \hat{\alpha}_{m_T})$. Recall that $w_t = \frac{\pi_e(A_t|C_t)}{\pi_t(A_t|C_t)}$ and $\varphi_t := \varphi_{P_e}(Z_t)$ as defined in Equation (3) and set $\bar{\Psi}_T = m_T^{-1} \sum_{t=m_T+1}^{2m_T} w_t \left(m_{\bar{\theta}}(Z_t) - \partial_\theta \ell_{\bar{\theta},\bar{\eta}}(Z_t)[\bar{\alpha}]\right)$. Define $A_T = \sum_{t=1}^{T} w_t$ and let*

$$\hat{S}_T := A_T^{-1} \sum_{t=1}^{T} w_t \left(m_{\hat{\theta}_t}(Z_t) - \partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}(Z_t)[\hat{\alpha}_t]\right), \tag{7}$$

$$\hat{V}_T := A_T^{-2} \sum_{t=2m_T+1}^{T} w_t^2 \left(m_{\hat{\theta}_t}(Z_t) - \partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}(Z_t)[\hat{\alpha}_t] - \bar{\Psi}_T\right)^2. \tag{8}$$

*Assume conditions (2a), (2b), and (2c) of Theorem 2 hold. Furthermore, assume that*

- *(3a)* $\|\bar{\theta} - \theta_{P^e}\|_\Theta = o_p(m_T^{-1/4})$, $\|\bar{\eta} - \eta_{P^e}\|_\Upsilon = o_p(m_T^{-1/4})$, *and* $\|\bar{\alpha} - \alpha_{P_e}\|_\Theta = o_p(m_T^{-1/4})$;
- *(3b)* $\sup_{t\in[T]} \sup_{c,a\in\mathcal{C}\times\mathcal{A}} \frac{\pi_e(a|c)}{\pi_t(a|c)} = o(m_T^{1/2})$ *almost surely;*
- *(3c)* $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[w_t^2\varphi_t^2 \mid \mathcal{F}_{t-1}] \xrightarrow{p} \kappa$ *to some random variable $\kappa$ which is finite and non-zero almost surely.*

*Then under Assumptions 1–4, $\hat{V}_T^{-1/2}\left(\hat{S}_T - \Psi_e(\theta_{P_e})\right) \xrightarrow{d} N(0,1)$.*

Note that the first $2m_T$ observations are only used to create $\bar{\Psi}_T$ which centers the empirical variance and is not a primary driver of asymptotic rates. Condition (3b) implies that the rate $m_T$ is allowed to grow at is driven by the rate at which the importance weights are allowed to inflate. If the weights are uniformly bounded over time, $m_T$ can be chosen to grow at $\log\log T$ rates or slower. On the other hand, if weights grow at $T^{1/3}$ rates, $m_T$ must be chosen to grow at $T^{2/3}$. We note a few example cases where condition (3c) is satisfied, but where $\kappa$ *is not* a deterministic constant in the following examples. Intuitively, a sufficient condition for this to hold is that there is a single asymptotic regime that dominates over the sample with only sublinear rounds of adaptive sampling.

**Example 1** (Deterministic sublinear exploration)**.** *Assume that an experimenter chooses the policy $\pi_t(\cdot|c)$ adaptively the first $T_0 = \lfloor T^\alpha \rfloor$ for $\alpha \in (0,1)$ rounds and then freezes it in all subsequent rounds. In this case, $\Pi_t = \Pi_{T_0}$ for all $t > T_0$. In this case, the conditional variance is constant for all $t > T_0$ and will converge to $\sigma^2_{T_0}$ so long as $\sigma_t$ is uniformly bounded across $t \in [T]$. Nonetheless, it is still a* random quantity.

**Example 2** (Policy switch at stopping times)**.** *For each $T$, denote a set of stopping times such that $0 < \tau_{1,T} < \tau_{2,T} < \ldots < \tau_{K_T} < T$. Suppose that policies change at stopping times and are piecewise constant between stopping times. That is $\Pi_t = \Pi_{\tau_{k,T}}$ for all $t \in [\tau_{k,T}, \tau_{k+1,T})$. In our setting, the variance within each block $[\tau_{k,T}, \tau_{k+1,T})$ is constant and almost surely bounded. Therefore, a sufficient condition for the asymptotic variance to be an almost surely finite random variable is that $T - \max_{k\in[K_T]}|\tau_{k+1,T} - \tau_{k,T}| = o_p(T)$.*

### 3.2 Estimating Conditional Variances

In cases where condition (3c) does not hold, whch can occur where there is adaptive sampling in every round $t \in [T]$, we must instead rely on Theorem 2 which requires us to construct a sequence of

$\mathcal{F}_{t-1}$ measurable estimators $\{\hat{\sigma}_t\}_{t=1}^T$. A natural strategy might be to directly estimate the functional,

$$\text{Var}\left(w_t\varphi_t \mid \mathcal{F}_{t-1}\right) = \mathbb{E}\left[w_t^2\varphi_t^2 \mid \mathcal{F}_{t-1}\right] - \mathbb{E}[w_t\varphi_t \mid \mathcal{F}_{t-1}]^2 = \mathbb{E}\left[w_t^2\varphi_t^2 \mid \mathcal{F}_{t-1}\right].$$

This approach would be burdensome because of the dependence of the estimand on several nuisance parameters: $\theta_{P_e}, \eta_{P_e}, \alpha_{P_e}, P_C, P_{Y|C,A}$. Instead, we construct $\hat{\sigma}_t$ by reweighting past observations to emulate draws from the evaluation policy at time $t$. This effectively reduces the problem to controlling a sequence of empirical averages over a policy class, which we handle via a maximal inequality.

We choose to target the following sequence of observable quantities:

$$\text{Var}\left(w_t\hat{\varphi}_t \mid \mathcal{F}_{t-1}\right) = \mathbb{E}\left[w_t^2\hat{\varphi}_t^2 \mid \mathcal{F}_{t-1}\right] - \mathbb{E}[w_t\hat{\varphi}_t \mid \mathcal{F}_{t-1}]^2 = \mathbb{E}\left[w_t^2\hat{\varphi}_t^2 \mid \mathcal{F}_{t-1}\right] - \mathbb{E}_{P_e}[\hat{\varphi}_t]^2.$$

At the oracle EIF, $\mathbb{E}_{P_e}[\varphi_t] = 0$ so the second term should be negligible. Our goal then is finding a sufficiently good estimate for $\mathbb{E}\left[w_t^2\hat{\varphi}_t^2 \mid \mathcal{F}_{t-1}\right]$. Now, note that for any $s < t$, $\hat{\varphi}_s(\cdot) := m_{\hat{\theta}_s}(\cdot) - \hat{\Psi}_s - \dot{\ell}_{\hat{\theta}_s,\hat{\eta}_s}(\cdot)[\hat{\alpha}_s])$ is composed of only $\mathcal{F}_{s-1}$ measurable known plug-in terms. Recalling that

$$\{\, f\,(Z_s) - \mathbb{E}[f(Z_s) \mid \mathcal{F}_{s-1}]\,\}_{s=1}^t$$

is a martingale difference sequence for any $\mathcal{F}_{s-1}$ measurable function $f$, consider some counterfactual policy $\pi$ and choose $f(z) := \frac{\pi_s(a|c)}{\pi(a|c)} w_s^2\hat{\varphi}_s^2(z) = \frac{\pi_e^2(a|c)}{\pi(a|c)\pi_s(a|c)}\hat{\varphi}_s^2(z)$. Intuitively, this map will change the measure to target the expectation of the reference policy because

$$\begin{aligned}\mathbb{E}\left[f(Z_s) \mid \mathcal{F}_{s-1}\right] &:= \mathbb{E}\left[\frac{\pi_e^2(A_s \mid C_s)}{\pi(A_s \mid C_s)\pi_s(A_s \mid C_s)}\hat{\varphi}_s^2(Z_s) \mid \mathcal{F}_{s-1}\right] \\ &= \mathbb{E}_{A\sim\pi}\left[\frac{\pi_e^2(A \mid C)}{\pi^2(A \mid C)}\hat{\varphi}_s^2(Z)\right].\end{aligned}$$

With this motivation in mind, a tempting way to target the quantity $\mathbb{E}\left[w_t^2\hat{\varphi}_t^2 \mid \mathcal{F}_{t-1}\right]$ would be to define the estimate

$$\hat{\sigma}_t^2 := \frac{1}{t-1}\sum_{s=1}^{t-1}\frac{\pi_e^2(A_s \mid C_s)}{\pi_t(A_s \mid C_s)\pi_s(A_s \mid C_s)}\hat{\varphi}_s^2(Z_s). \tag{9}$$

We justify the use of this estimator in Theorem 4. The key technical challenge apart from needing to derive uniform bounds over the policy class is that we are using prior influence function increments $\hat{\varphi}_s$ to target $\hat{\varphi}_t$ so we need to demonstrate these quantities converge sufficiently fast.

**Theorem 4.** *Assume all conditions of Theorem 2 are satisfied aside from condition (2d). Assume there exists a $c > 0$ such that $\min(\sigma_t, \hat{\sigma}_t) > c$ for all $t \in [T]$ almost surely. Recall that $\hat{\varphi}_s(\cdot) := m_{\hat{\theta}_s}(\cdot) - \hat{\Psi}_s - \dot{\ell}_{\hat{\theta}_s,\hat{\eta}_s}(\cdot)[\hat{\alpha}_s]$, where $\hat{\Psi}_s$ is any $\mathcal{F}_{s-1}$ measurable estimator such that $\frac{1}{T}\sum_{s=1}^T(\hat{\Psi}_s - \Psi_e)^2 = o_p(T^{-1/2})$ and $\sup_{s\in[T]}|\hat{\Psi}_s| = O_p(1)$.[1] Moreover assume that there exists a class of policies $\Pi_T$ such that $\pi_t \in \Pi_T$ almost surely for all $t \in [T]$ and*

*(4a) There exists $p \in [0, 2)$ such that for all $\epsilon \in (0, 1)$,*

$$\log N(\epsilon, \log\Pi_T, \|\cdot\|_\infty) \lesssim \begin{cases}\log(e/\epsilon), & p = 0, \\ \epsilon^{-p}, & p \in (0, 2),\end{cases},$$

*where $\log\Pi_T := \{(c, a) \mapsto \log\pi(a \mid c) : \pi \in \Pi_T\}$.*

*(4b) $\sup_{\pi\in\Pi}\sup_{c,a\in\mathcal{C}\times\mathcal{A}}\frac{\pi_e(a|c)}{\pi(a|c)} = O(T^{1/8})$.*

*Then, if $\hat{\sigma}_t^2$ is constructed as in Equation (9) with $\hat{\sigma}_1^2$ initialized as a fixed positive finite constant, condition (2d) is satisfied. That is, $\frac{1}{T}\sum_{t=1}^T \sigma_t^2/{\hat{\sigma}_t}^2 \xrightarrow{p} 1$.*

Condition (4a) controls the complexity of the policy class through an entropy bound; in practice, such conditions are satisfied by policy classes driven by parametric or other low-complexity models. We emphasize that this restrictions only applies to the *policy* class, nuisance and parameter estimates can still rely on flexible black box approaches. Condition (4b)'s stronger overlap requirement relative to Condition (2c) is the price of uniform empirical-process control over $\Pi$. However, it remains weaker than the assumptions in previous adaptive $M$-estimation work (Zhang et al., 2021; Guo and Xu, 2025; Leiner et al., 2026), which require importance weights that are uniformly bounded in time.

[1] Note that we provide an explicit construction of an estimator $\hat{\Psi}_t$ fulfilling these requirements in Appendix C.

## 4 Experiments

We discuss a dynamic pricing application as a prototypical example of the utility of our approach. We consider a sequential pricing problem in which, at each time period $t \in [T]$, the decision-maker selects a price $p_t \in \mathcal{A}$ and then observes a context $C_t \in \mathcal{C}$ corresponding to customer attributes and a reward $Y_t \in \{0, 1\}$ corresponding to either customer renewal ($Y_t = 1$) or attrition ($Y_t = 0$).[2] We model renewal decisions using a random utility framework. Let $\varepsilon_{0t}, \varepsilon_{1t} \overset{iid}{\sim} \text{Gumbel}(0, 1)$ and define,

$$U_t^1 = g^\star + \tau g_0(C_t) - \beta p_t + \varepsilon_{1t}, \qquad U_t^0 = \varepsilon_{0t}.$$

where $U_t^1$ denotes the utility of renewing and $U_t^0$ denotes the utility of attrition for individual $t$. The parameter $\beta > 0$ governs price sensitivity and is the primary target of inference. The function $g_0(X)$ captures baseline heterogeneity in demand, while $g^\star$ controls the overall level of demand. $\varepsilon_{0t}$ and $\varepsilon_{1t}$ are used to enforce idiosyncratic variation in the utility function of individuals. The observed outcome is

$$Y_t = \mathbf{1}\{U_t^1 \geq U_t^0\} = \mathbf{1}\{g^\star + \tau g_0(C_t) - \beta p_t + \varepsilon_{1t} - \varepsilon_{0t} \geq 0\}.$$

The difference of two independent standard Gumbel random variables follows a logistic distribution. Consequently, $\varepsilon_{1t} - \varepsilon_{0t} \sim \text{Logistic}(0, 1)$, and the renewal probability is $\Pr(Y_t = 1 \mid C_t, p_t) = \Lambda(g^\star + \tau g_0(C_t) - \beta p_t)$, where $\Lambda(u) = (1 + e^{-u})^{-1}$ is the sigmoid function.

**Sampling policies.** We primarily consider two sampling policies. The first, which we label *revenue maximization*, uses past data to learn a price that maximizes expected revenue over the population:

$$p^\star = \arg\max_{p \in \mathcal{A}} \mathbb{E}_{P_C} \left[ p \cdot \Lambda\big(g^\star + \tau g_0(C) - \beta p\big) \right].$$

The decision-maker uses an $\epsilon$-greedy allocation: the price with the highest estimated revenue receives probability $1 - \epsilon$, while the remaining probability $\epsilon$ is divided equally among the other prices.

The second policy, which we term *model auditing*, uses a planning model with price sensitivity $\beta_{\text{ref}} = 1$. At each policy update, the analyst tests $H_0 : \beta = \beta_{\text{ref}}$ at a nominal 5% significance level. If the null is rejected, prices are sampled uniformly to collect broader evidence about demand. Otherwise, the price recommended by the planning model receives probability $1-\epsilon$, and the remaining probability $\epsilon$ is divided equally among the other prices. We set $\epsilon = 0.1$ throughout. Further details and additional sampling rules appear in Appendix B.2.

**Implementation details.** We use $\mathcal{A} = \{0.5, 1, \ldots, 5\}$ and independent attributes $C_j = B_j G_j$, with $B_j \sim \text{Bernoulli}(1/2)$ and $G_j \sim N(0, 1)$. The fitted model is

$$m_{\alpha,\beta}(C, p) = \Lambda\{H(C)^\top \alpha - \beta p\}, \qquad H(C) = (1, C_1 C_2, \sin(C_2) C_3, C_4^2, C_5 C_6)^\top,$$

estimating all coefficients while keeping price sensitivity constant. For revenue maximization and auditing, true nuisance coefficients are $\alpha_0 = (2.5 + \log(1.5), 0.5, 0.5, 0.25, -0.125)^\top$. Correct specification uses $\beta = 1$; misspecification replaces it in the true utility by $\beta(C) = b + 1.2 \tanh(C_1)$, with $b \approx 1.426$. All methods target the logistic-loss projection $\beta_e \approx 1$ under uniform evaluation prices. Some implementations uses different intercept and sensitivity parameters that tie optimal revenues at prices 2 and 2.5, preserving $\beta_e \approx 1$. Each setting has 2000 replications at $T = 50000$. One uniform pilot uses 20% of observations to fit and initialize the methods requiring it; the remaining 80% enter their estimating equations. GLM and IPW regression use all $T$ observations.

**Results.** We compare equally weighted GLMs, IPW-weighted GLMs (labeled IPW), AdapTZ (Lin et al., 2025), self-normalization (Theorem 3), and stabilized IPW (Theorems 2 and 4). Both GLM regressions use sandwich standard errors. AdapTZ performs well under correct specification, while ordinary IPW achieves near-nominal coverage under revenue maximization even under misspecification. Model auditing induces undercoverage for IPW, and heterogeneous demand also leads to undercoverage for AdapTZ, outside its common-coefficient assumptions; stabilized IPW maintains near-nominal coverage across these settings.

[2] This setup differs slightly from the standard contextual bandit formulation, since the action space is restricted to context-independent pricing policies. We focus on this setting because fully personalized pricing (i.e., price discrimination based on customer attributes) is often infeasible in practice due to regulatory, fairness, or operational constraints.

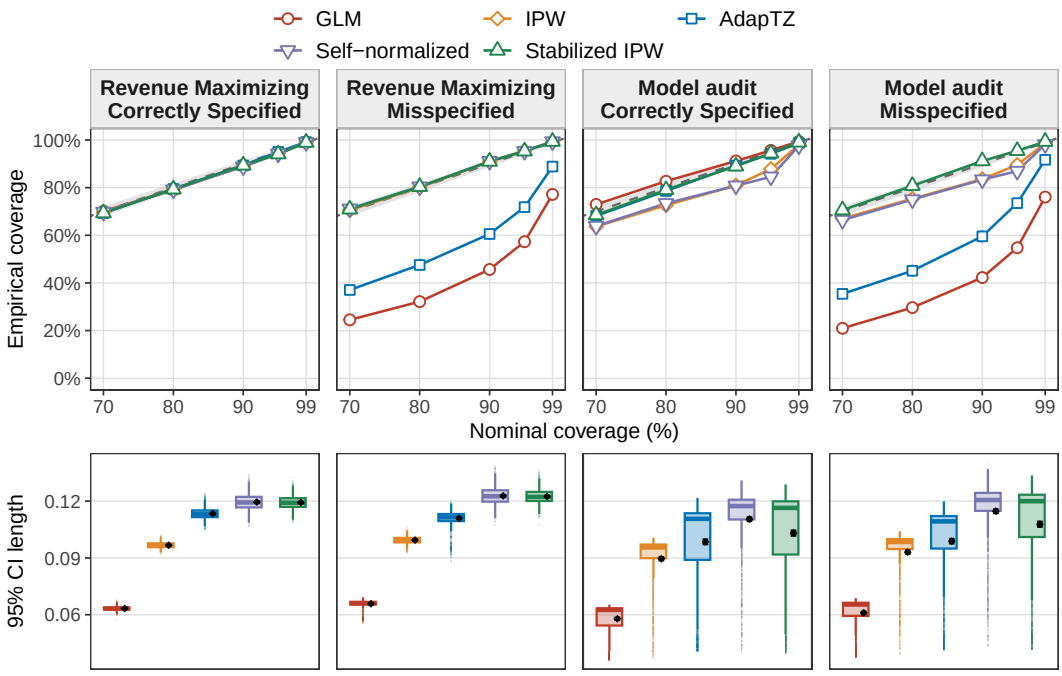


Figure 1: Dynamic pricing: empirical coverage (top) and 95% CI length (bottom), $T = 50000$, 2000 replications. Coverage bars are 95% Wilson intervals; shading marks two Monte Carlo standard errors around the diagonal. CI panels share a scale and follow the legend order. Boxes show medians and interquartile ranges; black diamonds and bars show means and 95% Monte Carlo intervals. All valid intervals are retained.

## 5 Conclusion

We have developed a unified framework for inference on smooth functionals of nonparametric M-estimands under adaptive data collection. Our two asymptotic normality results—one based on a novel reweighted plug-in variance estimator and one based on self-normalized influence function increments—provide valid confidence intervals across a range of adaptive sampling regimes. Our results permit flexible estimation of nuisance parameters and are valid under model misspecification.

Several directions remain open. Generalizing beyond inverse propensity reweighting to handle arbitrary time-varying loss reweighting would allow us to unify a larger subset of work in the adaptive inference literature. Moreover, although we have shown that our constructions are asymptotically normal, it is an open question whether they achieve semiparametric efficiency bounds. Finally, our results assume that sampling policies are known, a natural question is under what conditions the sampling policies can be estimated while preserving asymptotic normality.

### Acknowledgments

Aaditya Ramdas acknowledges the support of NSF Grant DMS-2610252.

## References


Afsar, M. M., T. Crump, and B. Far (2022). Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys 55*(7), 1–38.

Bibaut, A., A. Chambaz, M. Dimakopoulou, N. Kallus, and M. Laan (2021, 12). Post-contextual-bandit inference. *Advances in Neural Information Processing Systems 34*, 28548–28559.

Bickel, P. J. and C. A. Klaassen (1993). *Efficient and adaptive estimation for semiparametric models*, Volume 4. Springer.

Billingsley, P. (1961). The lindeberg-lévy theorem for martingales. *Proceedings of the American Mathematical Society 12*(5), 788–792.

