Carryover detection in switchback experimentation

Paul Missault
Amazon
Amazon, Luxemburg 1855, LU
pmissaul@amazon.com

Lorenzo Masoero
Amazon
Amazon, Seattle WA 98109
masoerl@amazon.com

Abstract

Switchback experiments assign units to treatment and control over time, yielding more precise causal estimates than fixed designs but risking bias from *carryover effects*, where past treatments influence future outcomes. Existing estimators require specifying an *influence period*, i.e. an upper bound on carryover duration, often guessed from intuition. We propose a statistical test that detects when this bound is underestimated by comparing estimators with different lag assumptions and rejecting results when significant differences indicate bias. The method has theoretical guarantees under standard assumptions and shows strong performance in simulations with synthetic effects on real-world covariates. This provides practitioners with a practical safeguard against erroneous conclusions in the presence of unaddressed carryover.

1 Introduction

Switchback designs vary treatment assignment of units over time, exposing each to both control and treatment conditions. These designs yield more precise estimates than fixed-assignment approaches [1, 2] but introduce potential bias through *carryover effects*—where past treatments influence future outcomes. Carryover effects have an *order* m, meaning treatment at time t affects outcomes until time $t + m$ with m possibly infinite. Recent work has addressed these designs in both theoretical [3, 4] and applied contexts [1, 5]. Most notably, [6] developed a method for unbiased causal estimates from switchback designs by assuming m is finite and experimenters can estimate an upper bound on the order m. Similarly, [3] allow for infinite order carryover, but assume effects decay exponentially according to a *mixing time*.

In general, all experimental designs can suffer from both temporal interference or cross-unit interference. Temporal interference is particularly relevant to switchback designs since units change treatment multiple times during an experiment period. General approaches to interference include multilevel designs [7] and interference networks [8], but these require prior knowledge of interaction patterns. For carryover effects specifically, [9] proposed linear mixed-effects models, while [10] modeled both cross-unit and temporal interference but offered no method to estimate key parameters.

A critical limitation of unbiased estimators of the causal effect in switchback designs is that they require experimenters to specify a carryover duration (or related concept such as mixing time) in advance. In practice, determining this parameter is challenging [11]. When underestimated, estimators are not guaranteed to be unbiased. While [6] proposed identifying carryover bounds through repeated experiments, this approach is resource-intensive and lacks theoretical justification.

We address this challenge by proposing a statistical test that detects when estimators are biased due to underestimated carryover effects, providing theoretical guarantees under an additive linear effects model.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CauScien: Uncovering Causality in Science.

2 Switchback Experimentation

Following [2, 6], we formally introduce switchback experiments. We consider a binary experiment with $[N] := \{1, 2, ..., N\}$ units and $[T] := \{1, 2, ..., T\}$ timesteps. Let $W \in \{0, 1\}^{N \times T}$ denote the treatment assignment matrix, where $W_{n,t}$ indicates whether unit n is treated at timestep t. The potential outcome $Y_{n,t}(W)$ represents the outcome of unit n at time t under treatment exposure W. We define $Y^C_{n,t} = Y_{n,t}(\mathbf{0}^{N \times T})$ as the outcome when all units are untreated, and $Y^T_{n,t} = Y_{n,t}(\mathbf{1}^{N \times T})$ when all are treated.

2.1 Experimental setup

In a switchback design, treatment assignment is randomly re-assigned for each unit at each timestep, as illustrated in Figure 1. From [6], experimental power increases with more frequent re-randomization.

	t_1	t_2	t_3	t_4	t_5	t_6
n_4	–	T	C	C	T	–
n_3	–	C	T	C	T	–
n_2	–	T	C	C	C	–
n_1	–	C	C	T	T	–

Figure 1: A switchback experiment with 4 units run for 4 timesteps. C=control, T=treatment.

However, as the time between randomizations decreases, the likelihood of significant carryover effects increases, potentially biasing estimators.

2.2 Carryover effects

The potential outcome $Y_{n,t}(W)$ may depend on the entire treatment history, not just the current assignment $W_{n,t}$. Outcomes can depend on prior assignments (e.g., a prior price reduction affecting current purchasing behavior) or assignments of other units. To rigorously handle carryover effects, we make two key assumptions.

Assumption 1. ***(No unit spillovers)****. The potential outcome of unit n at time t depends only on its own treatment history, not on the treatment of other units.*

This allows us to write $Y_{n,t}(W)$ as $Y_{n,t}(W_{n,\cdot})$, where $W_{n,\cdot}$ is the row-vector for unit n. This assumption is also known as SUTVA in literature, and is strong for real-world applications. We therefore note that whenever cross-unit interference is a concern, units can and should be clustered to block interference effects [12, 13].