Caria, A. S., G. Gordon, M. Kasy, S. Quinn, S. O. Shami, and A. Teytelboym (2024, 04). An adaptive targeted field experiment: Job search assistance for refugees in jordan. *Journal of the European Economic Association 22*(2), 781–836.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., W. K. Newey, and R. Singh (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica 90*(3), 967–1027.

Cook, T., A. Mishler, and A. Ramdas (2024). Semiparametric efficient inference in adaptive experiments. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, Volume 236 of *Proceedings of Machine Learning Research*, pp. 1033–1064.

Deshpande, Y., L. Mackey, V. Syrgkanis, and M. Taddy (2018). Accurate inference for adaptive linear models. In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80 of *Proceedings of Machine Learning Research*, pp. 1194–1203.

Dvoretzky, A. (1972). Asymptotic normality for sums of dependent random variables. *Berkeley Symp. on Math. Statist. and Prob.*.

Girard, S., A. Bibaut, J.-J. Vie, A. Gretton, N. Kallus, and H. Zenati (2026). Fast best-in-class regret for contextual bandits. In *Forty-Second Annual Conference on Uncertainty in Artificial Intelligence*.

Guo, Y. and Z. Xu (2025). Statistical inference for misspecified contextual bandits. *arXiv:2509.06287*.

Hadad, V., D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences 118*(15), e2014602118.

Hall, P. and C. C. Heyde (1980). *Martingale limit theory and its application*. Academic press.

Hardy, G., J. Littlewood, and G. Pólya (1953). Inequalities. *The Mathematical Gazette 37*(321), 236–236.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

Kato, M. (2020). Confidence interval for off-policy evaluation from dependent samples via bandit algorithm: Approach from standardized martingales. *arXiv:2006.06982*.

Kato, M., A. Oga, W. Komatsubara, and R. Inokuchi (2024). Active adaptive experimental design for treatment effect estimation with covariate choice. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 23291–23323.

Khamaru, K., Y. Deshpande, T. Lattimore, L. Mackey, and M. J. Wainwright (2025). Near-optimal inference in adaptive linear regression. *The Annals of Statistics 53*(6), 2329 – 2355.

Lai, T. L. and C. Z. Wei (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics 10*(1), 154–166.

Leiner, J., R. Dunn, and A. Ramdas (2026). Adaptive off-policy inference for m-estimators under model misspecification. *arXiv:2509.14218*.

Lin, L., K. Khamaru, and M. J. Wainwright (2025). Semiparametric inference based on adaptively collected data. *The Annals of Statistics 53*(3), 989 – 1014.

Liu, S., K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, and M. Feng (2020, Jul). Reinforcement learning for clinical decision support in critical care: Comprehensive review. *J Med Internet Res 22*(7), e18477.

Luedtke, A. R. and M. J. van der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics 44*(2), 713 – 742.

Rotnitzky, A., E. Smucler, and J. M. Robins (2021, 03). Characterization of parameters with a mixed bias property. *Biometrika 108*(1), 231–238.

Shin, J., A. Ramdas, and A. Rinaldo (2019). Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems*.

van der Laan, L., A. Bibaut, N. Kallus, and A. Luedtke (2026). Automatic debiased machine learning for smooth functionals of nonparametric m-estimands. *arXiv:2501.11868*.

van der Vaart, A. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Whitehouse, J., E. Brunskill, and V. Syrgkanis (2026). Inference on the value of policy improvement in adaptive experiments. *arXiv:2610.15789*.

Whitehouse, J., Q. Chen, M. Austern, and V. Syrgkanis (2026). Inference on optimal policy values and other irregular functionals via softmax smoothing. *arXiv:2507.11780*.

Ying, M., K. Khamaru, and C.-H. Zhang (2023). Adaptive linear estimating equations. In *Advances in Neural Information Processing Systems*, Volume 36, pp. 52051–52072.

Yom-Tov, E., G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, and I. Hochberg (2017, Oct). Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *J Med Internet Res 19*(10), e338.

Zhang, K., L. Janson, and S. Murphy (2020). Inference for batched bandits. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 9818–9829.

Zhang, K., L. Janson, and S. Murphy (2021). Statistical inference with M-estimators on adaptively collected data. In *Advances in Neural Information Processing Systems*, Volume 34, pp. 7460–7471.

Zhang, K. W., L. Janson, and S. A. Murphy (2023). Statistical inference after adaptive sampling for longitudinal data. *arXiv:2202.07098*.

Zhao, Y., M. R. Kosorok, and D. Zeng (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine 28*(26), 3294–3315.

# A Common Target Estimands and Relationship to Prior Work

In this section, we note how this framework can be used to target estimands considered in a variety of precursor work in the adaptive and contextual bandit settings. Table 2 summarizes some of the key differences in the assumptions of this work and precursor methods.

## A.1 Average Treatment Effect

Let $A \in \{0, 1\}$ be a binary treatment. Let $\Theta := L_2(P_e)$ and the nuisance space be trivial ($\Upsilon = \{0\}$) as there is no nuisance parameter. Assume a square loss function with $\ell_\theta(c, a, y) = (y - \theta(a, c))^2$. Then the $M$-estimand is $\theta_{P_e}(a, c) := \mathbb{E}[Y|A = a, C = c]$. Letting $m_\theta(c, a, y) := \theta(1, c) - \theta(0, c)$, the target becomes

$$\Psi_e(\theta_{P_e}) = \mathbb{E}_{P_e}\left[m_{\theta_{P_e}}(C, A, Y)\right] = \mathbb{E}_{P_C}\left[\theta_{P_e}(1, C) - \theta_{P_e}(0, C)\right].$$

This generalizes straightforwardly to any smooth functional of the conditional regression function, subsuming the frameworks of Bibaut et al. (2021) and Hadad et al. (2021).

### A.2 Parametric $M$-estimators

Let $\Theta = \mathbb{R}^d$ and the nuisance space be trivial (e.g. $\Upsilon = \{0\}$) as there is no nuisance parameter. Choose any loss $\ell_\theta(z)$ and let $\theta_{P_e} = \operatorname{argmin} \mathbb{E}_{P_e}[\ell_\theta(C, A, Y)]$. For any fixed $v \in \mathbb{R}^d$ (such as a basis vector $e_j$ to target a coordinate of $\theta$), we can then let $m_\theta = v^T\theta$ and consider $\Psi_e(\theta_{P_e}) = v^T\theta_{P_e}$. This subsumes the frameworks of Leiner et al. (2026) and Guo and Xu (2025).

We note that the approach of Zhang et al. (2021) cannot be put into this form as their construction uses square root inverse propensity weights. Under correct specification, the target estimand would be unchanged, but under misspecification, their procedure covers the time-varying random target

$$\operatorname*{argmin}_{\theta\in\Theta} \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{\frac{\pi_e(A_t \mid C_t)}{\pi_t(A_t \mid C_t)}}\ell_\theta(Z_t) \mid C_t, \mathcal{F}_{t-1}\right].$$

Generalizing our method to cover arbitrary time-varying weightings of the loss function would be an interesting extension for future work.

### A.3 Partial Linear and Generalized Linear Regression

We partition the covariates $X = (W, V)$, where $W \in \mathbb{R}^d$ enters the model linearly and $V$ enters the model non parametrically. Consider the model

$$Y = g\left(W^T\theta + \eta_{P_e}(V)\right) + \epsilon,$$

where $g : \mathbb{R} \to \mathbb{R}$ is a known inverse link function and $\mathbb{E}[\epsilon \mid C, A] = 0$. Then $\Theta = \mathbb{R}^d$ and $\Upsilon = L_2(P_C)$ and $\ell_{\theta,\eta}(c, a, y) = \left(y - g(w^T\theta + \eta(v))\right)^2$ where $(w, v)$ are components of (c,a). We then have $\theta_{P_e} = \operatorname{argmin}_{\theta\in\mathbb{R}^d} \mathbb{E}_{P_e}\left[\ell_{\theta,\eta_{P_e}}(C, A, Y)\right]$. For any fixed $v \in \mathbb{R}^d$, we can then let $m_\theta = v^T\theta$ and consider $\Psi_e(\theta_{P_e}) = v^T\theta_{P_e}$. This corresponds to the setting of Lin et al. (2025).

### A.4 Value of an Optimal Policy over a Policy Class

We first start with a naive solution. Let $\Theta := L_2(P_e)$ and the nuisance space be trivial ($\Upsilon = \{0\}$) so there is no nuisance parameter. Assume a square loss function with $\ell_\theta(c, a, y) = (y - \theta(a, c))^2$. As before, the $M$-estimand is $\theta_{P_e}(a, c) := \mathbb{E}[Y|A = a, C = c]$. Letting $\Psi_e(\theta_{P_e}) := \mathbb{E}_C\left[\max_{a\in\mathcal{A}} \theta_{P_e}(a, C)\right]$ correspond to the expected value of the optimal policy. Note that this functional is differentiable under a condition that ensures the optimal action given each context $C$ is unique almost surely.

However, this construction can be restrictive in that it requires the optimal action to be unique almost surely. This condition can be relaxed by optimizing over a policy class. Let $\Pi$ denote a class of policies, define $\theta(a, c) = \mathbb{E}[Y \mid A = a, C = c]$, and consider the functional

$$\Psi_e(\theta) = \sup_{\pi\in\Pi} \mathbb{E}_{P_C}\left[\int_{\mathcal{A}} \pi(a \mid C)\, \theta(a, C)\, \mu(da)\right].$$

This formulation avoids requiring a unique optimal action almost surely. Instead, it suffices that the population optimization problem over $\Pi$ admits a unique maximizer $\pi^\star$.

Under standard regularity conditions, the envelope theorem implies that for any perturbation $h \in \Theta$,

$$\partial_\theta\Psi_e(\theta)[h] = \mathbb{E}_{P_C}\left[\int_{\mathcal{A}} \pi^\star(a \mid C)\, h(a, C)\, \mu(da)\right].$$

The Riesz representer $\alpha_{P_e}$ can therefore be defined by the relation

$$\mathbb{E}_{P_C}\left[\int_{\mathcal{A}} \pi^\star_\theta(a \mid C)\, h(a, C)\, \mu(da)\right] = E_{P_e}\left[\partial^2_\theta\ell_{\theta_{P_e}}(Z)[\alpha_{P_e}, h]\right] \quad \text{for all } h \in \Theta.$$

Under the square loss $\ell_\theta(c, a, y) = (y - \theta(a, c))^2$, we have

$$E_{P_e}\left[\partial^2_\theta\ell_{\theta_{P_e}}(Z)[\alpha, h]\right] = 2\, E_{P_e}[\alpha(A, C)\, h(A, C)].$$

| Work | Target functional | Misspec. allowed | Adaptive data | Policy known | Overlap assumption | Other assumptions |
|---|---|---|---|---|---|---|
| Lai and Wei (1982) | linear coef. | no | yes | no | implicit | design stability |
| Luedtke and van der Laan (2016) | optimal value | n/a | no | no | uniform | variance estimable |
| Deshpande et al. (2018); Khamaru et al. (2025); Ying et al. (2023) | linear coef. | no | yes | no | implicit | online debiasing |
| Hadad et al. (2021) | arm means / policy value | n/a | yes | yes | vanishing | variance converges |
| Bibaut et al. (2021) | policy value / contrasts | n/a | yes | yes | vanishing | variance estimable |
| Zhang et al. (2020) | linear coef. | no | batched | no | uniform | fixed no. of batches |
| Zhang et al. (2021) | $M$-est. param. | no | yes | yes | uniform | square-root IPW |
| Zhang et al. (2023) | $Z$-projection | yes | yes | yes | uniform | longitudinal; fixed horizon |
| Lin et al. (2025) | partial-linear coef. | no | yes | moments | vanishing | nuisance consistency |
| Lin et al. (2025) | GLM coef. | no | yes | moments | vanishing | reference-arm floor; nuisance rates |
| Guo and Xu (2025) | $Z$-projection | yes | yes | yes | uniform | deterministic policy limit |
| Leiner et al. (2026) | $M$-projection | yes | yes | yes | uniform | variance estimable |
| This paper | smooth $M$-functionals | yes | yes | yes | vanishing | variance estimable |

Table 2: Positioning relative to related work. The target-functional column summarizes both the estimand and the generality of the working model. The overlap column distinguishes uniform positivity from permitted overlap that is permitted to vanish at some rate. The final column highlights some key assumptions for each method.

Matching terms yields the Riesz representer $\alpha_{P_e}(a,c) = \frac{1}{2}\frac{\pi^\star(a|c)}{\pi_e(a|c)}$. The corresponding efficient influence function is therefore

$$\varphi_{P_e}(Z) = \int_{\mathcal{A}} \pi^\star(a \mid C)\, \theta_{P_e}(a, C)\, \mu(da) - \Psi_e(\theta_{P_e}) + \frac{\pi^\star(A \mid C)}{\pi_e(A \mid C)}\big(Y - \theta_{P_e}(A, C)\big).$$

We note that this construction can potentially be extended to settings where the optimal policy is non-unique, although the value functional need not be pathwise differentiable in this case. Luedtke and van der Laan (2016) characterize conditions for pathwise differentiability of the optimal policy value and develop inference procedures even when these conditions fail. More recently, Whitehouse et al. (2026) develop an approach based on softmax smoothing to obtain inference on optimal policy values without requiring uniqueness, while Whitehouse et al. (2026) establish asymptotically normal inference for the value of policy improvement over a restricted policy class under adaptive sampling, also allowing non-unique optima.

## B Additional Simulation Details

This section describes the policy-update regimes, implementation, and sampling rules used in the dynamic pricing experiments.

### B.1 Different levels of adaptivity

We distinguish the *sampling rule*, which proposes a distribution over prices using past data, from the *update regime*, which determines when that proposal replaces the current distribution. Let

$m = \lfloor 0.2T \rfloor$ be the pilot size and $n = T - m$ the number of post-pilot observations. Index post-pilot rounds by $k = 1, \ldots, n$. The first 50 rounds use uniform prices; the first policy update occurs at $k = 51$. We consider four regimes:

1. **Sublinear adaptation followed by commitment.** Update at every round through $k = 51 + \lceil \sqrt{n} \rceil$, then hold the sampling distribution fixed. The updating period occupies a vanishing fraction of the post-pilot horizon. Although the policy eventually stops changing within a trajectory, its committed distribution depends on the observed history, so its associated variance can remain path dependent.
2. **Linear adaptation followed by commitment.** Update at every round through $k = 51 + \lfloor n/2 \rfloor$, then freeze the distribution. Feedback now continues for a nonvanishing fraction of the sample, allowing us to assess how inference changes when adaptation persists well beyond an initial learning period.
3. **Piecewise constant policies with random updates.** After the first update, independently refresh the policy with probability $0.3$ each round; otherwise retain the preceding distribution. Data continue to accumulate between updates. This creates locally fixed sampling distributions with repeated opportunities to change them. "Random" in the figures refers to these update times, not to uniformly random prices.
4. **Fully adaptive policies.** Refresh the distribution before every post-warm-up observation using all available past data. This permits ongoing feedback between estimation and sampling throughout the experiment.

An update recomputes the entire distribution, including its exploration probability and favored price. Freezing the distribution does not stop data collection or estimation.

### B.2 Implementation details

**Demand model and target.** We use the price menu $\mathcal{P} = \{0.5, 1, \ldots, 5\}$ and independent customer attributes $C_j = B_j G_j$, with $B_j \sim \text{Bernoulli}(1/2)$ and $G_j \sim N(0, 1)$. The fitted model is

$$m_\theta(C, p) = \Lambda\{H(C)^\top \alpha - \beta p\}, \qquad H(C) = (1, C_1 C_2, \sin(C_2) C_3, C_4^2, C_5 C_6)^\top,$$

where $\theta = (\alpha^\top, \beta)^\top$ and $\Lambda(u) = (1 + e^{-u})^{-1}$. We estimate all coefficients while keeping price sensitivity constant. The true nuisance coefficients are $\alpha_0 = (a, 0.5, 0.5, 0.25, -0.125)^\top$. Correct specification uses $\beta = 1$; misspecification replaces it in the true utility by $\beta(C) = b + 1.2 \tanh(C_1)$. All methods target the price-sensitivity coefficient $\beta_e \approx 1$ of the logistic-loss projection under uniform evaluation prices. For the estimation formulas below, write $X(C, p) = (H(C)^\top, -p)^\top$.

**Computing environment.** Simulations ran locally on an Apple M4 MacBook Air (10 CPU cores, 16 GB RAM; macOS 15.7.7), using R 4.6.0 and C++ via Rcpp 1.1.1.1.1. Replications ran in up to 16 CPU worker processes with fixed seeds.

### B.3 Sampling policies

At an eligible update, each rule below proposes a price distribution. The current price is then drawn from that distribution before observing the customer's attributes or purchase outcome. Between updates, prices are drawn from the previously stored distribution. Define

$$U(p) = \frac{1}{10}, \qquad E(p; q) = \begin{cases} 0.90, & p = q, \\ 0.10/9, & p \neq q. \end{cases}$$

Let $N_{t-1,p}$ and $S_{t-1,p}$ count previous offers and purchases at price $p$. The smoothed empirical revenue and its maximizer are

$$\widehat{r}_{t-1}(p) = p \frac{S_{t-1,p} + 1/2}{N_{t-1,p} + 1}, \qquad \widehat{p}_t = \arg\max_{p \in \mathcal{P}} \widehat{r}_{t-1}(p).$$

We add one-half to the purchase and non-purchase counts, shrinking small-sample purchase rates toward $1/2$ and ensuring that the estimate is defined even for prices with no observations. Counts include the pilot except for confidence deployment and probability matching, which start new revenue

counts after the pilot. This reset affects their allocation decisions; the pilot nuisance fits and variance history are retained.

The relevant sampling policies we test are enumerated below.

**Revenue maximization.** The decision maker concentrates on the current empirical revenue leader while retaining exploration at every price. At an update, set $\pi_t(p) = E(p; \widehat{p}_t)$. Thus the favored price receives probability exactly $0.90$, and each of the other nine prices receives $0.10/9$. This is the conventional revenue-driven benchmark for the statistical allocation rules that follow.

**Model auditing.** A retailer begins with a planning model that assumes sensitivity $\beta_{\text{ref}} = 1$. Fit its nuisance coefficients on the pilot with this sensitivity held fixed, obtaining $\widehat{\alpha}_{\text{ref}}$, and choose the planning price

$$p_{\text{ref}} = \arg\max_{p \in \mathcal{P}} \frac{p}{m} \sum_{i=1}^{m} \Lambda\{H(C_i)^\top \widehat{\alpha}_{\text{ref}} - \beta_{\text{ref}} p\}.$$

The benchmark is a demand coefficient, not a prespecified price. After $j$ post-pilot observations, form the running monitor

$$B_j = \frac{\sum_{s=1}^{j} w_{m+s} F_{m+s}}{\sum_{s=1}^{j} w_{m+s}}, \qquad s_j = \frac{\{\sum_{s=1}^{j} w_{m+s}^2 (F_{m+s} - B_j)^2\}^{1/2}}{\sum_{s=1}^{j} w_{m+s}}, \qquad Z_j = \frac{B_j - \beta_{\text{ref}}}{s_j}.$$

At round $k$, an eligible update sets $\pi_t = U$ if $|Z_{k-1}| > \Phi^{-1}(0.975)$ and $\pi_t = E(\cdot; p_{\text{ref}})$ otherwise. The retailer therefore uses its planning price while the benchmark remains compatible with the data, and broadens data collection when the model appears unreliable. The rule can switch back after later reviews. The monitor $B_j$ is an ordinary weighted one-step calculation; the final baseline labeled IPW is the weighted regression described above.

**Test-and-deploy.** This rule represents a different statistical objective: continue broad experimentation until the benchmark can be rejected, then deploy a price selected from the accumulated revenue data. Using the same monitor, set $\pi_t = U$ while $|Z_{k-1}| \le \Phi^{-1}(0.975)$, and $\pi_t = E(\cdot; \widehat{p}_t)$ otherwise. The rule is reversible, so a later review can return the retailer to uniform exploration.

**Permanent deployment.** Use the same test-and-deploy rule until its first rejection at an eligible update. Thereafter freeze the selected policy. Later updates can still change $\widehat{p}_t$ as revenue estimates change, but cannot restore uniform exploration. This represents a one-time decision to end broad experimentation while continuing to improve the deployed price.

**Confidence-based deployment.** Here deployment is based directly on uncertainty about revenue rankings. For each price, use a Jeffreys–Beta purchase-rate posterior, approximating the associated revenue by independent normal variables

$$R_p \sim N\left(p u_p, \frac{p^2 u_p (1 - u_p)}{N_p + 2}\right), \qquad u_p = \frac{S_p + 1/2}{N_p + 1}, \qquad \gamma_p = \Pr(R_p \ge R_q \text{ for every } q).$$

At an update, offer uniform prices if $\gamma_{\widehat{p}_t} < 0.95$; otherwise set $\pi_t(p) = 0.02/10 + 0.98\mathbf{1}\{p = \widehat{p}_t\}$. The rule can return to exploration if confidence subsequently falls. It models experimentation until there is sufficient evidence to deploy a revenue leader. The tied-revenue environment makes identifying a unique leader difficult, without requiring a policy to switch prices by design.

### B.4 Results across adaptivity levels and sampling policies

Figures 2–5 cross all sampling rules with the four update regimes and both correct and incorrect specifications at $T = 50000$, using 2000 replications per setting. Each policy has separate Correctly Specified and Misspecified plots, with update regimes as columns. Coverage curves appear above 95% CI lengths. CI boxes show medians and interquartile ranges; black diamonds and bars show means and 95% Monte Carlo intervals.

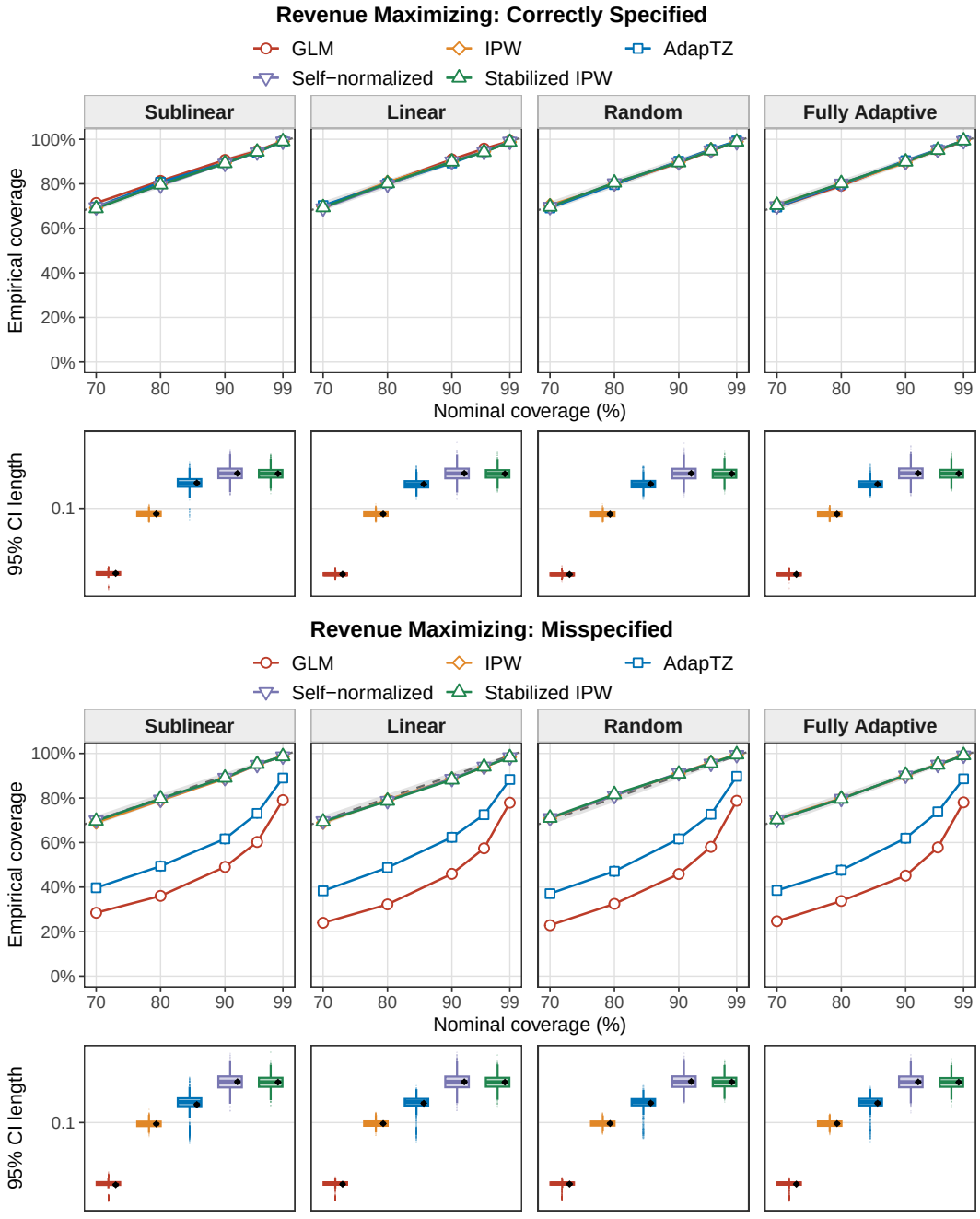


Figure 2: Revenue Maximizing: Correctly Specified (top) and Misspecified (bottom), $T = 50000$.

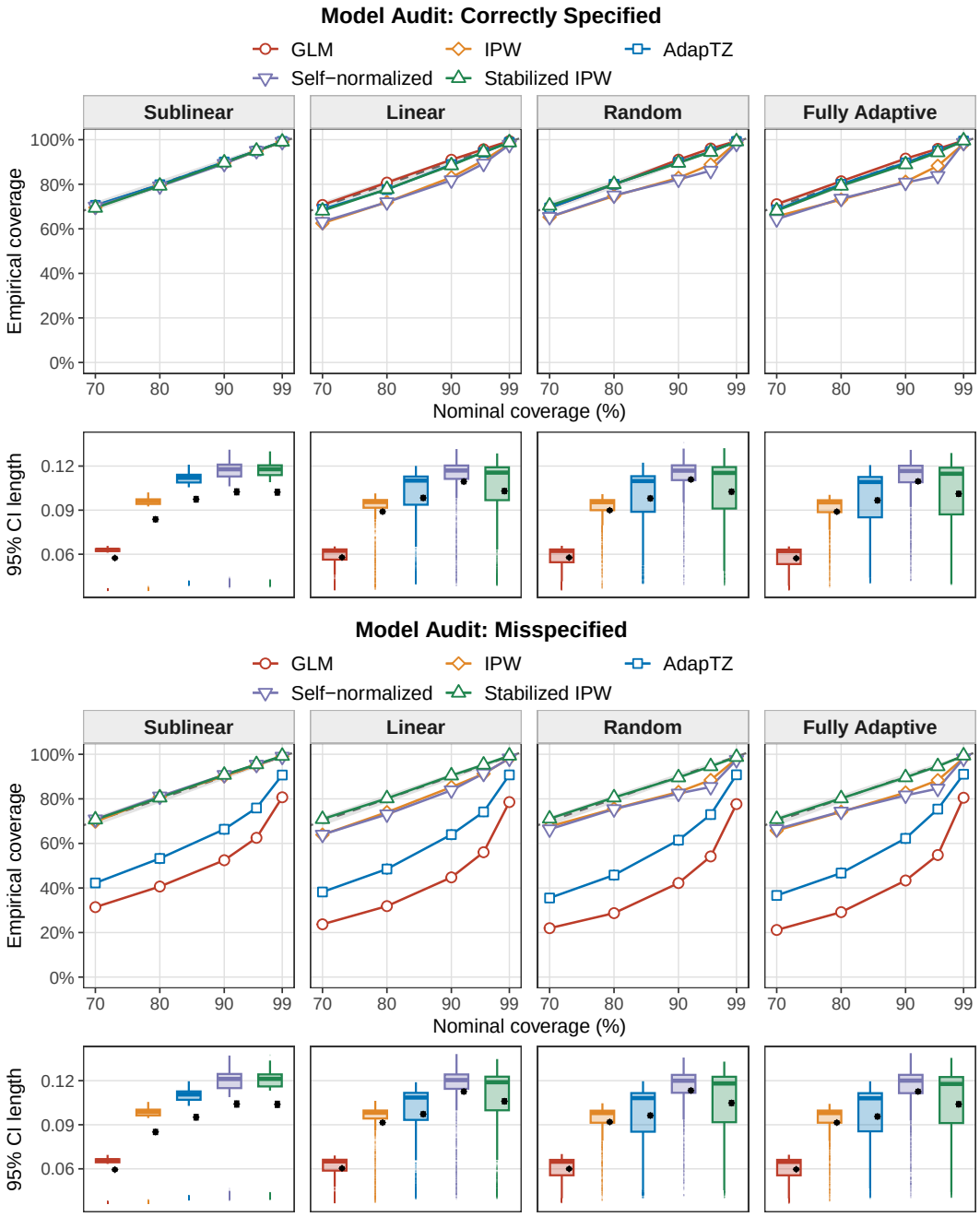


Figure 3: Model Audit: Correctly Specified (top) and Misspecified (bottom), $T = 50000$.

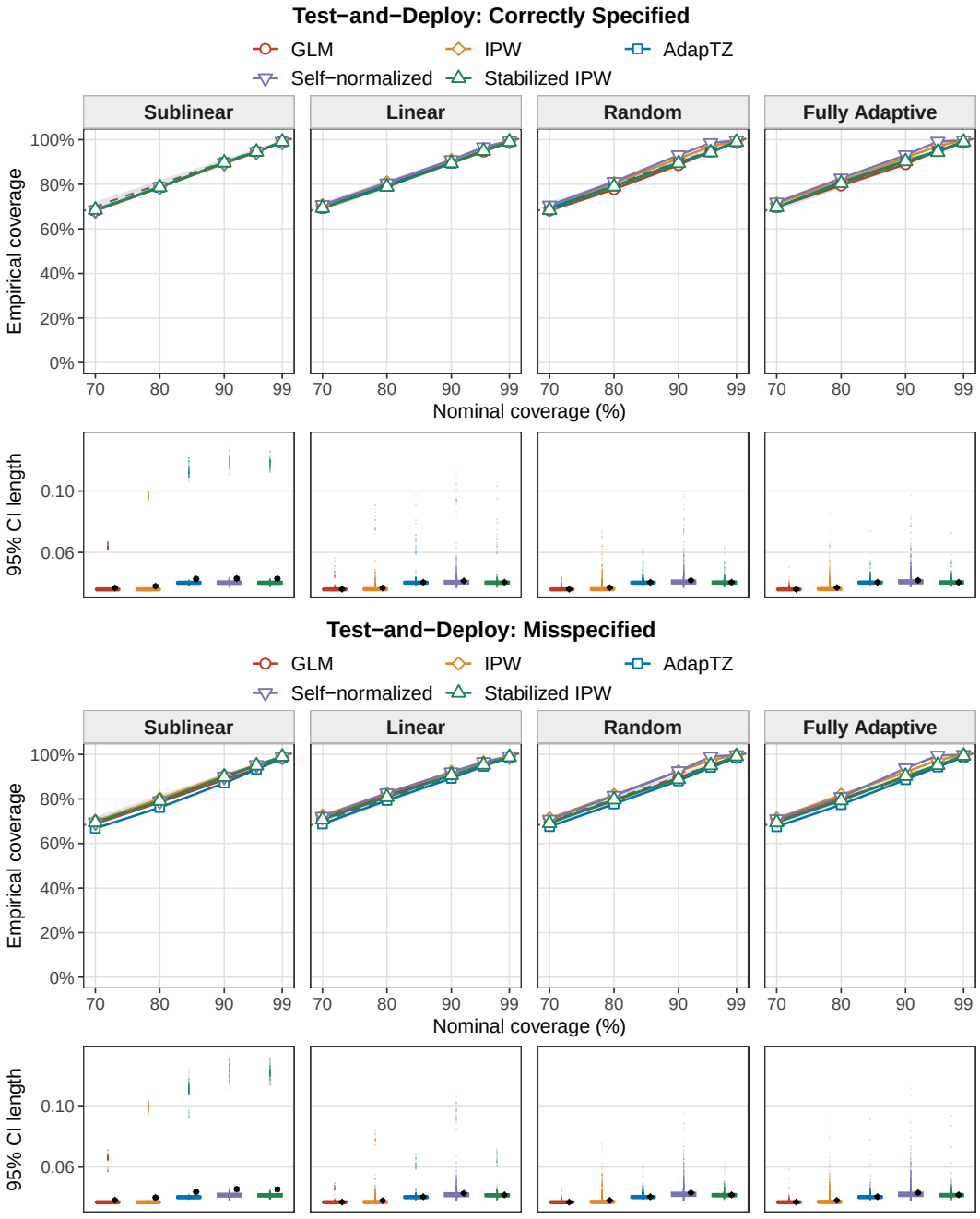


Figure 4: Test-and-Deploy: Correctly Specified (top) and Misspecified (bottom), $T = 50000$.

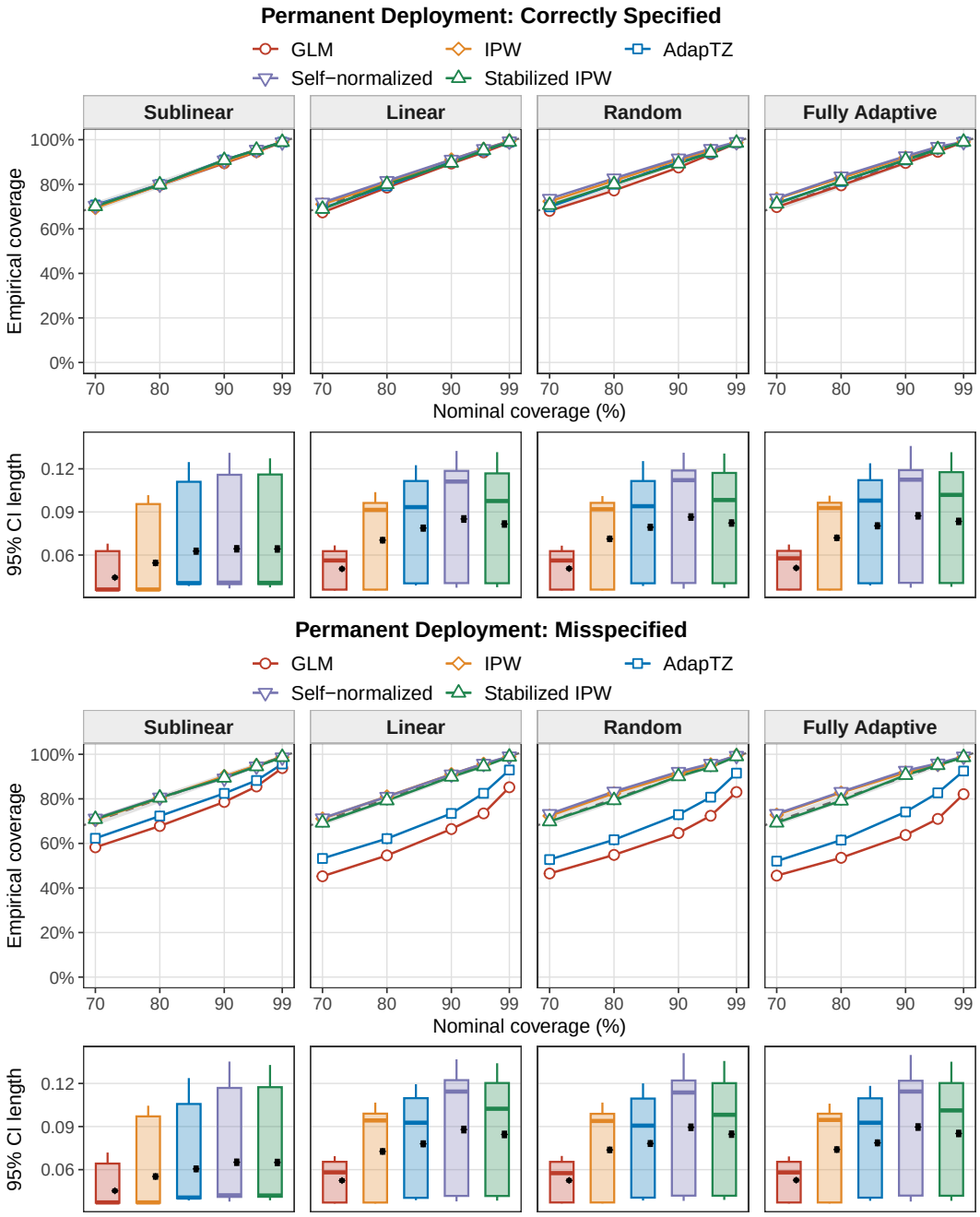


Figure 5: Permanent Deployment: Correctly Specified (top) and Misspecified (bottom), $T = 50000$.

## C Practical Considerations for Estimating Conditional Variances

We note that Theorem 4 requires the plug-in centering estimator $\hat{\Psi}_t$ to satisfy mild boundedness and consistency conditions. A convenient way to guarantee these properties is through sample splitting. Proposition 1 shows that if the nuisance and parameter estimates are trained on an initial burn-in block and the centering term is estimated on a separate holdout block, then the required rate condition follows automatically under the regularity assumptions of the paper. This is not a particularly onerous requirement, as we already recommend a burn-in period when applying Theorem 2.

Moreover, if $T_0$ is chosen to be sufficiently large, the resulting variance estimation problem becomes more stable. So long as the evaluation policy is chosen such that $\Pi_e(\cdot \mid c) \ll \Pi_t(\cdot \mid c)$, then $\hat{\sigma}_t^2$ will almost surely be bounded from below by a positive constant for all $t > T_0$.

**Proposition 1.** *Assume the conditions of Theorem 4, except for the requirements on the centering sequence $\hat{\Psi}_t$. Let $T_0 = \lfloor aT \rfloor$ and $T_1 = \lfloor bT \rfloor$, where $0 < a < b < 1$, and set $n = T_1 - T_0$. Suppose the $\mathcal{F}_{T_0}$-measurable pilot satisfies $\|\hat{\theta}_{T_0} - \theta_{P_e}\|_\Theta = o_p(T^{-1/4})$. Define*

$$\hat{\Psi} := \frac{1}{T_1 - T_0} \sum_{s=T_0+1}^{T_1} w_s m_{\hat{\theta}_{T_0}}(Z_s).$$

*Then*

$$(\hat{\Psi} - \Psi_e)^2 = o_p(T^{-1/2}).$$

*Consequently, setting $\hat{\Psi}_t = \hat{\Psi}$ for $t > T_1$ gives a predictable centering sequence on these rounds satisfying*

$$\frac{1}{T} \sum_{t=T_1+1}^{T} (\hat{\Psi}_t - \Psi_e)^2 = o_p(T^{-1/2}), \qquad \sup_{T_1 < t \le T} |\hat{\Psi}_t| = O_p(1).$$

*Proof.* Decompose

$$\hat{\Psi} - \Psi_e = \left(\hat{\Psi} - \mathbb{E}_{P_e}[m_{\hat{\theta}_{T_0}}]\right) + \left(\mathbb{E}_{P_e}[m_{\hat{\theta}_{T_0}}] - \mathbb{E}_{P_e}[m_{\theta_{P_e}}]\right).$$

By pilot consistency, the event

$$\mathcal{A}_T := \{\|\hat{\theta}_{T_0} - \theta_{P_e}\|_\Theta \le \epsilon_\Theta\} \in \mathcal{F}_{T_0}$$

has probability tending to one. We work on this event, where Assumption 4 applies.

Write

$$X_s := w_s m_{\hat{\theta}_{T_0}}(Z_s), \qquad \mu_0 := \mathbb{E}_{P_e}[m_{\hat{\theta}_{T_0}}].$$

Because the pilot is fixed before the holdout block, change of measure gives

$$\mathbb{E}[X_s \mid \mathcal{F}_{s-1}] = \mu_0, \qquad T_0 < s \le T_1.$$

Thus $\{X_s - \mu_0\}_{s=T_0+1}^{T_1}$ is a martingale difference sequence. Its distinct increments are uncorrelated conditionally on $\mathcal{F}_{T_0}$.

The overlap condition provides a bound $0 \le w_s \le K_T$, where $K_T = O(T^{1/8})$. Using $w_s^2 \le K_T w_s$ and change of measure again,

$$\begin{aligned}
\mathbb{E}[X_s^2 \mid \mathcal{F}_{s-1}] &\le K_T \mathbb{E}[w_s m_{\hat{\theta}_{T_0}}(Z_s)^2 \mid \mathcal{F}_{s-1}] \\
&= K_T \mathbb{E}_{P_e}[m^2_{\hat{\theta}_{T_0}}] \\
&\le K_T \mathbb{E}_{P_e}[C(Z)^2].
\end{aligned}$$

Consequently, on $\mathcal{A}_T$,

$$\begin{aligned}
\mathrm{Var}(\hat{\Psi} \mid \mathcal{F}_{T_0}) &= \frac{1}{n^2} \sum_{s=T_0+1}^{T_1} \mathrm{Var}(X_s \mid \mathcal{F}_{T_0}) \\
&\le \frac{K_T}{n} \mathbb{E}_{P_e}[C(Z)^2] = O(T^{-7/8}).
\end{aligned}$$

Conditional Chebyshev's inequality and $\Pr(\mathcal{A}_T) \to 1$ therefore yield

$$\left(\hat{\Psi} - \mathbb{E}_{P_e}[m_{\hat{\theta}_{T_0}}]\right)^2 = O_p(T^{-7/8}) = o_p(T^{-1/2}).$$

For the second term, the derivative envelope gives, on $\mathcal{A}_T$,

$$|m_{\hat{\theta}_{T_0}}(z) - m_{\theta_{P_e}}(z)| \le C(z) \|\hat{\theta}_{T_0} - \theta_{P_e}\|_\Theta.$$

Hence, by Jensen's inequality and the pilot rate,

$$\left|\mathbb{E}_{P_e}[m_{\hat{\theta}_{T_0}} - m_{\theta_{P_e}}]\right|^2 \leq \mathbb{E}_{P_e}\left[(m_{\hat{\theta}_{T_0}} - m_{\theta_{P_e}})^2\right]$$
$$\leq \mathbb{E}_{P_e}[C(Z)^2]\,\|\hat{\theta}_{T_0} - \theta_{P_e}\|_\Theta^2 = o_p(T^{-1/2}).$$

Combining the two bounds proves the first conclusion.