Assumption 2. ***(m-Carryover Effects)****. There exists a fixed duration m such that for any unit n and time t:*

$$Y_{n,t+m+1}(W) = Y_{n,t+m+1}(W').$$

for any treatment assignments W and W' that differ only at or before time t for unit n. In other words, a treatment at time t affects outcomes only up to time $t + m$.

This assumption states that treatment assignments more than m timesteps in the past do not affect current outcomes. We define our causal estimand of interest, the *lag-ℓ* average causal effect:

$$\tau_\ell = \frac{1}{N(T-\ell)} \sum_{n=1}^{N} \sum_{t=\ell+1}^{T} (Y^T_{n,t} - Y^C_{n,t}). \tag{1}$$

When $\ell = 0$, this is the standard Average Treatment Effect (ATE). Using $\ell > 0$ accounts for carryover effects by discarding initial "burn-in" timesteps. We define $\mathbb{I}_{n,t}(\ell) = I(W_{n,t-\ell:t} = 1^{1\times\ell+1})$ and $\mathbb{O}_{n,t}(\ell) = I(W_{n,t-\ell:t} = 0^{1\times\ell+1})$ where $1^{1\times\ell+1}$ and $0^{1\times\ell+1}$ are vectors of 1's or 0's respectively. To estimate τ_ℓ, we use the estimator from [2]:

$$\hat{\tau}_\ell(W) = \frac{1}{N(T-\ell)} \sum_{n=1}^{N} \sum_{t=\ell+1}^{T} \left[\frac{\mathbb{I}_{n,t}(\ell)}{Prob(\mathbb{I}_{n,t}(\ell))} Y_{n,t}(W) - \frac{\mathbb{O}_{n,t}(\ell)}{Prob(\mathbb{O}_{n,t}(\ell))} Y_{n,t}(W) \right]. \tag{2}$$

This estimator is unbiased for τ_ℓ when $\ell \geq m$. However, determining an appropriate ℓ that upper bounds m is challenging in practice. Instead of requiring experimenters to fully understand all sources of carryover effects, we propose a statistical test to detect when ℓ is insufficient.

3 Testing for bias

To detect when ℓ underestimates m, we note that for any $\ell \geq m$ (where m is the true carryover duration), the estimator in eq. (2) provides an unbiased estimate of τ_ℓ, provided the randomization is not degenerate (i.e., $Prob(\mathbb{O}_{i,t}(\ell)) \neq 0$ and $Prob(\mathbb{I}_{i,t}(\ell)) \neq 0$). We adopt a linear effects model widely used in causal inference literature [14, 6]:

$$Y_{i,t}(W_{i,1:t}) = \mu_i + \alpha_{i,t} + \delta^{(0)} W_{i,t} + \delta^{(1)} W_{i,t-1} + ... + \delta^{(m)} W_{i,t-m} + \epsilon_{i,t} \tag{3}$$

where $\delta^{(j)}$ represents the effect of treatment j timesteps in the past.

We introduce an offset estimator that will help us detect when $\ell < m$:

$$\hat{\tau}_{\ell,\ell'}(W) = \frac{1}{N(T-\ell')} \sum_{i=1}^{N} \sum_{t=\ell'+1}^{T} \left[\frac{\mathbb{I}_{i,t}(\ell)}{Prob(\mathbb{I}_{i,t}(\ell))} Y_{i,t}(W) - \frac{\mathbb{O}_{i,t}(\ell)}{Prob(\mathbb{O}_{i,t}(\ell))} Y_{i,t}(W) \right] \tag{4}$$

Note that $\hat{\tau}_{\ell,\ell}(W) = \hat{\tau}_\ell(W)$. The key insight comes from examining the difference between these estimators:

Theorem 1 (Expected Difference Under Linear Effects Model)**.** *Under the linear effects model eq.* (3)*, experiment duration of T timesteps, fair coin randomization and $\ell' \geq \ell + 1$ with $\ell' < T$:*

- *If $\ell \geq m$, then $\mathbb{E}[\hat{\tau}_{\ell,\ell'}(W) - \hat{\tau}_{\ell+1,\ell'}(W)] = 0$*
- *If $\ell < m$, then $\mathbb{E}[\hat{\tau}_{\ell,\ell'}(W) - \hat{\tau}_{\ell+1,\ell'}(W)] = -\delta^{(\ell+1)}$*

This theorem is proven in appendix A.1 and provides the foundation for our test: when $\ell \geq m$, the expected difference between estimators is zero; when $\ell < m$, we expect a non-zero difference proportional to the unaccounted carryover effect $\delta^{(\ell+1)}$. Appendix A.2 generalizes it to $\mathbb{E}[\hat{\tau}_{\ell,\ell+J}(W) - \hat{\tau}_{\ell+J}(W)] = \sum_{j=1}^{J} -\delta^{(\ell+j)}$.