Finally, $\hat{\Psi}$ is $\mathcal{F}_{T_1}$-measurable and $\hat{\Psi} = \Psi_e + o_p(1) = O_p(1)$. Therefore, for $t > T_1$, $\hat{\Psi}_t = \hat{\Psi}$ is predictable, uniformly bounded in probability, and

$$\frac{1}{T}\sum_{t=T_1+1}^{T}(\hat{\Psi}_t - \Psi_e)^2 = \frac{T-T_1}{T}(\hat{\Psi} - \Psi_e)^2 = o_p(T^{-1/2}).$$

□

# D Deferred Proofs

## D.1 Deferred Assumptions

Prior to presenting proofs, we briefly state, discuss and justify assumptions that are not discussed in the main body.

**Assumption 4** (Envelope condition)**.** *There exist constants $\epsilon_\Theta, \epsilon_\Upsilon > 0$ and a measurable function $C : \mathcal{Z} \to \mathbb{R}_+$ such that $\mathbb{E}_{P_e}[C(Z)^4] < \infty$, and the following conditions hold whenever $\|\theta - \theta_{P_e}\|_\Theta \leq \epsilon_\Theta$, $\|\theta' - \theta_{P_e}\|_\Theta \leq \epsilon_\Theta$, $\|\eta - \eta_{P_e}\|_\Upsilon \leq \epsilon_\Upsilon$, and $\|\eta' - \eta_{P_e}\|_\Upsilon \leq \epsilon_\Upsilon$.*

1. $|m_\theta(z)| \leq C(z)$;
2. $|\partial_\theta m_\theta(z)[h]| \leq C(z)\|h\|_\Theta$;
3. $|\partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[h]| \leq C(z)\|h\|_\Theta$
4. $|\partial_\theta^2 \ell_{\theta,\eta}(z)[h,u]| \leq C(z)\|h\|_\Theta\|u\|_\Theta$;
5. $|\partial_\eta\partial_\theta \ell_{\theta,\eta}(z)[h,v]| \leq C(z)\|h\|_\Theta\|v\|_\Upsilon$.

*Furthermore, the derivatives are locally Lipschitz in this neighborhood:*

1. $|\partial_\theta m_\theta(z)[h] - \partial_\theta m_{\theta'}(z)[h]| \leq C(z)\,\|\theta - \theta'\|_\Theta\|h\|_\Theta$;
2. $|\partial_\eta\partial_\theta \ell_{\theta,\eta}(z)[h,v] - \partial_\eta\partial_\theta \ell_{\theta',\eta'}(z)[h,v]| \leq C(z)\,(\|\theta - \theta'\|_\Theta + \|\eta - \eta'\|_\Upsilon)\,\|h\|_\Theta\|v\|_\Upsilon$;
3. $\left|\partial_\theta^2 \ell_{\theta,\eta}(z)[h,u] - \partial_\theta^2 \ell_{\theta',\eta'}(z)[h,u]\right| \leq C(z)\,(\|\theta - \theta'\|_\Theta + \|\eta - \eta'\|_\Upsilon)\,\|h\|_\Theta\|u\|_\Theta$.

Assumption 4 is used to translate consistency of the parameter estimates into asymptotic negligibility of the function-valued quantities that enter the one-step expansion. In particular, once $(\hat{\theta}, \hat{\eta}, \hat{\alpha})$ lies in a local neighborhood of $(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})$, the envelope ensures that the associated influence function increments are dominated by an integrable random variable. This allows the population Taylor remainders in Theorem 1 and the martingale remainder terms in the asymptotic normality proofs (Theorem 2, Theorem 3) to be controlled by products of estimation errors. The local Lipschitz bounds play the analogous role for differences of derivative terms: they ensure that replacing $(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})$ by consistent estimates changes the induced functionals by an amount proportional to the corresponding estimation error. In finite-dimensional models, these Lipschitz conditions are implied by bounded third derivatives; here we state them directly in operator form so that the same argument applies to general nonparametric $M$-estimands. Similar sets of restrictions are common in standard proofs of asymptotic normality for $M$-estimators such as in van der Vaart (2000).

**Lemma 1** (Completion in the Hessian norm)**.** *Under Assumptions 3 and 4, let $\bar{\Theta}$ be the completion of $\Theta$ under $\|h\|_{H_{P_e}} := \{H_{P_e}(h,h)\}^{1/2}$. The norm $\|\cdot\|_\Theta$ extends to an equivalent norm on $\bar{\Theta}$. The derivatives entering the one-step expansion extend uniquely and continuously in their $\Theta$-direction arguments,*

*preserving their envelope and local Lipschitz bounds. In particular,* $L(h) := \mathbb{E}_{P_e}[\partial_\theta m_{\theta_{P_e}}(Z)[h]]$ *has a unique Riesz representer* $\alpha_{P_e} \in \bar{\Theta}$*, and*

$$\alpha_{P_e} = \arg\min_{\alpha \in \bar{\Theta}} \left\{ \tfrac{1}{2} H_{P_e}(\alpha, \alpha) - L(\alpha) \right\}.$$

*Proof.* Let $M := \mathbb{E}_{P_e}[C(Z)] < \infty$. Coercivity and the Hessian envelope give

$$c\|h\|_\Theta^2 \le \|h\|_{H_{P_e}}^2 \le M\|h\|_\Theta^2, \qquad h \in \Theta.$$

Thus both norms have the same Cauchy sequences, and $\|h\|_\Theta := \lim_n \|h_n\|_\Theta$ is well defined whenever $h_n \in \Theta$ converges to $h \in \bar{\Theta}$. The Hessian is symmetric by the stated differentiability and local Lipschitz assumptions, so its positive definite extension makes $\bar{\Theta}$ a Hilbert space. Moreover,

$$|L(h)| \le M\|h\|_\Theta \le Mc^{-1/2}\|h\|_{H_{P_e}},$$

so $L$ extends continuously and Riesz representation applies.

For the derivative evaluations, each envelope bound makes the images of approximating direction sequences Cauchy. For example, $|\partial_\theta m_\theta(z)[h_n - h_k]| \le C(z)c^{-1/2}\|h_n - h_k\|_{H_{P_e}}$. The bilinear maps extend by approximation in each $\Theta$-direction argument; their product bounds make the limits independent of the approximating sequences. The loss derivative at a nearby $(\theta, \eta)$ is also bounded: integrating the Hessian and mixed derivative along the segment from $(\theta_{P_e}, \eta_{P_e})$ gives

$$|\partial_\theta \ell_{\theta,\eta}(z)[h]| \le C(z)\big(1 + \|\theta - \theta_{P_e}\|_\Theta + \|\eta - \eta_{P_e}\|_\Upsilon\big)\|h\|_\Theta.$$

Taking limits preserves the envelope and Lipschitz inequalities. Their integrable bounds justify passing expectations to the limit, including the first-order and orthogonality identities. These extensions act only on direction arguments; $\theta$ and $\eta$ remain in their original spaces. Finally, $L(\alpha) = H_{P_e}(\alpha_{P_e}, \alpha)$ implies

$$\tfrac{1}{2}H_{P_e}(\alpha, \alpha) - L(\alpha) = \tfrac{1}{2}\|\alpha - \alpha_{P_e}\|_{H_{P_e}}^2 - \tfrac{1}{2}\|\alpha_{P_e}\|_{H_{P_e}}^2,$$

which proves the claimed unique minimizer. □

### D.2 Proof of Theorem 1

*Proof.* Recalling that $H_{P_e}(h, u) := \mathbb{E}_{P_e}\big[\partial_\theta^2 \ell_{\theta_{P_e},\eta_{P_e}}(Z)[h, u]\big]$, we construct a functional Taylor expansion at the evaluation law. Since $m_\theta$ is Fréchet differentiable, we have the expansion, we have that

$$\begin{aligned}\Psi_e(\bar{\theta}) - \Psi_e(\theta_{P_e}) &= \mathbb{E}_{P_e}[m_{\bar{\theta}} - m_{\theta_{P_e}}] \\ &= \mathbb{E}_{P_e}\left[\partial_\theta m_{\theta_{P_e}}(z)[\bar{\theta} - \theta_{P_e}]\right] + R_1,\end{aligned}$$

where $r_1(z) = \int_0^1 \left\{\partial_\theta m_{\theta_{P_e} + u(\bar{\theta} - \theta_{P_e})}(z) - \partial_\theta m_{\theta_{P_e}}(z)\right\}[\bar{\theta} - \theta_{P_e}]\, du.$ and and $R_1 = \mathbb{E}_{P_e}[r_1(Z)]$. By Assumption 4,

$$|r_1(z)| \le \int_0^1 C(z)u\|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2\, du \le C(z)\|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2.$$

Therefore

$$R_1 = \mathbb{E}_{P_e}[r_1(Z)] \le |\mathbb{E}_{P_e}[r_1(Z)]| \le \mathbb{E}_{P_e}[C(Z)]\|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2.$$

If $m_\theta$ is linear in $\theta$, then $r_1(z) = 0$ and thus $R_1 = 0$.

By the Riesz representation identity, $\mathbb{E}_{P_e}\left[\partial_\theta m_{\theta_{P_e}}(Z)[h]\right] = H_{P_e}(\alpha_{P_e}, h)$ for all $h \in \Theta$, so

$$\Psi_e(\bar{\theta}) - \Psi_e(\theta_{P_e}) = H_{P_e}(\alpha_{P_e}, \bar{\theta} - \theta_{P_e}) + R_1.$$

We now expand $\mathbb{E}_{P_e}[\partial_\theta \ell_{\theta,\eta}(Z)[\alpha]]$ around $(\theta_{P_e}, \eta_{P_e})$:

$$\partial_\theta \ell_{\bar{\theta},\bar{\eta}}(z)[\bar{\alpha}] = \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\bar{\alpha}] + \partial_\theta^2 \ell_{\theta_{P_e},\eta_{P_e}}(z)[\bar{\theta} - \theta_{P_e}, \bar{\alpha}] + \partial_\eta \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\bar{\alpha}, \bar{\eta} - \eta_{P_e}] + r_\ell(z),$$

where $\theta_u = \theta_{P_e} + u(\bar{\theta} - \theta_{P_e})$ and $\eta_u = \eta_{P_e} + u(\bar{\eta} - \eta_{P_e})$, and the remainder is given by

$$r_2(z) = \int_0^1 \left[\partial_\theta^2 \ell_{\theta_u,\eta_u}(z)[\bar{\theta} - \theta_{P_e}, \bar{\alpha}] - \partial_\theta^2 \ell_{\theta_{P_e},\eta_{P_e}}(z)[\bar{\theta} - \theta_{P_e}, \bar{\alpha}]\right] du$$

$$+\int_0^1 \left[\partial_\eta \partial_\theta \ell_{\theta_u,\eta_u}(z)[\bar{\alpha}, \bar{\eta} - \eta_{P_e}] - \partial_\eta \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\bar{\alpha}, \bar{\eta} - \eta_{P_e}]\right] du.$$

Taking expectation,

$$\mathbb{E}_{P_e}\left[\partial_\theta \ell_{\bar{\theta},\bar{\eta}}(z)[\bar{\alpha}]\right] = \mathbb{E}_{P_e}[\partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(Z)[\bar{\alpha}]] + H_{P_e}(\bar{\theta} - \theta_{P_e}, \bar{\alpha}) + \mathbb{E}_{P_e}[\partial_\eta \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(Z)[\bar{\alpha}, \bar{\eta} - \eta_{P_e}]] + R_2,$$

where $R_2 := \mathbb{E}_{P_e}[r_2(Z)]$.

Under the envelope bound of Assumption 4, there exists $C(z)$ such that

$$|\partial_\theta^2 \ell_{\theta,\eta}(z)[h,u]| \leq C(z)\|h\|_\Theta \|u\|_\Theta,$$

$$|\partial_\eta \partial_\theta \ell_{\theta,\eta}(z)[h,u]| \leq C(z)\|h\|_\Theta \|u\|_\Upsilon.$$

Noting that the integrand is constant in the case that $\delta_{\text{quad}} = 0$, we have that

$$|r_2(z)| \lesssim C(z)\|\bar{\alpha}\|_\Theta \left(\delta_{\text{quad}}\|\bar{\theta} - \theta_{P_e}\|_\Theta^2 + \|\bar{\theta} - \theta_{P_e}\|_\Theta \|\bar{\eta} - \eta_{P_e}\|_\Upsilon + \|\bar{\eta} - \eta_{P_e}\|_\Upsilon^2\right).$$

Taking expectations,

$$|R_2| \lesssim \|\bar{\alpha}\|_\Theta \left(\delta_{\text{quad}}\|\bar{\theta} - \theta_{P_e}\|_\Theta^2 + \|\bar{\theta} - \theta_{P_e}\|_\Theta \|\bar{\eta} - \eta_{P_e}\|_\Upsilon + \|\bar{\eta} - \eta_{P_e}\|_\Upsilon^2\right).$$

By the first order conditions, we have that Assumption 3

$$\mathbb{E}_{P_e}[\partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(Z)[h]] = 0, \quad \text{and} \quad \mathbb{E}_{P_e}[\partial_\eta \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(Z)[h,u]] = 0.$$

Thus,

$$\begin{aligned}\mathbb{E}_{P_e}\left[\partial_\theta \ell_{\bar{\theta},\bar{\eta}}(z)[\bar{\alpha}]\right] &= H_{P_e}(\bar{\theta} - \theta_{P_e}, \bar{\alpha}) + R_2 \\ &= H_{P_e}(\alpha_{P_e}, \bar{\theta} - \theta_{P_e}) + H_{P_e}(\bar{\alpha} - \alpha_{P_e}, \bar{\theta} - \theta_{P_e}) + R_2\end{aligned}$$

Note that under bilinearity $H_{P_e}(\alpha_{P_e}, \bar{\theta} - \theta_{P_e}) = \mathbb{E}_{P_e}\left[\partial_\theta \ell_{\bar{\theta},\bar{\eta}}(z)[\bar{\alpha}]\right] - H_{P_e}(\bar{\alpha} - \alpha_{P_e}, \bar{\theta} - \theta_{P_e}) - R_2$
Putting everything together, we have that

$$\Psi_e(\bar{\theta}) - \Psi_e(\theta_{P_e}) = \mathbb{E}_{P_e}[\partial_\theta \ell_{\bar{\theta},\bar{\eta}}(Z)[\bar{\alpha}]] - H_{P_e}(\bar{\alpha} - \alpha_{P_e}, \bar{\theta} - \theta_{P_e}) + R_1 - R_2.$$

Combining the bounds on $R_1$ and $R_2$ yields the result. □

### D.3 Proof of Theorem 2

*Proof.* Let us rewrite

$$\frac{\hat{A}_T}{\sqrt{T}}(\hat{\Psi}_T^{\text{os}} - \Psi_e(\theta_{P_e})) = \frac{1}{\sqrt{T}}\sum_{t=1}^T \hat{\sigma}_t^{-1} w_t \left(m_{\hat{\theta}_t}(Z_t) - \dot{\ell}_{\hat{\theta}_t,\hat{\eta}_t}(Z_t)[\hat{\alpha}_t] - \Psi_e(\theta_{P_e})\right)$$

Let $F(\theta, \eta, \alpha)(z) := m_\theta(z) - \dot{\ell}_{\theta,\eta}(z)[\alpha]$. Add and subtract $F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})$ inside the summand. We then have that

$$\frac{\hat{A}_T}{\sqrt{T}}(\hat{\Psi}_T^{\text{os}} - \Psi(P_e)) = \underbrace{\frac{1}{\sqrt{T}}\sum_{t=1}^T \hat{\sigma}_t^{-1} w_t \varphi_t}_{\xrightarrow{d} N(0,1)} + \underbrace{\frac{1}{\sqrt{T}}\sum_{t=1}^T \hat{\sigma}_t^{-1} w_t \left(F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t)(Z_t) - F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})(Z_t)\right)}_{\text{Remainder}}.$$

The first term converges to $N(0,1)$ immediately by applying a martingale central limit theorem (Lemma 2). We enumerate some sufficient conditions for the remainder to be asymptotically negligible in Lemma 4.

Conditions (a) of Lemma 4 is true by assumption, so all that remains is verifying conditions (b) - (e). To this end, we can apply Lemma 5. We verify the conditions of Lemma 5 for each term one-by one:

**Showing** $P_T^a\Big(\dot{m}_{\theta_{P_e}}(\cdot)[\hat{\theta}_t - \theta_{P_e}] - \ddot{\ell}_{\theta_{P_e},\eta_{P_e}}(\cdot)[\hat{\theta}_t - \theta_{P_e}, \alpha_{P_e}]\Big) = o_p(T^{-1/2})$ We note that because $\hat{\theta}_t$ is $\mathcal{F}_{t-1}$ measurable, it is a constant conditional on $\mathcal{F}_{t-1}$. Therefore, if we let $f_t := \dot{m}_{\theta_{P_e}}(\cdot)[\hat{\theta}_t - \theta_{P_e}] - \ddot{\ell}_{\theta_{P_e},\eta_{P_e}}(\cdot)[\hat{\theta}_t - \theta_{P_e}, \alpha_{P_e}]$, we have:

$$\begin{aligned}
\mathbb{E}\left[w_t f_t(Z_t) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[w_t\left(\dot{m}_{\theta_{P_e}}(Z_t)[\hat{\theta}_t - \theta_{P_e}] - \ddot{\ell}_{\theta_{P_e},\eta_{P_e}}(Z_t)[\hat{\theta}_t - \theta_{P_e}, \alpha_{P_e}]\right) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}_{p_e}\left[\dot{m}_{\theta_{P_e}}(Z)[\hat{\theta}_t - \theta_{P_e}] - \ddot{\ell}_{\theta_{P_e},\eta_{P_e}}(Z)[\hat{\theta}_t - \theta_{P_e}, \alpha_{P_e}]\right] \\
&= 0,
\end{aligned}$$

where for the last step we use the the properties of the Riesz representer $\alpha_{P_e}$ . Condition (b) is true by assumption. For condition c, we have that $\sup_{(c,a)} w_t < K_T$ for some $K_T = O(T^{1/3})$. Therefore,

$$\begin{aligned}
\frac{1}{T^2}\sum_{t=1}^{T}\left\{\mathbb{E}[w_t^2 C(Z_t)^2 \mid \mathcal{F}_{t-1}]\right\}^2 &\leq \frac{1}{T^2}\sum_{t=1}^{T}\left\{K_T\mathbb{E}[w_t C(Z_t)^2 \mid \mathcal{F}_{t-1}]\right\}^2 \\
&= \frac{K_T^2}{T}(\mathbb{E}_{P_e}[C^2])^2 = O(T^{-1/3}) = O(1).
\end{aligned}$$

**Showing** $P_T^a\Big(\partial_\eta\partial_\theta\ell_{\theta_{P_e},\eta_{P_e}}(\cdot)[\alpha_{P_e}, \hat{\eta}_t - \eta_{P_e}]\Big) = o_p(T^{-1/2})$ The arguments are the same as for the previous term, with the only step differing in showing $\mathbb{E}[w_t f_t(Z_t) \mid \mathcal{F}_{t-1}] = 0$ for each $t \in [T]$. Here, we have:

$$\begin{aligned}
\mathbb{E}\left[w_t f_t(Z_t) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[w_t\partial_\eta\partial_\theta\ell_{\theta_{P_e},\eta_{P_e}}(Z_t)[\alpha_{P_e}, \hat{\eta}_t - \eta_{P_e}] \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}_{p_e}\left[\partial_\eta\partial_\theta\ell_{\theta_{P_e},\eta_{P_e}}(Z)[\alpha_{P_e}, \hat{\eta}_t - \eta_{P_e}]\right] \\
&= 0,
\end{aligned}$$

with the last line following from the Neyman orthogonality condition in Assumption 3 that $\mathbb{E}_{P_e}\left[\partial_\eta\partial_\theta\ell_{\theta_{P_e},\eta_{P_e}}(Z)[h, u]\right] = 0$ for all $h \in \Theta$ and $u \in \Upsilon$.

**Showing** $P_T^a\Big(\partial_\theta\ell_{\theta_{P_e},\eta_{P_e}}(\cdot)[\hat{\alpha}_t - \alpha_{P_e}]\Big) = o_p(T^{-1/2})$ The arguments are the same as for the previous term, with the only step differing in showing $\mathbb{E}[w_t f_t(Z_t) \mid \mathcal{F}_{t-1}] = 0$ for each $t \in [T]$. Here, we have:

$$\begin{aligned}
\mathbb{E}\left[w_t f_t(Z_t) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[w_t\partial_\theta\ell_{\theta_{P_e},\eta_{P_e}}(Z_t)[\hat{\alpha}_t - \alpha_{P_e}] \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}_{p_e}\left[\partial_\theta\ell_{\theta_{P_e},\eta_{P_e}}(Z_t)[\hat{\alpha}_t - \alpha_{P_e}]\right] \\
&= 0,
\end{aligned}$$

with the last line following from the fact that the first order condition of Assumption 3 states that $\mathbb{E}_{P_e}\left[\partial\theta\ell_{\theta_{P_e},\eta_{P_e}}(Z)[h]\right] = 0$ for all $h \in \Theta$. Finally, all that remains is showing condition (e) of Lemma 4 is satisfied.

**Showing** $\frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\left[w_t^4 C(Z_t)^4 \mid \mathcal{F}_{t-1}\right] = O_p(1)$**.** This is almost the same as in the prior step. All we need to show is:

$$\begin{aligned}
\frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}[w_t^4 C(Z_t)^4 \mid \mathcal{F}_{t-1}] &\leq \frac{K_T^3}{T^2}\sum_{t=1}^{T}\mathbb{E}[w_t C(Z_t)^4 \mid \mathcal{F}_{t-1}] \\
&= \frac{K_T^3}{T}\mathbb{E}_{P_e}[C^4] = O(1).
\end{aligned}$$

Thus, the remainder is asymptotically negligible and asymptotic normality holds. □

### D.4 Proof of Theorem 3

*Proof.* Define $F(\theta,\eta,\alpha)(z) := m_\theta(z) - \partial_\theta \ell_{\theta,\eta}(z)[\alpha]$, $F_0(z) := F(\theta_{P_e},\eta_{P_e},\alpha_{P_e})(z)$ and set $\Psi_e := \Psi_e(\theta_{P_e})$, and $D_t := w_t\big(F_0(Z_t) - \Psi_e\big)$ Define the plug-in quantities

$$\hat{D}_t := w_t\big(F(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(Z_t) - \bar{\Psi}_T\big), \quad \text{for } t > 2m_T,$$

and set $\hat{D}_t := 0$ for $t \le 2m_T$. Then we can write

$$\begin{aligned}\hat{V}_T^{-1/2}(\hat{S}_T - \Psi_e) &= \frac{A_T^{-1}\sum_{t=1}^T\{w_t F(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(Z_t) - w_t\Psi_e\}}{A_T^{-1}(\sum_{t=2m_T+1}^T \hat{D}_t^2)^{1/2}} \\ &= \sqrt{T}\,\frac{T^{-1}\sum_{t=1}^T\{w_t F(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(Z_t) - w_t\Psi_e\}}{(T^{-1}\sum_{t=2m_T+1}^T \hat{D}_t^2)^{1/2}}.\end{aligned}$$

Define

$$\tilde{S}_T = \frac{1}{T}\sum_{t=1}^T\{w_t F(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(Z_t) - w_t\Psi_e(\theta_{P_e})\}, \qquad \widetilde{V}_T = \frac{1}{T}\sum_{t=2m_T+1}^T \hat{D}_t^2.$$

It now suffices to verify the conditions of Lemma 3 for $\{D_t, \hat{D}_t, \tilde{S}_t\}$ in order to demonstrate asymptotic normality.

By construction, $\{D_t\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{F}_{t-1}\}_{t=1}^T$. Assumptions (a) and (c) are true by condition (3c) of Theorem 3 and condition (2b) of Theorem 2 respectively. We now verify the remaining pieces of the lemma, namely

1. (Realized Variance) $(TV_T)^{-1}\sum_{t=1}^T D_t^2 \xrightarrow{p} 1$;
2. (Plug-in Rate 1) $\tilde{S}_T - S_T = o_p(T^{-1/2})$
3. (Plug-in Rate 2) $\frac{1}{T}\sum_{t=1}^T(\hat{D}_t - D_t)^2 = o_p(1)$.

**Realized Variance** We wish to show $(TV_T)^{-1}\sum_{t=1}^T D_t^2 \xrightarrow{p} 1$. Writing this out,

$$\frac{1}{T}\sum_{t=1}^T D_t^2 - V_T = \frac{1}{T}\sum_{t=1}^T\left\{D_t^2 - E[D_t^2 \mid \mathcal{F}_{t-1}]\right\}.$$

We have that

$$\mathbb{E}[\{D_t^2 - \mathbb{E}[D_t^2 \mid \mathcal{F}_{t-1}]\}^2 \mid \mathcal{F}_{t-1}] \le E[D_t^4 \mid \mathcal{F}_{t-1}].$$

By Lemma 8, the conditional fourth moment is $O_p(K_T^3)$, so this untruncated variance bound is $O_p(K_T^3/T)$ and need not vanish at the allowed endpoint. We retain the centered-square argument but truncate its increments using the assumed Lindeberg condition. For $\epsilon > 0$ define

$$U_{T,t} = D_t^2\mathbf{1}\{|D_t| \le \epsilon\sqrt{T}\}, \qquad L_T(\epsilon) = \frac{1}{T}\sum_t \mathbb{E}[D_t^2\mathbf{1}\{|D_t| > \epsilon\sqrt{T}\} \mid \mathcal{F}_{t-1}].$$

Then $L_T(\epsilon) \to_p 0$. For a nonnegative adapted sum $X_T$ with predictable sum $A_T$, stopping before the next predictable increment would make $A_T$ exceed $b$ and applying Markov's inequality gives

$$\Pr(X_T > \eta) \le b/\eta + \Pr(A_T > b).$$

Apply this to $X_T = T^{-1}\sum_t D_t^2\mathbf{1}\{|D_t| > \epsilon\sqrt{T}\}$ and $A_T = L_T(\epsilon)$: the realized tail also tends to zero in probability. For the retained increments,

$$\begin{aligned}\frac{1}{T^2}\sum_t \mathbb{E}[\{U_{T,t} - \mathbb{E}[U_{T,t} \mid \mathcal{F}_{t-1}]\}^2 \mid \mathcal{F}_{t-1}] &\le \frac{1}{T^2}\sum_t \mathbb{E}[D_t^4\mathbf{1}\{|D_t| \le \epsilon\sqrt{T}\} \mid \mathcal{F}_{t-1}] \\ &\le \epsilon^2\frac{1}{T}\sum_t \mathbb{E}[D_t^2 \mid \mathcal{F}_{t-1}] = \epsilon^2 V_T.\end{aligned}$$

The stopped martingale second-moment bound consequently gives

$$\Pr\left(\left|T^{-1}\sum_t\{U_{T,t}-\mathbb{E}[U_{T,t}\mid\mathcal{F}_{t-1}]\}\right|>\eta\right)\le\epsilon^2L/\eta^2+\Pr(V_T>L).$$

First choose $L$ large using $V_T\to_p\kappa<\infty$, then take $T\to\infty$ and $\epsilon\downarrow 0$. Combining retained increments, their predictable tails, and their realized tails proves the centered-square average is $o_p(1)$. The preceding truncated second-moment argument yields

$$\frac{1}{T}\sum_{t=1}^{T}\left\{D_t^2-E[D_t^2\mid\mathcal{F}_{t-1}]\right\}=o_p(1).$$

Therefore, $\frac{1}{T}\sum_{t=1}^{T}D_t^2=V_T+o_p(1)$. Since $V_T\xrightarrow{p}\kappa$ with $\kappa>0$ almost surely, it follows that $\frac{1}{TV_T}\sum_{t=1}^{T}D_t^2=1+o_p(1)$.

**Showing $\tilde{S}_T-S_T=o_p(T^{-1/2})$** Note that $S_T=T^{-1}\sum_{t=1}^{T}\{w_tF_0(Z_t)-w_t\Psi_e\}$. Therefore,

$$\tilde{S}_T-S_T=\frac{1}{T}\sum_{t=1}^{T}w_t\{F(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(Z_t)-F_0(Z_t)\}.$$

Demonstrating this quantity is $o_p(T^{-1/2})$ is precisely the same argument as showing the remainder in Theorem 2 is $o_p(T^{-1/2})$. The only difference is that in this case we let $\hat{\sigma}_t=1$ for all $t\in[T]$ which is allowed because the only requirement is that the weights are almost surely bounded. Repeating these arguments with this choice of $\hat{\sigma}_t$ yields the result.

**Showing $\frac{1}{T}\sum_{t=1}^{T}(\hat{D}_t-D_t)^2=o_p(1)$** We first show that $\bar{\Psi}_T-\Psi_e=o_p(1)$. Note that

$$\begin{aligned}\bar{\Psi}_T-\Psi_e&=\frac{1}{m_T}\sum_{t=m_T+1}^{2m_T}w_tF(\bar{\theta},\bar{\eta},\bar{\alpha})(Z_t)-\Psi_e\\&=\underbrace{\left(\frac{1}{m_T}\sum_{t=m_T+1}^{2m_T}w_tF(\bar{\theta},\bar{\eta},\bar{\alpha})(Z_t)-\mathbb{E}_{P_e}\left[F(\bar{\theta},\bar{\eta},\bar{\alpha})(Z)\right]\right)}_{R_1}+\underbrace{\left(\mathbb{E}_{P_e}\left[F(\bar{\theta},\bar{\eta},\bar{\alpha})(Z)\right]-\mathbb{E}_{P_e}\left[F_0(Z)\right]\right)}_{R_2}\end{aligned}$$

To bound $R_1$, note that $\mathbb{E}\left[w_tF(\bar{\theta},\bar{\eta},\bar{\alpha})(Z_t)\mid\mathcal{F}_{t-1}\right]=\mathbb{E}_{P_e}\left[F(\bar{\theta},\bar{\eta},\bar{\alpha})(Z)\right]$ so the centered summands are martingale differences. On the predictable localization event of Lemma 8, $|F(\bar{\theta},\bar{\eta},\bar{\alpha})|\le LC$. Thus $\mathbb{E}[(w_t\bar{F}-\mathbb{E}_{P_e}[\bar{F}])^2\mid\mathcal{F}_{t-1}]\le K_T\mathbb{E}_{P_e}[\bar{F}^2]\le K_TL^2\mathbb{E}_{P_e}[C^2]$. Summing over the $m_T$ centering observations and dividing by $m_T^2$ gives $R_1=O_p(\sqrt{K_T/m_T})$. There is no shrinking-function factor here to justify a little-o root-$m_T$ bound.

To bound $R_2$, we can apply Theorem 1 because this term is precisely

$$\begin{aligned}\mathbb{E}_{P_e}\left[F(\bar{\theta},\bar{\eta},\bar{\alpha})(Z)\right]-\mathbb{E}_{P_e}\left[F_0(Z)\right]&=\Psi_e(\bar{\theta})-\Psi_e(\theta_{P_e})-\mathbb{E}_{P_e}\left[\partial_\theta\ell_{\bar{\theta},\bar{\eta}}[\bar{\alpha}]\right]\\&=-H_{P_e}[\bar{\alpha}-\alpha_{P_e},\,\bar{\theta}-\theta_{P_e}]+O\left((1+\|\bar{\alpha}\|_\Theta)\|\bar{\theta}-\theta_{P_e}\|_\Theta^2\right)\\&\quad+O\left(\|\bar{\alpha}\|_\Theta\left(\|\bar{\theta}-\theta_{P_e}\|_\Theta\|\bar{\eta}-\eta_{P_e}\|_\Upsilon+\|\bar{\eta}-\eta_{P_e}\|_\Upsilon^2\right)\right).\end{aligned}$$

The requirements that $\|\bar{\theta}-\theta_{P_e}\|_\Theta=o_p(m_T^{-1/4})$,$\|\bar{\eta}-\eta_{P_e}\|_\Upsilon=o_p(m_T^{-1/4})$, $\|\bar{\alpha}-\alpha_{P_e}\|_\Theta=o_p(m_T^{-1/4})$ is sufficient when combined with the results of Lemma 6 which shows that $\|\bar{\alpha}\|_\Theta$ is bounded in probability to conclude that $R_2=o_p(m_T^{-1/2})$. Together, this shows $\bar{\Psi}_T-\Psi_e=O_p(\sqrt{K_T/m_T})+o_p(m_T^{-1/2})$, and hence $K_T(\bar{\Psi}_T-\Psi_e)^2=O_p(K_T^2/m_T)+o_p(K_T/m_T)=o_p(1)$.

We now move to showing the result. First, note that $\frac{1}{T}\sum_{t=1}^{2m_T}D_t^2=o_p(1)$, because, for the stated fixed martingale sequence, condition (3c) also gives $V_{2m_T}\to_p\kappa$. Therefore $T^{-1}\sum_{t\le 2m_T}\sigma_t^2=(2m_T/T)V_{2m_T}=o_p(1)$, since $m_T=o(T)$. The nonnegative-process bound established above gives the realized conclusion. This derives the needed negligibility from the existing assumptions; no

additional discarded-block condition is required. For a horizon-dependent array, terminal variance convergence alone would not justify substituting $2m_T$ for $T$ in this way.

For $t > 2m_T$,

$$\hat{D}_t - D_t = w_t\{F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t)(Z_t) - F_0(Z_t)\} + w_t(\Psi_e - \bar{\Psi}_T).$$

Therefore,

$$(\hat{D}_t - D_t)^2 \lesssim w_t^2\{F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t) - F_0\}^2(Z_t) + w_t^2(\bar{\Psi}_T - \Psi_e)^2.$$

Denote $\sup_{t,c,a} \pi_e/\pi_t \le K_T$. Since $\mathbb{E}[w_t^2 \mid \mathcal{F}_{t-1}] \le K_T$, Markov's inequality gives $T^{-1}\sum_t w_t^2 = O_p(K_T)$. We can therefore write.

$$\begin{aligned}\frac{1}{T}\sum_{t=2m_T+1}^{T}(\hat{D}_t - D_t)^2 &\lesssim \frac{1}{T}\sum_{t=2m_T+1}^{T} w_t^2\{F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t) - F_0\}^2(Z_t) + (\bar{\Psi}_T - \Psi_e)^2\frac{1}{T}\sum_{t=2m_T+1}^{T} w_t^2 \\ &\lesssim \left(\frac{K_T}{T}\sum_{t=2m_T+1}^{T} w_t\{F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t) - F_0\}^2(Z_t)\right) + O_p(K_T)(\bar{\Psi}_T - \Psi_e)^2.\end{aligned}$$

To control the first term, use Lemma 9, whose retained Taylor calculation gives

$$\frac{1}{T}\sum_t \mathbb{E}_{P_e}\left[\{F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t) - F_0\}^2\right] = o_p(T^{-1/2}).$$

The summands are predictable as functions of $Z$, so the nonnegative-process bound applied after multiplying by $\sqrt{T}$ yields

$$\frac{1}{T}\sum_{t>2m_T} w_t\{F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t) - F_0\}^2(Z_t) = o_p(T^{-1/2}).$$

Combining this with the centering calculation gives

$$\frac{1}{T}\sum_{t>2m_T}(\hat{D}_t - D_t)^2 = o_p(K_T T^{-1/2}) + O_p(K_T^2/m_T) + o_p(K_T/m_T).$$

Thus, $\frac{1}{T}\sum_{t=1}^{T}(\hat{D}_t - D_t)^2 = o_p(1)$ so long as $K_T = o(m_T^{1/2})$.

□

### D.5 Proof of Theorem 4

*Proof.* First, for each fixed candidate $\pi$, define the predictable random function

$$G_s(\pi) := \mathbb{E}\left[\frac{\pi_e^2(A_s \mid C_s)}{\pi(A_s \mid C_s)\pi_s(A_s \mid C_s)}\hat{\varphi}_s^2(Z_s)\,\middle|\,\mathcal{F}_{s-1}\right] = \mathbb{E}_{P_e}\left[\frac{\pi_e}{\pi}\hat{\varphi}_s^2\right].$$

Set $\bar{\sigma}_t^2 = (t-1)^{-1}\sum_{s<t} G_s(\pi_t)$ for $t \ge 2$. In general $G_s(\pi_t)$ is not the conditional expectation obtained by leaving the future random $\pi_t$ inside the conditioning on $\mathcal{F}_{s-1}$. The initialized first-round squared error is $O_p(K_T^2/T) = o_p(1)$, since $\sigma_1^2 \le K_T\mathbb{E}_{P_e}[\varphi_{P_e}^2]$; below all historical averages start at $t = 2$.

Consider the following decomposition

$$\hat{\sigma}_t^2 - \sigma_t^2 = \left(\hat{\sigma}_t^2 - \mathbb{E}\left[\hat{\varphi}_t^2 w_t^2 | \mathcal{F}_{t-1}\right]\right) + \left(\mathbb{E}\left[w_t^2\left(\hat{\varphi}_t^2 - \varphi_t^2\right)|\mathcal{F}_{t-1}\right]\right).$$

Define $w_{t,s} = \frac{\pi_e(A_s|C_s)}{\pi_t(A_s|C_s)}$to be the transported importance weights defined at time $t$ transported to observation $Z_s$. Note that $\frac{\pi_e(A_s|C_s)^2}{\pi_t(A_s|C_s)\pi_s(A_s|C_s)} = w_{t,s}w_s$. Summing over $t \in [T]$ and applying the Cauchy–Schwarz inequality twice yields

$$\begin{aligned}
\frac{1}{T}\sum_{t=2}^{T}(\hat{\sigma}_t^2-\sigma_t^2)^2 &= \frac{1}{T}\sum_{t=2}^{T}\left\{(\hat{\sigma}_t^2-\mathbb{E}[w_t^2\hat{\varphi}_t^2\mid\mathcal{F}_{t-1}])+\mathbb{E}[w_t^2(\hat{\varphi}_t^2-\varphi_t^2)\mid\mathcal{F}_{t-1}]\right\}^2\\
&\le \frac{2}{T}\sum_{t=2}^{T}(\hat{\sigma}_t^2-\mathbb{E}[w_t^2\hat{\varphi}_t^2\mid\mathcal{F}_{t-1}])^2+\frac{2}{T}\sum_{t=2}^{T}\{\mathbb{E}[w_t^2(\hat{\varphi}_t^2-\varphi_t^2)\mid\mathcal{F}_{t-1}]\}^2\\
&\le \underbrace{\frac{4}{T}\sum_{t=2}^{T}\left\{\frac{1}{t-1}\sum_{s<t}[\hat{\varphi}_s^2(Z_s)w_sw_{t,s}-G_s(\pi_t)]\right\}^2}_{R_1}\\
&\quad+\underbrace{\frac{4}{T}\sum_{t=2}^{T}(\bar{\sigma}_t^2-\mathbb{E}[w_t^2\hat{\varphi}_t^2\mid\mathcal{F}_{t-1}])^2}_{R_2}+\underbrace{\frac{2}{T}\sum_{t=2}^{T}\{\mathbb{E}[w_t^2(\hat{\varphi}_t^2-\varphi_t^2)\mid\mathcal{F}_{t-1}]\}^2}_{R_3}.
\end{aligned}$$

$R_1$ is an empirical process term whose summands we can bound uniformly over time by an appeal to the maximal inequality in Fact 2, with a log-policy entropy argument as in Girard et al. (2026). $R_2$ is an approximation error drift term. $R_3$ can be bound with roughly the same asymptotic arguments as described in Theorem 2. We now turn to each piece separately.

**Showing $R_1 = o_p(1)$** Choose a deterministic $t_0 = o(T)$ with $t_0 \to \infty$. Let $R_1 = 4T^{-1}\sum_{t=2}^{T} A_t$, where

$$A_t = \left\{\frac{1}{t-1}\sum_{s<t}\left[\hat{\varphi}_s^2(Z_s)w_sw_{t,s}-G_s(\pi_t)\right]\right\}^2.$$

We use the envelope established in Lemma 8. Its bounds are in probability, so before applying unconditional moment inequalities we stop predictably at the first round whose fitted norms or absolute centering value exceed a fixed bound $L$, or whose nuisance estimates leave the stated neighborhoods. Set subsequent fitted scores to zero. The stopping indicator at round $s$ is $\mathcal{F}_{s-1}$-measurable. For every $\eta > 0$, $L$ can be chosen so that the probability of altering any score has limit superior at most $\eta$. For this stopped process, after enlarging the constant $L$ if necessary,

$$|\hat{\varphi}_s(z)| \le LC(z), \qquad C \ge 1, \qquad \mathbb{E}_{P_e}[C^4] < \infty$$

holds for every $s, z$. Constants below may depend on fixed $L$, but not on $T$. We first prove the bounds for this process and then remove the stopping.