3.1 Algorithm

We propose a statistical test to determine if the assumed lag ℓ is insufficient:

Algorithm 1 Test if ℓ upper bounds m

```
procedure TestUpperBound(W, ℓ, α)
    ℓ' = ℓ + 1
    Calculate τ̂_{ℓ,ℓ'}(W) and τ̂_{ℓ'}(W)
    δ ← τ̂_{ℓ,ℓ'}(W) − τ̂_{ℓ'}(W)
    Calculate standard error SE(δ)
    t ← δ/SE(δ)
    p-value ← 2 · (1 − Φ(|t|))                    ▷ Two-tailed test
    if p-value < α then
        reject null hypothesis of no difference
    else
        fail to reject null hypothesis
    end if
end procedure
```

This test serves as a diagnostic tool for switchback experiments. While it cannot confirm that ℓ is a tight upper bound on m, it can reliably detect when ℓ is insufficient. In practice, we recommend applying this test as a validation check for any switchback experiment. The standard error can be calculated through bootstrapping the errors. When the test rejects the null hypothesis, experimenters should be warned that their results may be biased due to unaccounted carryover effects.

The power of this test depends on the magnitude of $\delta^{(\ell+1)}$ relative to the variance of the estimator difference. When carryover effects decay gradually, the test becomes more sensitive as the unaccounted effects grow larger.

4 Simulation and empirical verification

Since ground truth is unobservable outside of simulated experiments, we use simulation to evaluate algorithm 1 by introducing controlled effects on data sampled from a retail store. We randomly select 50,000 unique products that were sold on a particular day and track those products over 14 subsequent days.

Using the data generating mechanism in eq. (3), we simulate scenarios with direct effect $\delta^{(0)} = 0.2$ and two carryover structures on top of this sales data:

- **Cumulative effects**: $\delta^{(j)} = 0.2$ for all $1 \leq j \leq m$ (all lags up to m have equal effect)
- **Specific lag effects**: $\delta^{(m)} = 0.2$ and $\delta^{(j)} = 0$ for $j > 0, j \neq m$ (only lag m has an effect)

We vary the maximum carryover lag m between 1-4 and test with assumed upper bounds $\ell \in \{1, 2\}$. For each configuration, we perform 500 iterations with randomly sampled treatment assignments $W \in \{0, 1\}^{50,000 \times 14}$. Table 1 shows the reject rates of our test.

Table 1: Simulation results showing reject rates and false positive rates across different carryover effect configurations. Each row represents results from 500 randomizations.

ℓ	**Effect type**	$\delta^{(0)}$	$\delta^{(1)}$	$\delta^{(2)}$	$\delta^{(3)}$	$\delta^{(4)}$	**Reject rate**
1	None	0.0	0.0	0.0	0.0	0.0	0.05
1	Direct only	0.2	0.0	0.0	0.0	0.0	0.04
1	Cumulative	0.2	0.2	0.0	0.0	0.0	0.03
1	Cumulative	0.2	0.2	0.2	0.0	0.0	0.75
1	Cumulative	0.2	0.2	0.2	0.2	0.0	0.86
1	Cumulative	0.2	0.2	0.2	0.2	0.2	0.82
1	Specific lag	0.2	0.0	0.0	0.0	0.2	0.02
2	None	0.0	0.0	0.0	0.0	0.0	0.04
2	Cumulative	0.2	0.2	0.2	0.0	0.0	0.02
2	Cumulative	0.2	0.2	0.2	0.2	0.0	0.47
2	Cumulative	0.2	0.2	0.2	0.2	0.2	0.54

The results confirm our theoretical predictions. When $\ell \geq m$ (i.e., the assumed upper bound is correct), our test maintains appropriate false positive rates around 0.05. When $\ell < m$, the test successfully detects bias with high probability (0.47-0.86) for cumulative effects. For specific lag effects, detection power depends on the pattern of carryover effects, with distant isolated effects being harder to detect. These findings demonstrate that our test effectively identifies underestimated carryover durations in practical settings, particularly when effects are substantial and not isolated at distant lags.