We also choose a truncation threshold $M_T$ and decompose

$$\hat{\varphi}_s^2 w_s w_{t,s} = \hat{\varphi}_s^2 w_s w_{t,s}\mathbf{1}\{w_s\hat{\varphi}_s^2(Z_s) \le M_T\} + \hat{\varphi}_s^2 w_s w_{t,s}\mathbf{1}\{w_s\hat{\varphi}_s^2(Z_s) > M_T\}.$$

For a fixed candidate policy $\pi$, write $w_{\pi,s} = \pi_e(A_s \mid C_s)/\pi(A_s \mid C_s)$ and define

$$\begin{aligned}
C_t^{(M)}(\pi) :=& \frac{1}{t-1}\sum_{s<t}\Big(\hat{\varphi}_s^2(Z_s)w_sw_{\pi,s}\mathbf{1}\{w_s\hat{\varphi}_s^2(Z_s)\le M_T\}\\
&\qquad -\mathbb{E}[\hat{\varphi}_s^2(Z_s)w_sw_{\pi,s}\mathbf{1}\{w_s\hat{\varphi}_s^2(Z_s)\le M_T\}\mid\mathcal{F}_{s-1}]\Big),\\
C_t^{(\mathrm{tail})}(\pi) :=& \frac{1}{t-1}\sum_{s<t}\Big(\hat{\varphi}_s^2(Z_s)w_sw_{\pi,s}\mathbf{1}\{w_s\hat{\varphi}_s^2(Z_s)> M_T\}\\
&\qquad -\mathbb{E}[\hat{\varphi}_s^2(Z_s)w_sw_{\pi,s}\mathbf{1}\{w_s\hat{\varphi}_s^2(Z_s)> M_T\}\mid\mathcal{F}_{s-1}]\Big).
\end{aligned}$$

All compensators are defined for fixed $\pi$; $C_t^{(M)}(\pi_t)$ and $C_t^{(\mathrm{tail})}(\pi_t)$ denote evaluation of the resulting random maps. This avoids treating $\pi_t$ as known at time $s-1$. The two indicators are complementary in every empirical term and compensator.

Since $\pi_s \in \Pi$ and $\sup_{\pi,c,a}\pi_e(a \mid c)/\pi(a \mid c) \le K_T = O(T^{1/8})$, we have

$$0 \le \hat{\varphi}_s^2 w_s w_{\pi,s}\mathbf{1}\{w_s\hat{\varphi}_s^2(Z_s)\le M_T\} \le K_T M_T =: B_T.$$

For $d(\pi,\pi') = \|\log\pi - \log\pi'\|_\infty \le 1$,

$$\begin{aligned}|w_{\pi,s} - w_{\pi',s}| &= w_{\pi,s}|1 - \exp\{\log\pi(A_s \mid C_s) - \log\pi'(A_s \mid C_s)\}| \\ &\le eK_T d(\pi,\pi').\end{aligned}$$

Multiplication by $w_s\hat\varphi_s^2\mathbf{1}\{w_s\hat\varphi_s^2 \le M_T\} \le M_T$ gives the Lipschitz bound $eB_T d(\pi,\pi')$ needed in Fact 2. That fact, proved below with its moment bound, yields

$$\mathbb{E}\sup_{\pi\in\Pi}|C_t^{(M)}(\pi)|^2 \lesssim \frac{B_T^2}{t-1}, \qquad \Pr\left\{\sup_{\pi\in\Pi}|C_t^{(M)}(\pi)| > \frac{cB_T}{\sqrt{t-1}}(1+\sqrt{\log(1/\delta)})\right\} \le \delta.$$

Applying the probability bound with $\delta = T^{-2}$ and a union bound over $t \in \{t_0,\dots,T\}$ gives

$$\sup_{t\ge t_0}\sup_{\pi\in\Pi}|C_t^{(M)}(\pi)| = O_p\left(B_T\sqrt{\frac{\log T}{t_0}}\right).$$

Thus the late-round supremum is $o_p(1)$ when $B_T^2\log T/t_0 \to 0$. The factor $K_T$ in $B_T = K_T M_T$ must remain in this condition.

We have $A_t \le 2\{C_t^{(M)}(\pi_t)\}^2 + 2\{C_t^{(\mathrm{tail})}(\pi_t)\}^2$. Therefore,

$$\begin{aligned}R_1 &\le \frac{8}{T}\sum_{t=2}^T\{C_t^{(M)}(\pi_t)\}^2 + \frac{8}{T}\sum_{t=2}^T\{C_t^{(\mathrm{tail})}(\pi_t)\}^2 \\ &\le \frac{8}{T}\sum_{2\le t<t_0}\{C_t^{(M)}(\pi_t)\}^2 + 8\sup_{t\ge t_0,\pi\in\Pi}|C_t^{(M)}(\pi)|^2 + \frac{8}{T}\sum_{t=2}^T\{C_t^{(\mathrm{tail})}(\pi_t)\}^2.\end{aligned}$$

For the first term, the deterministic bound $|C_t^{(M)}| \le 2B_T$ is too crude at the desired overlap rate. The second-moment part of the same maximal inequality instead gives

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{2\le t<t_0}\{C_t^{(M)}(\pi_t)\}^2\right] &\le \frac{1}{T}\sum_{2\le t<t_0}\mathbb{E}\sup_{\pi\in\Pi}|C_t^{(M)}(\pi)|^2 \\ &\lesssim \frac{B_T^2}{T}\sum_{n=1}^{t_0-2}\frac{1}{n} \lesssim \frac{B_T^2\log(et_0)}{T}.\end{aligned}$$

Markov's inequality now proves that the early-round term is $o_p(1)$ when $B_T^2\log T/T \to 0$. The second term is controlled by the preceding late-round supremum. It remains to bound the third, tail term.

For the tail term, note that

$$0 \le \hat\varphi_s^2 w_s w_{\pi,s}\mathbf{1}\{w_s\hat\varphi_s^2(Z_s) > M_T\} \le K_T w_s\hat\varphi_s^2(Z_s)\mathbf{1}\{w_s\hat\varphi_s^2(Z_s) > M_T\} =: \zeta_s.$$

Because the increments are nonnegative, with $\mu_s = \mathbb{E}[\zeta_s \mid \mathcal{F}_{s-1}]$,

$$|C_t^{(\mathrm{tail})}(\pi_t)| \le \frac{1}{t-1}\sum_{s<t}\zeta_s + \frac{1}{t-1}\sum_{s<t}\mu_s.$$

Therefore,

$$\frac{1}{T}\sum_{t=2}^T\{C_t^{(\mathrm{tail})}(\pi_t)\}^2 \le \frac{2}{T}\sum_{t=2}^T\left(\frac{1}{t-1}\sum_{s<t}\zeta_s\right)^2 + \frac{2}{T}\sum_{t=2}^T\left(\frac{1}{t-1}\sum_{s<t}\mu_s\right)^2.$$

By change of measure and the fourth moment of the envelope,

$$\begin{aligned}\mu_s &\le \frac{K_T}{M_T}\mathbb{E}[w_s^2\hat\varphi_s^4(Z_s) \mid \mathcal{F}_{s-1}] \\ &\le \frac{K_T^2}{M_T}\mathbb{E}_{P_e}[\hat\varphi_s^4] \le \frac{L^4K_T^2}{M_T}\mathbb{E}_{P_e}[C^4] =: b_T.\end{aligned}$$

Consequently the predictable part is bounded by $2b_T^2 = O(K_T^4/M_T^2)$.

It remains to control the empirical average. Expanding the square,

$$\frac{1}{T}\sum_{t=2}^{T}\left(\frac{1}{t-1}\sum_{s<t}\zeta_s\right)^2 = \frac{1}{T}\sum_{t=2}^{T}\frac{1}{(t-1)^2}\sum_{s,r<t}\zeta_s\zeta_r.$$

We first consider the off-diagonal terms. For $s < r$, $\zeta_s$ is $\mathcal{F}_{r-1}$-measurable, so

$$\mathbb{E}[\zeta_s\zeta_r] = \mathbb{E}[\zeta_s\mu_r] \le b_T\mathbb{E}\zeta_s \le b_T^2.$$

It follows that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=2}^{T}\frac{1}{(t-1)^2}\sum_{\substack{s,r<t\\ s\neq r}}\zeta_s\zeta_r\right] &= \frac{2}{T}\sum_{t=2}^{T}\frac{1}{(t-1)^2}\sum_{s<r<t}\mathbb{E}[\zeta_s\mu_r]\\ &\le \frac{b_T^2}{T}\sum_{t=2}^{T}\frac{(t-1)(t-2)}{(t-1)^2}\\ &\le b_T^2 = O(K_T^4/M_T^2).\end{aligned}$$

For the diagonal terms, all overlap factors must again be retained:

$$\zeta_s^2 \le L^4K_T^2w_s^2C(Z_s)^4 \le L^4K_T^3w_sC(Z_s)^4.$$

Thus $\mathbb{E}\zeta_s^2 \le L^4K_T^3\mathbb{E}_{P_e}[C^4]$, and

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=2}^{T}\frac{1}{(t-1)^2}\sum_{s<t}\zeta_s^2\right] &\le \frac{L^4K_T^3\mathbb{E}_{P_e}[C^4]}{T}\sum_{t=2}^{T}\frac{1}{t-1}\\ &= O\left(\frac{K_T^3\log T}{T}\right).\end{aligned}$$

Combining the diagonal and off-diagonal parts and applying Markov's inequality gives

$$\frac{1}{T}\sum_{t=2}^{T}\left(\frac{1}{t-1}\sum_{s<t}\zeta_s\right)^2 = O_p\left(\frac{K_T^4}{M_T^2} + \frac{K_T^3\log T}{T}\right).$$

The same rate, with a different constant, bounds the average squared tail process.

Collecting the rate requirements up to this point, we need

$$\begin{gathered}M_T\to\infty, \quad t_0\to\infty, \quad t_0/T\to 0,\\ \frac{K_T^2M_T^2\log T}{T}\to 0, \quad \frac{K_T^2M_T^2\log T}{t_0}\to 0, \quad \frac{K_T^4}{M_T^2}\to 0, \quad \frac{K_T^3\log T}{T}\to 0.\end{gathered}$$

For $K_T = O(T^{1/8})$, choose $M_T = T^{5/16}$ and $t_0 = \lceil T^{15/16}\rceil$. In the same order, the four nontrivial bounds are

$$O(T^{-1/8}\log T), \qquad O(T^{-1/16}\log T), \qquad O(T^{-1/8}), \qquad O(T^{-5/8}\log T),$$

all tending to zero. Hence $R_1 = o_p(1)$ for the stopped process. The original process agrees with it except on an event of asymptotic probability at most $\eta$; letting $\eta\downarrow 0$ removes the stopping and proves the claim.

**Showing $R_2 = o_p(1)$** Note that by a change of measure argument, we can rewrite

$$\begin{aligned}
\bar{\sigma}_t^2 &= \frac{1}{t-1}\sum_{s<t} G_s(\pi_t) \\
&= \frac{1}{t-1}\sum_{s=1}^{t-1}\int_{\mathcal{C}}\int_{\mathcal{A}}\int_{\mathcal{Y}} \frac{\pi_e^2(a \mid c)}{\pi_t(a \mid c)\pi_s(a \mid c)}\hat{\varphi}_s^2(c,a,y)\, p_Y(dy \mid c,a)\, \pi_s(a \mid c)\, \mu(da \mid c)\, P_C(dc) \\
&= \frac{1}{t-1}\sum_{s=1}^{t-1}\int_{\mathcal{C}}\int_{\mathcal{A}}\int_{\mathcal{Y}} \frac{\pi_e^2(a \mid c)}{\pi_t(a \mid c)}\hat{\varphi}_s^2(c,a,y)\, p_Y(dy \mid c,a)\, \mu(da \mid c)\, P_C(dc) \\
&= \frac{1}{t-1}\sum_{s=1}^{t-1}\int_{\mathcal{C}}\int_{\mathcal{A}}\int_{\mathcal{Y}} \left(\frac{\pi_e(a \mid c)}{\pi_t(a \mid c)}\right)^2 \hat{\varphi}_s^2(c,a,y)\, p_Y(dy \mid c,a)\, \pi_t(a \mid c)\, \mu(da \mid c)\, P_C(dc) \\
&= \frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{E}\left[w_t^2\hat{\varphi}_s^2 \,\middle|\, \mathcal{F}_{t-1}\right].
\end{aligned}$$

In the last conditional expectation, the historical fitted function $\hat{\varphi}_s$ is evaluated at the current observation $Z_t$; its fitted coefficients are already $\mathcal{F}_{t-1}$-measurable. Therefore, we have that

$$\begin{aligned}
\frac{1}{T}\sum_{t=2}^{T}(\bar{\sigma}_t^2 - \mathbb{E}\left[\hat{\varphi}_t^2 w_t^2 | \mathcal{F}_{t-1}\right])^2 &= \frac{1}{T}\sum_{t=2}^{T}\left(\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{E}\left[w_t^2\hat{\varphi}_s^2 \,\middle|\, \mathcal{F}_{t-1}\right] - \mathbb{E}\left[\hat{\varphi}_t^2 w_t^2 | \mathcal{F}_{t-1}\right]\right)^2 \\
&= \frac{1}{T}\sum_{t=2}^{T}\left(\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{E}\left[w_t^2(\hat{\varphi}_s^2 - \hat{\varphi}_t^2) \,\middle|\, \mathcal{F}_{t-1}\right]\right)^2 \\
&\leq \frac{1}{T}\sum_{t=2}^{T}\left(\frac{1}{t-1}\sum_{s=1}^{t-1}\|w_t\|_\infty \mathbb{E}_{P_e}[|\hat{\varphi}_s^2 - \hat{\varphi}_t^2|]\right)^2 \\
&\leq \frac{K_T^2}{T}\sum_{t=2}^{T}\left(\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{E}_{P_e}\left[|\hat{\varphi}_s^2 - \hat{\varphi}_t^2|\right]\right)^2
\end{aligned}$$

By the triangle inequality,

$$\mathbb{E}_{P_e}[|\hat{\varphi}_s^2 - \hat{\varphi}_t^2|] \leq \mathbb{E}_{P_e}\left[|\hat{\varphi}_s^2 - \varphi_{P_e}^2|\right] + \mathbb{E}_{P_e}\left[|\hat{\varphi}_t^2 - \varphi_{P_e}^2|\right].$$

We now apply Hardy's inequality (Hardy et al., 1953) which we recount here for our reader's convenience.

**Fact 1** (Hardy's Inequality). *If $\{a_n\}_{n\in\mathbb{N}}$ is a sequence of non-negative real numbers, then for every real number $p > 1$,*

$$\sum_{n=1}^{\infty}\left(\frac{a_1 + a_2 + \cdots + a_n}{n}\right)^p \leq \left(\frac{p}{p-1}\right)^p \sum_{n=1}^{\infty} a_n^p, \quad \textit{for } p > 1.$$

Fact 1 implies that $\frac{1}{T}\sum_{t=2}^{T}\left(\frac{1}{t-1}\sum_{s=1}^{t-1} a_s\right)^2 \leq \frac{4}{T}\sum_{t=1}^{T} a_t^2$. Therefore, applying Cauchy–Schwarz once more

$$
\begin{aligned}
&\frac{K_T^2}{T}\sum_{t=2}^{T}\left\{\frac{1}{t-1}\sum_{s<t}\mathbb{E}_{P_e}[|\hat{\varphi}_s^2-\hat{\varphi}_t^2|]\right\}^2\\
&\le \frac{K_T^2}{T}\sum_{t=2}^{T}\left\{\mathbb{E}_{P_e}[|\hat{\varphi}_t^2-\varphi_{P_e}^2|]+\frac{1}{t-1}\sum_{s<t}\mathbb{E}_{P_e}[|\hat{\varphi}_s^2-\varphi_{P_e}^2|]\right\}^2\\
&\le \frac{2K_T^2}{T}\sum_{t=2}^{T}\{\mathbb{E}_{P_e}[|\hat{\varphi}_t^2-\varphi_{P_e}^2|]\}^2\\
&\quad+\frac{2K_T^2}{T}\sum_{t=2}^{T}\left\{\frac{1}{t-1}\sum_{s<t}\mathbb{E}_{P_e}[|\hat{\varphi}_s^2-\varphi_{P_e}^2|]\right\}^2\\
&\le \frac{2K_T^2+8K_T^2}{T}\sum_{t=1}^{T}\{\mathbb{E}_{P_e}[|\hat{\varphi}_t^2-\varphi_{P_e}^2|]\}^2\\
&= \frac{10K_T^2}{T}\sum_{t=1}^{T}\{\mathbb{E}_{P_e}[|\hat{\varphi}_t^2-\varphi_{P_e}^2|]\}^2\\
&\le 10K_T^2\sup_{t\le T}\mathbb{E}_{P_e}[(\hat{\varphi}_t+\varphi_{P_e})^2]\cdot\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_e}[(\hat{\varphi}_t-\varphi_{P_e})^2].
\end{aligned}
$$

Since $K_T = O(T^{1/8})$ it suffices to show that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_e}\left[(\hat{\varphi}_t-\varphi_t)^2\right]=o_p(T^{-1/4})$, and $\sup_{t\le T}\mathbb{E}_{P_e}\left[(\hat{\varphi}_t+\varphi_t)^2\right]=O_p(1)$ in order to show the entire term is $o_p(1)$. However, Lemma 9 combined with the requirement that $\frac{1}{T}\sum_{t=1}^{T}(\hat{\Psi}_t-\Psi_e(\theta_{P_e}))^2=o_p(T^{-1/2})$ immediately yield the stronger claim that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_e}\left[(\hat{\varphi}_t-\varphi_t)^2\right]=o_p(T^{-1/2})$. For the second term, note that by Cauchy-Schwarz and Jensen's inequality that,

$$
\sup_{t\in[T]}\mathbb{E}_{P_e}\left[(\hat{\varphi}_t+\varphi_t)^2\right]\le\left(\sup_{t\in[T]}\mathbb{E}_{P_e}\left[(\hat{\varphi}_t+\varphi_t)^4\right]\right)^{1/2}\le\left(\sup_{t\in[T]}8\mathbb{E}_{P_e}\left[\hat{\varphi}_t^4\right]+\sup_{t\in[T]}8\mathbb{E}_{P_e}\left[\varphi_t^4\right]\right)^{1/2}.
$$

We can then invoke Lemma 8 to show $\sup_{t\in[T]}\mathbb{E}_{P_e}\left[\hat{\varphi}_t^4\right]=O_p(1)$ and $\sup_{t\in[T]}\mathbb{E}_{P_e}\left[\varphi_t^4\right]=O(1)$.

**Showing** $R_3=o_p(1)$ Using $\hat{\varphi}_t^2-\varphi_t^2=(\hat{\varphi}_t-\varphi_t)(\hat{\varphi}_t+\varphi_t)$ and Cauchy–Schwarz inequality,

$$
\begin{aligned}
\frac{2}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[w_t^2(\hat{\varphi}_t-\varphi_t)(\hat{\varphi}_t+\varphi_t)\mid\mathcal{F}_{t-1}\right]\right)^2&\le\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\left[w_t^2(\hat{\varphi}_t-\varphi_t)^2\mid\mathcal{F}_{t-1}\right]\mathbb{E}\left[w_t^2(\hat{\varphi}_t+\varphi_t)^2\mid\mathcal{F}_{t-1}\right]\\
&\le\frac{2K_T^2}{T}\sum_{t=1}^{T}\mathbb{E}_{P_e}\left[(\hat{\varphi}_t-\varphi_t)^2\right]\mathbb{E}_{P_e}\left[(\hat{\varphi}_t+\varphi_t)^2\right]
\end{aligned}
$$

The remaining argument is the same as what was shown in demonstrating $R_2=o_p(1)$

On the event that $\min_{t\in[T]}\hat{\sigma}_t^2>c$, observe that

$$
\frac{\sigma_t^2}{\hat{\sigma}_t^2}-1=\frac{\sigma_t^2-\hat{\sigma}_t^2}{\hat{\sigma}_t^2}.
$$

Hence, on the event $\min_{t\in[T]}\hat{\sigma}_t^2\ge c$,

$$
\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sigma_t^2}{\hat{\sigma}_t^2}-1\right)\right|\le\frac{1}{c}\cdot\frac{1}{T}\sum_{t=1}^{T}|\hat{\sigma}_t^2-\sigma_t^2|.
$$

Applying Cauchy–Schwarz gives

$$
\frac{1}{T}\sum_{t=1}^{T}|\hat{\sigma}_t^2-\sigma_t^2|\le\left(\frac{1}{T}\sum_{t=1}^{T}(\hat{\sigma}_t^2-\sigma_t^2)^2\right)^{1/2}.
$$

Therefore,

$$\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sigma_t^2}{\hat{\sigma}_t^2}-1\right)\right| \le \frac{1}{c}\left(\frac{1}{T}\sum_{t=1}^{T}(\hat{\sigma}_t^2-\sigma_t^2)^2\right)^{1/2} = o_p(1),$$

□

# E Auxiliary Lemmas

**Fact 2** (Theorem 6.6 of Girard et al. (2026))**.** *Let* $\Pi$ *satisfy*

$$\log N(\epsilon, \log\Pi, \|\cdot\|_\infty) \lesssim \begin{cases}\log(e/\epsilon), & p=0,\\ \epsilon^{-p}, & p\in(0,2).\end{cases}$$

*For each* $s \le t$*, let* $g_s : \mathcal{Z} \to [0,B]$ *be* $\mathcal{F}_{s-1}$*-measurable, and define*

$$f_{\pi,s}(Z_s) = \frac{g_s(Z_s)}{\pi(A_s \mid C_s)}.$$

*Then there exists a universal constant* $C>0$ *such that, for any* $t\ge 2$ *and* $\delta\in(0,1)$*, with probability at least* $1-\delta$*,*

$$\sup_{\pi\in\Pi}\left|\frac{1}{t}\sum_{s=1}^{t}\{f_{\pi,s}(Z_s)-\mathbb{E}[f_{\pi,s}(Z_s)\mid\mathcal{F}_{s-1}]\}\right| \le C\left[r_{B,t}^2+\tilde{\rho}_t(\pi)\,t^{-1/2}+\hat{\rho}_t(\pi)\sqrt{\frac{\log(et/\delta)}{t}}+B\frac{\log(et/\delta)}{t}\right],$$

*uniformly over* $\pi\in\Pi$*, where*

$$\tilde{\rho}_t(\pi) := \begin{cases}\left(\frac{1}{t}\sum_{s=1}^{t} f_{\pi,s}(Z_s)^2\right)^{\frac{1-p/2}{2}} & \text{when } p\in(0,2),\\ \left(\frac{1}{t}\sum_{s=1}^{t} f_{\pi,s}(Z_s)^2\right)^{1/2}\sqrt{\log t} & \text{when } p=0.\end{cases}$$

*and*

$$r_{B,t} \asymp \begin{cases}\left(\frac{B^{2-p}}{t}\right)^{\frac{2}{2+p}} & \text{when } p\in(0,2),\\ B\sqrt{\frac{\log t}{t}} & \text{when } p=0.\end{cases}$$

*Proof.* This is a direct application of Theorem 6.6 of Girard et al. (2026) with only minor notation changes to match our setting. □

**Lemma 2.** *Let* $\sigma_t$ *be defined as in Equation (4). Assume that the following conditions hold.*

*(a) A sequence of estimators* $\{\hat{\sigma}_t\}_{t=1}^T$ *adapted to* $\mathcal{F}_{t-1}$ *can be constructed such that there exists a constant* $c>0$ *such that* $\min_{t\in[T]}\hat{\sigma}_t^2 > c$ *almost surely and* $\frac{1}{T}\sum_{t=1}^{T}\sigma_t^2/{\hat{\sigma}_t}^2 - 1 = o_p(1)$*.*

*(b) For every* $\varepsilon>0$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[w_t^2\varphi_t^2\mathbf{1}\{|w_t\varphi_t|>\varepsilon\sqrt{T}\}\mid\mathcal{F}_{t-1}\right]\xrightarrow{p}0. \tag{10}$$

*Then under Assumptions 1-2,*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{\sigma}_t^{-1}w_t\varphi_t \xrightarrow{d} N(0,1). \tag{11}$$

*Proof.* Define $\zeta_t := \hat{\sigma}_t^{-1}w_t\varphi_t$. By Theorem 3.4 of Hall and Heyde (1980), we need to demonstrate three conditions:

(a) **Conditional Expectation** $\mathbb{E}[\zeta_t|\mathcal{F}_{t-1}] = 0$ for all $t \in [T]$;

(b) **Conditional Variance** $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\zeta_t^2|\mathcal{F}_{t-1}] = 1$

(c) **Conditional Lindeberg** $\frac{1}{T}; \sum_{t=1}^{T}\mathbb{E}[\zeta_t^2\mathbf{1}\Big\{|\zeta_t > \varepsilon\sqrt{T}\Big\}\,|\mathcal{F}_{t-1}] = o_p(1)$.

The first condition follows from the fact that $\hat{\sigma}_t$ is $\mathcal{F}_{t-1}$ measurable so

$$E[\hat{\sigma}^{-1}w_t\varphi_t \mid \mathcal{F}_{t-1}] = \hat{\sigma}^{-1}E[w_t\varphi_t \mid \mathcal{F}_{t-1}] = \hat{\sigma}^{-1}E_{P_e}[\varphi_{P_e}(Z)] = 0.$$

For the second condition, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\hat{\sigma}^{-2}w_t^2\varphi_t \mid \mathcal{F}_{t-1}] = \frac{1}{T}\sum_{t=1}^{T}\frac{\sigma_t^2}{\hat{\sigma}_t^2} = 1 + o_p(1).$$

For the third condition, we have a constant $c > 0$ such that $\zeta_t^2 = \hat{\sigma}_t^{-2}w_t^2\varphi_t^2 \leq \frac{w_t^2}{c^2}\varphi_t^2$. We therefore have for some $\varepsilon > 0$ that,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\zeta_t^2\mathbf{1}\{|\zeta_t| > \varepsilon\sqrt{T}\}\,\Big|\,\mathcal{F}_{t-1}\Big] \;\leq\; c^{-2}\cdot\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[w_t^2\varphi_t^2\mathbf{1}\left\{|w_t\varphi_t| > \varepsilon\sqrt{T}\right\}\Big|\,\mathcal{F}_{t-1}\Big] = o_p(1).$$

□

**Lemma 3** (Self-normalized CLT for plug-in quantities)**.** *Let $\{D_t\}_{t=1}$ be a martingale difference sequence and $\hat{D}_t$ be a $\mathcal{F}_{t-1}$ measurable estimate of $D_t$. Define*

$$S_T := \frac{1}{T}\sum_{t=1}^{T}D_t, \quad V_T := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[D_t^2 \mid \mathcal{F}_{t-1}], \quad \hat{V}_T := \frac{1}{T}\sum_{t=1}^{T}\hat{D}_t^2.$$

*Let $\hat{S}_T$ be a $\mathcal{F}_{t-1}$ measurable estimator of $S_T$. Suppose that the following conditions hold,*

*(a) (Conditional Variance) $V_T \xrightarrow{p} \kappa$ to some non-zero random variable $\kappa$ which is finite almost surely;*

*(b) (Realized Variance) $(TV_T)^{-1}\sum_{t=1}^{T}D_t^2 \xrightarrow{p} 1$;*

*(c) (Conditional Lindeberg) For every $\varepsilon > 0$; $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[D_t^2\mathbf{1}\{|D_t| > \varepsilon\sqrt{T}\}\,\Big|\,\mathcal{F}_{t-1}\Big] \to_p 0$;*

*(d) (Plug-In Rates) $\hat{S}_T - S_T = o_p(T^{-1/2})$, and $\frac{1}{T}\sum_{t=1}^{T}(\hat{D}_t - D_t)^2 = o_p(1)$.*

*Then $\sqrt{T}\frac{\hat{S}_T}{\sqrt{\hat{V}_T}} \xrightarrow{d} N(0,1)$.*

*Proof.* Let $R_T := \frac{1}{T}\sum_{t=1}^{T}D_t^2$. By assumption, $\frac{R_T}{V_T} \xrightarrow{p} 1$. Since condition (a) states that $V_T \to_p \kappa$, with $\kappa$ almost surely finite $R_T \xrightarrow{p} \kappa$. Now define the martingale difference array $X_{T,t} := \frac{D_t}{\sqrt{T}}$. Note that

$$\sum_{t=1}^{T}\mathbb{E}[X_{T,t}^2 \mid \mathcal{F}_{t-1}] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[D_t^2 \mid \mathcal{F}_{t-1}] = V_T \xrightarrow{p} \kappa.$$

Moreover, the conditional Lindeberg assumption ensure that

$$\sum_{t=1}^{T}\mathbb{E}\big[X_{T,t}^2\mathbf{1}\{|X_{T,t}| > \varepsilon\} \mid \mathcal{F}_{t-1}\big] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[D_t^2\mathbf{1}\{|D_t| > \varepsilon\sqrt{T}\} \mid \mathcal{F}_{t-1}\Big] \xrightarrow{p} 0.$$

Hence, by Theorem 3.2 of Hall and Heyde (1980) we have that,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} D_t \xrightarrow{st} \sqrt{\kappa}\, Z,$$

where $Z \sim N(0,1)$ and $\xrightarrow{st}$ denotes stable convergence. In particular, conditional on $\kappa$, the limit is Gaussian with variance $\kappa$, so the unconditional characteristic function is$\mathbb{E}\big[\exp\big(-\frac{1}{2}\kappa u^2\big)\big]$, corresponding to a mixture of centered Gaussian distributions.

Since $R_T \xrightarrow{p} \kappa$, stable convergence and convergence in probability imply joint convergence:

$$\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} D_t,\ R_T\right) \xrightarrow{d} (\sqrt{\kappa}\, Z,\ \kappa).$$

Because $\kappa > 0$ almost surely, the continuous mapping theorem yields

$$\sqrt{T}\,\frac{S_T}{\sqrt{R_T}} = \frac{T^{-1/2}\sum_{t=1}^{T} D_t}{\sqrt{R_T}} \xrightarrow{d} \frac{\sqrt{\kappa}\, Z}{\sqrt{\kappa}} = Z \sim N(0,1).$$

It remains to replace $(S_T, R_T)$ with $(\hat{S}_T, \hat{V}_T)$. By condition (d), we have that$\sqrt{T}\,(\hat{S}_T - S_T) \xrightarrow{p} 0$. Next, observe that

$$\hat{V}_T - R_T = \frac{1}{T}\sum_{t=1}^{T}(\hat{D}_t^2 - D_t^2) = \frac{1}{T}\sum_{t=1}^{T}(\hat{D}_t - D_t)^2 + \frac{2}{T}\sum_{t=1}^{T} D_t(\hat{D}_t - D_t).$$

The first term is $o_p(1)$ by condition (d). For the second term, by Cauchy–Schwarz,

$$\frac{1}{T}\left|\sum_{t=1}^{T} D_t(\hat{D}_t - D_t)\right| \le \left(\frac{1}{T}\sum_{t=1}^{T} D_t^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}(\hat{D}_t - D_t)^2\right)^{1/2} = R_T^{1/2}\cdot o_p(1) = o_p(1),$$

since $R_T \xrightarrow{p} \kappa$. Hence, $\hat{V}_T - R_T = o_p(1)$ and $\hat{V}_T \xrightarrow{p} \kappa$. Finally,

$$\sqrt{T}\,\frac{\hat{S}_T}{\sqrt{\hat{V}_T}} = \sqrt{T}\,\frac{S_T}{\sqrt{R_T}}\left(\frac{R_T}{\hat{V}_T}\right)^{1/2} + \sqrt{T}\,\frac{\hat{S}_T - S_T}{\sqrt{\hat{V}_T}}.$$

We have shown that $\sqrt{T}\,\frac{S_T}{\sqrt{R_T}} \xrightarrow{d} N(0,1)$, $\frac{R_T}{\hat{V}_T} \xrightarrow{p} 1$, and $\sqrt{T}\,\frac{\hat{S}_T - S_T}{\sqrt{\hat{V}_T}} \xrightarrow{p} 0$. Thus, by Slutsky's theorem, $\sqrt{T}\,\frac{\hat{S}_T}{\sqrt{\hat{V}_T}} \xrightarrow{d} N(0,1)$. □

**Lemma 4.** *For any measurable $f$, define the operator $P_T^a f := \frac{1}{T}\sum_{t=1}^{T} a_t^{-1} w_t f(Z_t)$, where the weights $\{a_t\}_{t=1}^{T}$ are $\mathcal{F}_{t-1}$ measurable and satisfy $a_t \ge c > 0$ uniformly in $t$, almost surely. Let*

$$F(\theta, \eta, g)(z) := m_\theta(z) - \dot{\ell}_{\theta,\eta}(z)[g]$$

*and define*

$$R_T := P_T^a\left(F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t) - F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})\right).$$

*Suppose the following conditions hold:*

*(a)* $\frac{1}{T}\sum_{t=1}^{T}\|\hat{\theta}_t - \theta_{P_e}\|_\Theta^4 = o_p(T^{-1})$, $\frac{1}{T}\sum_{t=1}^{T}\|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^4 = o_p(T^{-1})$, *and* $\frac{1}{T}\sum_{t=1}^{T}\|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^4 = o_p(T^{-1})$.

*(b)*

$$P_T^a\Big(\dot{m}_{\theta_{P_e}}(\cdot)[\hat{\theta}_t - \theta_{P_e}] - \ddot{\ell}_{\theta_{P_e},\eta_{P_e}}(\cdot)[\hat{\theta}_t - \theta_{P_e}, \alpha_{P_e}]\Big) = o_p(T^{-1/2}).$$

*(c)*

$$P_T^a\Big(\partial_\eta\partial_\theta\ell_{\theta_{P_e},\eta_{P_e}}(\cdot)[\alpha_{P_e}, \hat{\eta}_t - \eta_{P_e}]\Big) = o_p(T^{-1/2}).$$

*(d)*

$$P_T^a\Big(\partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(\cdot)[\hat{\alpha}_t - \alpha_{P_e}]\Big) = o_p(T^{-1/2}).$$

*(e)*

$$\frac{1}{T^2}\sum_{t=1}^{T} \mathbb{E}\left[w_t^4 C(Z_t)^4 \mid \mathcal{F}_{t-1}\right] = O_p(1).$$

*Then, under Assumptions 1–4, $R_T = o_p(T^{-1/2})$.*

*Proof.* A first-order functional Taylor expansion of $F$ around $(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})$ yields

$$\begin{aligned}
&F(\hat{\theta}_t, \hat{\eta}_t, \hat{\alpha}_t)(z) - F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})(z) \\
&\quad = \partial_\theta F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})(z)[\hat{\theta}_t - \theta_{P_e}] \\
&\qquad + \partial_\eta F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})(z)[\hat{\eta}_t - \eta_{P_e}] \\
&\qquad + \partial_g F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})(z)[\hat{\alpha}_t - \alpha_{P_e}] + R_{2,t}(z),
\end{aligned}$$

where $R_{2,t}(z)$ is a second-order remainder term. In particular,

$$\begin{aligned}
\partial_\theta F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})(z)[h] &= \dot{m}_{\theta_{P_e}}(z)[h] - \ddot{\ell}_{\theta_{P_e},\eta_{P_e}}(z)[h, \alpha_{P_e}], \\
\partial_\eta F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})(z)[u] &= -\partial_\eta \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\alpha_{P_e}, u],
\end{aligned}$$

and

$$\partial_g F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})(z)[h] = -\dot{\ell}_{\theta_{P_e},\eta_{P_e}}(z)[h].$$

Condition (a) implies that the maximum estimation error over $t \le T$ converges to zero in probability for each of the three estimators. Thus, with probability tending to one, the local bounds in Assumption 4 apply simultaneously for all $t \le T$. On this event, after enlarging $C$ by a fixed constant if necessary,

$$\begin{aligned}
|R_{2,t}(Z_t)| \le C(Z_t)\Big(&\|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2 + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^2 \\
&+ \|\hat{\theta}_t - \theta_{P_e}\|_\Theta \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon \\
&+ \|\hat{\theta}_t - \theta_{P_e}\|_\Theta \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta \\
&+ \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta\Big).
\end{aligned}$$

Note that there is no quadratic term in $\|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta$, since $F(\theta, \eta, g)$ is affine in $g$.

Applying the operator $P_T^a$ to the Taylor expansion gives

$$\begin{aligned}
R_T = &P_T^a\left[\dot{m}_{\theta_{P_e}}(\cdot)[\hat{\theta}_t - \theta_{P_e}] - \ddot{\ell}_{\theta_{P_e},\eta_{P_e}}(\cdot)[\hat{\theta}_t - \theta_{P_e}, \alpha_{P_e}]\right] \\
&- P_T^a\left[\partial_\eta \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(\cdot)[\alpha_{P_e}, \hat{\eta}_t - \eta_{P_e}]\right] \\
&- P_T^a\left[\dot{\ell}_{\theta_{P_e},\eta_{P_e}}(\cdot)[\hat{\alpha}_t - \alpha_{P_e}]\right] + P_T^a R_{2,t}.
\end{aligned}$$

The first three terms on the right-hand side are each $o_p(T^{-1/2})$ by conditions (b)–(d). It therefore remains to control the final term.

Using the above bound on $R_{2,t}$,

$$|P_T^a R_{2,t}| \le \frac{1}{T}\sum_{t=1}^{T} a_t^{-1} w_t C(Z_t)\delta_t,$$

where

$$\begin{aligned}
\delta_t := &\|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2 + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^2 \\
&+ \|\hat{\theta}_t - \theta_{P_e}\|_\Theta \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon \\
&+ \|\hat{\theta}_t - \theta_{P_e}\|_\Theta \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta \\
&+ \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta.
\end{aligned}$$

Since $w_t \geq 0$, the Cauchy–Schwarz inequality gives

$$|P_T^a R_{2,t}| \leq \left(\frac{1}{T}\sum_{t=1}^{T} a_t^{-2} w_t C(Z_t)^2\right)^{1/2} \left(\frac{1}{T}\sum_{t=1}^{T} w_t \delta_t^2\right)^{1/2}.$$

It therefore suffices to show that the first average is $O_p(1)$ and the second is $o_p(T^{-1})$.

**Showing $\frac{1}{T}\sum_{t=1}^{T} a_t^{-2} w_t C(Z_t)^2 = O_p(1)$.** By predictability and the change-of-measure identity,

$$\mathbb{E}\left[a_t^{-2} w_t C(Z_t)^2 \mid \mathcal{F}_{t-1}\right] = a_t^{-2}\mathbb{E}_{P_e}[C(Z)^2] \leq c^{-2}\mathbb{E}_{P_e}[C(Z)^2].$$

Since $C \in L^2(P_e)$, taking expectations and applying Markov's inequality gives

$$\frac{1}{T}\sum_{t=1}^{T} a_t^{-2} w_t C(Z_t)^2 = O_p(1).$$

**Showing $\frac{1}{T}\sum_{t=1}^{T} w_t \delta_t^2 = o_p(T^{-1})$.** Since $(a_1 + \cdots + a_5)^2 \leq 5(a_1^2 + \cdots + a_5^2)$,

$$\begin{aligned}\delta_t^2 \lesssim\ & \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^4 + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^4 \\ & + \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2 \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^2 \\ & + \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2 \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^2 \\ & + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^2 \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^2.\end{aligned}$$

By condition (a), the averages of the fourth-power terms are $o_p(T^{-1})$. The averages of the mixed terms are also $o_p(T^{-1})$ by Cauchy–Schwarz. For example,

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T} \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2 \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^2 \\ &\quad \leq \left(\frac{1}{T}\sum_{t=1}^{T} \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^4\right)^{1/2} \left(\frac{1}{T}\sum_{t=1}^{T} \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^4\right)^{1/2} = o_p(T^{-1}),\end{aligned}$$

and similarly for the remaining mixed terms. Hence

$$\frac{1}{T}\sum_{t=1}^{T} \delta_t^2 = o_p(T^{-1}), \qquad \text{or equivalently} \qquad \sum_{t=1}^{T} \delta_t^2 = o_p(1).$$

To obtain the corresponding weighted bound, note that $\delta_t$ is $\mathcal{F}_{t-1}$ measurable and $\mathbb{E}[w_t \mid \mathcal{F}_{t-1}] = 1$. For any $b > 0$, predictability therefore gives

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T} w_t \delta_t^2 \mathbf{1}\left\{\sum_{s=1}^{t} \delta_s^2 \leq b\right\}\right] \\ &\quad = \mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \mathbf{1}\left\{\sum_{s=1}^{t} \delta_s^2 \leq b\right\}\right] \leq b.\end{aligned}$$

Consequently, for any $\varepsilon, b > 0$, Markov's inequality yields

$$\Pr\left(\sum_{t=1}^{T} w_t \delta_t^2 > \varepsilon\right) \leq \Pr\left(\sum_{t=1}^{T} \delta_t^2 > b\right) + \frac{b}{\varepsilon}.$$

Taking $T \to \infty$ and then $b \downarrow 0$ proves $\sum_{t=1}^{T} w_t \delta_t^2 = o_p(1)$, and thus

$$\frac{1}{T}\sum_{t=1}^{T} w_t \delta_t^2 = o_p(T^{-1}).$$

Combining the two bounds gives $P_T^a R_{2,t} = o_p(T^{-1/2})$. Together with conditions (b)–(d), this proves $R_T = o_p(T^{-1/2})$. □

**Lemma 5.** *Let $\{f_t\}_{t=1}^T$ be a predictable sequence of measurable maps $f_t : \mathcal{Z} \to \mathbb{R}$ such that*

*(a)* $\mathbb{E}[w_t f_t(Z_t) \mid \mathcal{F}_{t-1}] = 0$ *for each* $t \in [T]$ *almost surely.*

*(b) There exists a measurable function* $C : \mathcal{Z} \to \mathbb{R}_+$ *and nonnegative* $\mathcal{F}_{t-1}$ *measurable variables* $\delta_t$ *such that* $|f_t(z)| \le C(z)\delta_t$ *for all* $z \in \mathcal{Z}$ *and* $\frac{1}{T}\sum_{t=1}^T \delta_t^4 = o_p(T^{-1})$.