5 Conclusion

We address a core challenge in switchback experimentation, namely detecting bias from underestimated carryover effects. We provide (i) theoretical foundations showing when causal effect estimators differ, (ii) a statistical test that flags insufficient lag without extra experiments or full prior knowledge, and (iii) simulation evidence on real-world data confirming strong detection across effect structures.

Our test offers practitioners a safeguard against erroneous conclusions, with sensitivity influenced by effect size and pattern. Future work should extend our results to more general effects models and quantify the sensitivity in terms of ℓ, N, T, and m, enabling optimal tradeoffs for experimental power and robustness to carryovers.

References

[1] Edward J Mills, An-Wen Chan, Ping Wu, Andy Vail, Gordon H Guyatt, and Douglas G Altman. Design, analysis, and presentation of crossover trials. *Trials*, 10(1):27, 2009.

[2] Paul Missault, Lorenzo Masoero, Christian Delbé, Thomas Richardson, and Guido Imbens. Robust and efficient multiple-unit switchback experimentation, 2025.

[3] Yuchen Hu and Stefan Wager. Switchback experiments under geometric mixing. *arXiv preprint arXiv:2209.00197*, 2022.

[4] Zikun Ye, Zhiqi Zhang, Dennis Zhang, Heng Zhang, and Renyu Philip Zhang. Deep-learning-based causal inference for large-scale combinatorial experiments: Theory and empirical evidence. *Available at SSRN 4375327*, 2023.

[5] Joe Cooprider and Shima Nassiri. Science of price experimentation at amazon. In *AEA 2023, NABE 2023*, 2023.

[6] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777, 2023.

[7] Betsy Sinclair, Margaret McConnell, and Donald P Green. Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56(4):1055–1069, 2012.

[8] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1):20150021, 2017.

[9] S Gwynn Sturdevant and Thomas Lumley. Statistical methods for testing carryover effects: A mixed effects model approach. *Contemporary Clinical Trials Communications*, 22:100711, 2021.

[10] Su Jia, Nathan Kallus, and Christina Lee Yu. Clustered switchback experiments: Near-optimal rates under spatiotemporal interference. *arXiv preprint arXiv:2312.15574*, 2023.

[11] Ruoxuan Xiong, Alex Chin, and Sean Taylor. Bias-variance tradeoffs for designing simultaneous temporal experiments. In *The KDD'23 Workshop on Causal Discovery, Prediction and Decision*, pages 115–131. PMLR, 2023.

[12] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[13] Alan S Gerber and Donald P Green. Field experiments: Design, analysis, and interpretation. *(No Title)*, 2012.

[14] A Hedayat and K Afsarinejad. Repeated measurements designs, ii. *The Annals of Statistics*, pages 619–628, 1978.

A proofs

A.1 Proof of Theorem 1

We first introduce the below lemma:

Lemma 1 (difference of misspecified estimators)**.** *For any $\ell' \geq \ell + 1$, under fair-coin randomization with $P_{\ell+1} = \frac{1}{2}P_\ell$,*

$$\hat{\tau}_{\ell,\ell'}(W) - \hat{\tau}_{\ell+1,\ell'}(W) = \frac{1}{P_\ell}\frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell) + \mathbb{O}_{i,t}(\ell)\right)Y_{i,t}(W).$$

Proof. Fix any $\ell' \geq \ell + 1$. We begin by writing out the difference of offset estimators:

$$\begin{aligned}
\hat{\tau}_{\ell,\ell'}(W) - \hat{\tau}_{\ell+1,\ell'}(W) &= \frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}\left[\frac{\mathbb{I}_{i,t}(\ell)}{P_\ell} - \frac{\mathbb{O}_{i,t}(\ell)}{P_\ell}\right]Y_{i,t}(W) \\
&\quad - \frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}\left[\frac{\mathbb{I}_{i,t}(\ell+1)}{P_{\ell+1}} - \frac{\mathbb{O}_{i,t}(\ell+1)}{P_{\ell+1}}\right]Y_{i,t}(W) \\
&= \frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}\left(\frac{\mathbb{I}_{i,t}(\ell)}{P_\ell} - \frac{\mathbb{O}_{i,t}(\ell)}{P_\ell}\right) - \left(\frac{\mathbb{I}_{i,t}(\ell+1)}{P_{\ell+1}} - \frac{\mathbb{O}_{i,t}(\ell+1)}{P_{\ell+1}}\right)Y_{i,t}(W)
\end{aligned}$$