*(c)* $\frac{1}{T^2}\sum_{t=1}^T \left(\mathbb{E}[w_t^2 C(Z_t)^2 \mid \mathcal{F}_{t-1}]\right)^2 = O_p(1)$.

*Let $\{a_t\}_{t=1}^T$ be an $\mathcal{F}_{t-1}$ measurable sequence such that $\inf_{t\in[T]} a_t > c$ almost surely for some constant $c > 0$. Then, under Assumptions 1–3,*

$$\frac{1}{T}\sum_{t=1}^T a_t^{-1} w_t f_t(Z_t) = o_p(T^{-1/2}).$$

*Proof.* Define $\zeta_t := a_t^{-1} w_t f_t(Z_t)$. Since $a_t^{-1}$ is $\mathcal{F}_{t-1}$ measurable, condition (a) gives $\mathbb{E}[\zeta_t \mid \mathcal{F}_{t-1}] = 0$. Thus $\{\zeta_t\}_{t=1}^T$ is a martingale difference sequence. We first show that its normalized predictable variance satisfies

$$V_T := \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}] = o_p(1).$$

By condition (b),

$$|\zeta_t|^2 = a_t^{-2} w_t^2 f_t(Z_t)^2 \le a_t^{-2} w_t^2 C(Z_t)^2 \delta_t^2.$$

Taking conditional expectations and using predictability gives

$$\mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}] \le a_t^{-2}\delta_t^2 \mathbb{E}[w_t^2 C(Z_t)^2 \mid \mathcal{F}_{t-1}].$$

Moreover, $a_t^{-2} \le c^{-2}$ uniformly, so

$$\mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}] \le c^{-2}\delta_t^2 \mathbb{E}[w_t^2 C(Z_t)^2 \mid \mathcal{F}_{t-1}].$$

Applying Cauchy–Schwarz,

$$\begin{aligned} V_T &\le \frac{c^{-2}}{T}\sum_{t=1}^T \delta_t^2 \mathbb{E}[w_t^2 C(Z_t)^2 \mid \mathcal{F}_{t-1}] \\ &\le c^{-2}\left(\sum_{t=1}^T \delta_t^4\right)^{1/2}\left(\frac{1}{T^2}\sum_{t=1}^T \left(\mathbb{E}[w_t^2 C(Z_t)^2 \mid \mathcal{F}_{t-1}]\right)^2\right)^{1/2}. \end{aligned}$$

By condition (b), the first factor is $o_p(1)$, while condition (c) makes the second factor $O_p(1)$. Hence $V_T = o_p(1)$.

Finally, for any $\varepsilon, b > 0$, a stopped martingale second-moment bound gives

$$\Pr\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^T \zeta_t\right| > \varepsilon\right) \le \Pr(V_T > b) + \frac{b}{\varepsilon^2}.$$

Indeed, retain the increment $\zeta_t$ only when $T^{-1}\sum_{s=1}^t \mathbb{E}[\zeta_s^2 \mid \mathcal{F}_{s-1}] \le b$. This restriction is predictable, and the retained sum divided by $\sqrt{T}$ has second moment at most $b$. It agrees with the original sum whenever $V_T \le b$, giving the displayed bound by Chebyshev's inequality.

Taking $T \to \infty$ and then $b \downarrow 0$ yields $T^{-1/2}\sum_{t=1}^T \zeta_t = o_p(1)$. Therefore,

$$\frac{1}{T}\sum_{t=1}^T a_t^{-1} w_t f_t(Z_t) = o_p(T^{-1/2}),$$

as required. □

**Lemma 6.** *Under the conditions of Theorem 2, if* $\|\alpha_{P_e}\|_\Theta < \infty$*, then*

$$\sup_{t\in[T]} \|\hat{\alpha}_t\|_\Theta = O_p(1).$$

*Proof.* Since all terms are nonnegative,

$$\sup_{t\in[T]} \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^4 \le \sum_{t=1}^T \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^4 = T \cdot \frac{1}{T}\sum_{t=1}^T \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^4 = o_p(1).$$

Taking fourth roots gives that $\sup_{t\in[T]} \|\hat{\alpha}_t - \alpha_{P_e}\| = o_p(1)$. The final claim follows from

$$\sup_{t\le T} \|\hat{\alpha}_t\|_\Theta \le \|\alpha_{P_e}\|_\Theta + \sup_{t\in[T]} \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta.$$

□

**Lemma 7.** *Assume the conditions of Theorem 2. Let* $\hat{\varphi}_t(z) := m_{\hat{\theta}_t}(z) - \hat{\Psi}_t - \partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}(z)[\hat{\alpha}_t]$*, where* $\hat{\Psi}_t$ *is an* $\mathcal{F}_{t-1}$ *measurable estimator such that* $\sup_{t\in[T]} |\hat{\Psi}_t| = O_p(1)$*. Assume the overlap condition* $\sup_{\pi\in\Pi} \sup_{c,a} \frac{\pi_e(a|c)}{\pi(a|c)} \le K_T$ *for some* $K_T < \infty$*. Writing* $\varphi_t = \varphi_{P_e}(Z_t)$ *for the oracle influence function evaluated at* $Z_t$*, for any fixed* $k \ge 1$*,*

$$\sup_{t\in[T]} \mathbb{E}\left[w_t^k \hat{\varphi}_t^4(Z_t) \mid \mathcal{F}_{t-1}\right] = O_p(K_T^{k-1}) \qquad \textit{and} \qquad \sup_{t\in[T]} \mathbb{E}\left[w_t^k \varphi_t^4 \mid \mathcal{F}_{t-1}\right] = O_p(K_T^{k-1}).$$

*Proof.* By Lemma 6,

$$\sup_{t\in[T]} \|\hat{\alpha}_t\|_\Theta = O_p(1).$$

Furthermore, by condition (2a) of Theorem 2,

$$\sup_{t\in[T]} \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^4 \le \sum_{t=1}^T \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^4 = o_p(1),$$

and

$$\sup_{t\in[T]} \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^4 \le \sum_{t=1}^T \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^4 = o_p(1).$$

Taking fourth roots gives uniform consistency of both estimators. Therefore, with probability tending to one, $(\hat{\theta}_t, \hat{\eta}_t)$ lies in the neighborhoods where Assumption 4 holds, simultaneously for all $t \in [T]$. In particular, on this event,

$$|m_{\hat{\theta}_t}(z)| \le C(z), \qquad \text{for all } t \in [T].$$

Next, decompose

$$\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}(z)[\hat{\alpha}_t] = \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\hat{\alpha}_t] + R_t(z),$$

where, by the derivative bounds in Assumption 4,

$$|R_t(z)| \le C(z)\|\hat{\alpha}_t\|_\Theta \left(\|\hat{\theta}_t - \theta_{P_e}\|_\Theta + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon\right).$$

For any $\varepsilon > 0$, we can choose a fixed $L < \infty$ such that, for all sufficiently large $T$, with probability at least $1 - \varepsilon$, the estimators lie in these neighborhoods and, uniformly over $t \in [T]$,

$$\|\hat{\alpha}_t\|_\Theta \le L, \quad \|\hat{\theta}_t - \theta_{P_e}\|_\Theta \le 1, \quad \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon \le 1, \quad |\hat{\Psi}_t| \le L.$$

On this event, $|R_t(z)| \le 2LC(z)$ and

$$\left|\partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\hat{\alpha}_t]\right| \le C(z)\|\hat{\alpha}_t\|_\Theta \le LC(z).$$

Combining these bounds gives

$$|\hat{\varphi}_t(z)| \le L'(1 + C(z))$$

for some finite constant $L'$ depending only on $L$, uniformly over $t \in [T]$.

Since the estimators are $\mathcal{F}_{t-1}$ measurable, the overlap bound and change-of-measure identity give

$$\mathbb{E}\left[w_t^k \hat{\varphi}_t^4(Z_t) \mid \mathcal{F}_{t-1}\right] \leq K_T^{k-1} \mathbb{E}\left[w_t \hat{\varphi}_t^4(Z_t) \mid \mathcal{F}_{t-1}\right] = K_T^{k-1} \mathbb{E}_{P_e}[\hat{\varphi}_t^4(Z)].$$

Consequently, on the preceding event,

$$\begin{aligned}\sup_{t\in[T]} \mathbb{E}\left[w_t^k \hat{\varphi}_t^4(Z_t) \mid \mathcal{F}_{t-1}\right] &\leq K_T^{k-1} \sup_{t\in[T]} \mathbb{E}_{P_e}[\hat{\varphi}_t^4(Z)] \\ &\leq (L')^4 K_T^{k-1} \mathbb{E}_{P_e}[(1+C(Z))^4].\end{aligned}$$

By Assumption 4, $\mathbb{E}_{P_e}[(1+C(Z))^4] < \infty$. Since $\varepsilon > 0$ was arbitrary, this proves

$$\sup_{t\in[T]} \mathbb{E}\left[w_t^k \hat{\varphi}_t^4(Z_t) \mid \mathcal{F}_{t-1}\right] = O_p(K_T^{k-1}).$$

For the oracle influence function, the same envelope bounds give

$$|\varphi_{P_e}(z)| = \left|m_{\theta_{P_e}}(z) - \Psi_e - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\alpha_{P_e}]\right| \leq L_0(1+C(z))$$

for some fixed finite constant $L_0$. Repeating the change-of-measure argument yields

$$\sup_{t\in[T]} \mathbb{E}\left[w_t^k \varphi_t^4 \mid \mathcal{F}_{t-1}\right] \leq L_0^4 K_T^{k-1} \mathbb{E}_{P_e}[(1+C(Z))^4] = O_p(K_T^{k-1}),$$

which completes the proof. □

**Lemma 8.** *Assume the conditions of Theorem 2. Let* $\hat{\varphi}_t(z) := m_{\hat{\theta}_t}(z) - \hat{\Psi}_t - \partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}(z)[\hat{\alpha}_t]$ *where* $\hat{\Psi}_t$ *is a* $\mathcal{F}_{t-1}$ *measurable estimator such that* $\sup_{t\in[T]} \hat{\Psi}_t = O_p(1)$. *Assume the overlap condition* $\sup_{\pi\in\Pi} \sup_{c,a} \frac{\pi_e}{\pi} \leq K_T$ *for some* $K_T < \infty$. *Then, for any fixed* $k \geq 1$,

$$\sup_{t\in[T]} \mathbb{E}[w_t^k \hat{\varphi}_t^4(Z_t) \mid \mathcal{F}_{t-1}] = O_p(1) \qquad \textit{and} \qquad \sup_{t\in[T]} \mathbb{E}[w_t^k \varphi_t^4(Z_t) \mid \mathcal{F}_{t-1}] = O_p(1).$$

*Proof.* By Lemma 6,

$$\sup_{t\in[T]} \|\hat{\alpha}_t\|_\Theta = O_p(1).$$

Furthermore, repeating the arguments in Lemma 6 note that by condition (2a) of Theorem 2, we have that

$$\sup_{t\in[T]} \left\|\hat{\theta}_t - \theta_{P_e}\right\|_\Theta^4 \leq \sum_{t=1}^T \left\|\hat{\theta}_t - \theta_{P_e}\right\|_\Theta^4 = o_p(1)$$

$$\sup_{t\in[T]} \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^4 \leq \sum_{t=1}^T \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^4 = o_p(1)$$

by the same arguments. Then, by taking fourth roots, we have that $\sup_{t\in[T]} \left\|\hat{\theta}_t - \theta_{P_e}\right\|_\Theta = o_p(1)$. Therefore, $(\hat{\theta}_t, \hat{\eta}_t)$ lies in a neighborhood where the envelope condition of Assumption 3 holds.

In particular,

$$|m_{\hat{\theta}_t}(z)| \leq C(z), \qquad \text{for all } t \in [T].$$

Next, decompose

$$\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}(z)[\hat{\alpha}_t] = \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\hat{\alpha}_t] + R_t(z),$$

where, by the same arguments as Theorem 1,

$$|R_t(z)| \leq C(z)\|\hat{\alpha}_t\|_\Theta \left(\|\hat{\theta}_t - \theta_{P_e}\|_\Theta + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon\right).$$

Therefore, on an event with probability tending to 1, there exists a constant $L < \infty$ such that, uniformly over $t \in [T]$,

$$\|\hat{\alpha}_t\|_\Theta \leq L, \quad \|\hat{\theta}_t - \theta_{P_e}\|_\Theta \leq 1, \quad \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon \leq 1, \quad |\hat{\Psi}_t| \leq L.$$

Thus, $|R_t(z)| \le LC(z)$ and $|\partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(z)[\hat{\alpha}_t]| \le C(z)\|\hat{\alpha}_t\|_\Theta \le LC(z)$. Combining everything, we have that

$$|\hat{\varphi}_t(z)| = |m_{\hat{\theta}_t}(z) - \hat{\Psi}_t - \partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}(z)[\hat{\alpha}_t]| \le L'(1 + C(z))$$

for some finite constant $L'$, uniformly over $t \in [T]$ on an event with probability tending to 1. Therefore, on this event,

$$\begin{aligned}
\sup_{t\in[T]} \mathbb{E}\left[w_t^k \hat{\varphi}_t^4(Z_t) \mid \mathcal{F}_{t-1}\right] &\le \sup_{t\in[T]} (L')^4 \mathbb{E}\left[w_t^k (1 + C(Z_t))^4 \mid \mathcal{F}_{t-1}\right] \\
&\le \sup_{t\in[T]} (L')^4 \mathbb{E}\left[\left(\sup_{t\in[T]} w_t^{k-1}\right) w_t (1 + C(Z_t))^4 \mid \mathcal{F}_{t-1}\right] \\
&\le \sup_{t\in[T]} (L')^4 K_T^{k-1} \mathbb{E}_{P_e}\left[(1 + C(Z))^4\right].
\end{aligned}$$

However, $\mathbb{E}_{P_e}\left[(1 + C(Z))^4\right]$ is bounded by assumption so the entire term is $O_p(1)$. For the oracle influence function, applying the envelope condition again gives us

$$|\varphi_t| = |m_{\theta_{P_e}}(Z_t) - \Psi_e - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}(Z_t)[\alpha_{P_e}]| \le L(1 + C(Z_t))$$

for some finite constant $L$. Repeating the same steps as above gives us that this term is $O_p(1)$.

□

**Lemma 9.** *Let $F(\theta,\eta,\alpha)(z) := m_{\theta_t}(z) - \partial_\theta \ell_{\theta,\eta}(z)[\alpha]$ and $F_0(z) := F(\theta_{P_e}, \eta_{P_e}, \alpha_{P_e})$. Then under the conditions of Theorem 2,*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{P_e}\left[\left(F(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(Z) - F_0(Z)\right)^2\right] = o_p(T^{-1/2}).$$

*Proof.* We decompose

$$F(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(z) - F_0(z) = \{m_{\hat{\theta}_t}(z) - m_{\theta_{P_e}}(z)\} - \left\{\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}[\hat{\alpha}_t](z) - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}[\alpha_{P_e}](z)\right\}.$$

We have using Assumption 4 that

$$|m_{\hat{\theta}_t}(z) - m_{\theta_{P_e}}(z)| \le \int_0^1 C(z)\|\hat{\theta}_t - \theta_{P_e}\|_\Theta \, ds = C(z)\|\hat{\theta}_t - \theta_{P_e}\|_\Theta.$$

Therefore,

$$\mathbb{E}_{P_e}[(m_{\hat{\theta}_t} - m_{\theta_{P_e}})^2] \le \mathbb{E}_{P_e}[C(Z)^2]\|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2.$$

Next, write

$$\begin{aligned}
\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}[\hat{\alpha}_t] - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}[\alpha_{P_e}] &= \partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t}[\Delta_{\alpha,t}] + \left\{\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t} - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}\right\}[\alpha_{P_e}] \\
&= \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}[\Delta_{\alpha,t}] + \left\{\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t} - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}\right\}[\Delta_{\alpha,t}] + \left\{\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t} - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}\right\}[\alpha_{P_e}]
\end{aligned}$$

For the first term, we have by Assumption 4 that $\mathbb{E}_{P_e}\left[\left\{\partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}[\alpha_t - \alpha_{P_e}]\right\}^2\right] \lesssim \|\alpha_t - \alpha_{P_e}\|_\Theta^2$.
For the second and third term, consider that

$$\begin{aligned}
\left\{\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t} - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}\right\}[h] = &\underbrace{\int_0^1 \partial_\theta^2 \ell_{\theta_{P_e}+s\Delta_{\theta,t},\eta_{P_e}+s\Delta_{\eta,t}}[\Delta_{\theta,t}, h]\, ds}_{R_1 t} \\
&+ \underbrace{\int_0^1 \partial_\eta \partial_\theta \ell_{\theta_{P_e}+s\Delta_{\theta,t},\eta_{P_e}+s\Delta_{\eta,t}}[h, \Delta_{\eta,t}]\, ds}_{R_{2t}}.
\end{aligned}$$

Again applying the envelope condition Assumption 4,

$$\left|\left\{\partial_\theta \ell_{\hat{\theta}_t,\hat{\eta}_t} - \partial_\theta \ell_{\theta_{P_e},\eta_{P_e}}\right\}[h]\right| \lesssim C(Z)\left(\|\hat{\theta}_t - \theta_{P_e}\|_\Theta + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon\right)\|h\|_\Theta.$$

Taking $h = \hat{\alpha}_t - \alpha_{P_e}$ yields

$$\begin{aligned}\mathbb{E}_{P_e}[R_{1t}^2] \lesssim\ & \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^2 \\ & + \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^2\left(\|\hat{\theta}_t - \theta_{P_e}\|_\Theta + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon\right)^2.\end{aligned}$$

Similarly, taking $h = \alpha_{P_e}$ gives

$$\mathbb{E}_{P_e}[R_{2t}^2] \lesssim \|\alpha_{P_e}\|_\Theta^2\left(\|\hat{\theta}_t - \theta_{P_e}\|_\Theta + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon\right)^2.$$

Combining the preceding bounds,

$$\begin{aligned}\mathbb{E}_{P_e}\left[\left(F_t(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(Z) - F_0(Z)\right)^2\right] \lesssim\ & \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2 + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^2 + \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^2 \\ & + \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^2\left(\|\hat{\theta}_t - \theta_{P_e}\|_\Theta + \|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon\right)^2.\end{aligned}$$

Averaging over $t$, condition (2a) of Theorem 2 implies

$$\frac{1}{T}\sum_{t=1}^T \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2 \le \left(\frac{1}{T}\sum_{t=1}^T \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^4\right)^{1/2} = o_p(T^{-1/2}),$$

and similarly for $\hat{\eta}_t$ and $\hat{\alpha}_t$. Moreover,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^2\|\hat{\theta}_t - \theta_{P_e}\|_\Theta^2 &\le \left(\frac{1}{T}\sum_{t=1}^T \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^4\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^T \|\hat{\theta}_t - \theta_{P_e}\|_\Theta^4\right)^{1/2} \\ &= o_p(T^{-1}),\end{aligned}$$

and the same argument yields $\frac{1}{T}\sum_{t=1}^T \|\hat{\alpha}_t - \alpha_{P_e}\|_\Theta^2\|\hat{\eta}_t - \eta_{P_e}\|_\Upsilon^2 = o_p(T^{-1})$.

Therefore,

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{P_e}\left[\left(F_t(\hat{\theta}_t,\hat{\eta}_t,\hat{\alpha}_t)(Z) - F_0(Z)\right)^2\right] = o_p(T^{-1/2}).$$

□

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract's claims are made more explicit in the "Summary of Contributions" paragraph, which provides direct reference to the key theorems.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, this is discussed in the context of the assumptions required as well as in the concluding remarks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes, all Theorem statements have proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, and the code used to run the experiments is attached.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, and the code is available in the supplement.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Yes, and the code is available in the supplement.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Yes, you can see error bars in the empirical results for CI length of the competing methods.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, see Section B.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Yes. No human subjects or private datasets were used. This is theoretical work with low potential for misuse.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: I do not see any direct path to misuse as this is a generic inferential method.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: All data is simulated or open source

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: All data is simulated or open source

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are released

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: There were no human subjects or crowd sourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: There were no human subjects or crowd sourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The original version of this manuscript was developed without substantive LLM assistance, apart from routine writing and editing support. Subsequent revisions utilized LLMs to identify potential gaps in proof arguments, assist with improvements to simulation code, and help address suggestions raised during peer review. All mathematical proofs were written and carefully reviewed by the authors, with LLM assistance limited to identifying potential gaps and suggesting improvements. Similarly, all code was reviewed by the authors, who take full responsibility for the correctness of the mathematical arguments, computational results, and any remaining errors.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.