Now, we simplify using $P_{\ell+1} = P_\ell/2$.

$$= \frac{1}{P_\ell}\frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}\left(\mathbb{I}_{i,t}(\ell) - \mathbb{O}_{i,t}(\ell)\right) - 2\left(\mathbb{I}_{i,t}(\ell+1) - \mathbb{O}_{i,t}(\ell+1)\right)Y_{i,t}(W)$$

We use the identities $\mathbb{I}_{i,t}(\ell+1) = \mathbb{I}_{i,t}(\ell)I(W_{i,t-\ell-1} = 1)$ and $\mathbb{O}_{i,t}(\ell+1) = \mathbb{O}_{i,t}(\ell)I(W_{i,t-\ell-1} = 0)$. Substituting these gives:

$$\begin{aligned}
&\left(\mathbb{I}_{i,t}(\ell) - \mathbb{O}_{i,t}(\ell)\right) - 2\left(\mathbb{I}_{i,t}(\ell+1) - \mathbb{O}_{i,t}(\ell+1)\right) \\
&= \left(\mathbb{I}_{i,t}(\ell) - \mathbb{O}_{i,t}(\ell)\right) - 2\left(\mathbb{I}_{i,t}(\ell)I(W_{i,t-\ell-1} = 1) - \mathbb{O}_{i,t}(\ell)I(W_{i,t-\ell-1} = 0)\right) \\
&= \mathbb{I}_{i,t}(\ell)\left(1 - 2I(W_{i,t-\ell-1} = 1)\right) - \mathbb{O}_{i,t}(\ell)\left(1 - 2I(W_{i,t-\ell-1} = 0)\right) \\
&= \mathbb{I}_{i,t}(\ell)\left(I(W_{i,t-\ell-1} = 0) - I(W_{i,t-\ell-1} = 1)\right) + \mathbb{O}_{i,t}(\ell)\left(I(W_{i,t-\ell-1} = 0) - I(W_{i,t-\ell-1} = 1)\right) \\
&= \left(\mathbb{I}_{i,t}(\ell) + \mathbb{O}_{i,t}(\ell)\right)\left(I(W_{i,t-\ell-1} = 0) - I(W_{i,t-\ell-1} = 1)\right) \\
&= (-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell) + \mathbb{O}_{i,t}(\ell)\right)
\end{aligned}$$

Putting it all together, the difference of the total causal effect estimators is:

$$\hat{\tau}_{\ell,\ell'}(W) - \hat{\tau}_{\ell+1,\ell'}(W) = \frac{1}{P_\ell}\frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell) + \mathbb{O}_{i,t}(\ell)\right)Y_{i,t}(W)$$

□

This allows us to prove theorem 1 as follows:

Proof. From Lemma 1,

$$\hat{\tau}_{\ell,\ell'}(W) - \hat{\tau}_{\ell+1,\ell'}(W) = \frac{1}{P_\ell}\frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell) + \mathbb{O}_{i,t}(\ell)\right)Y_{i,t}(W)$$

Expectation $E[\frac{1}{P_\ell}\frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell) + \mathbb{O}_{i,t}(\ell)\right)Y_{i,t}(W)]$

Substituting the linear effects model:

$$E\left[\frac{1}{P_\ell}\frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell)+\mathbb{O}_{i,t}(\ell)\right)Y_{i,t}(W)\right]$$
$$=\frac{1}{P_\ell}\frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}E\left[(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell)+\mathbb{O}_{i,t}(\ell)\right)\left(\mu_i+\alpha_{i,t}+\sum_{j=0}^{\min(t-1,m)}\delta^{(j)}W_{i,t-j}+\epsilon_t\right)\right] \tag{5}$$

The terms μ_i, $\alpha_{i,\ell+1}$, and $\epsilon_{\ell+1}$ have expected value 0 when multiplied by the balanced difference $(\mathbb{I}_{i,t}(\ell)-\mathbb{O}_{i,t}(\ell))$ due to the symmetry of randomization. Consider e.g. for the term containing μ_i:

$$E\left[(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell)+\mathbb{O}_{i,t}(\ell)\right)\mu_i\right]=\mu_i E\left[(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell)+\mathbb{O}_{i,t}(\ell)\right)\right]=\mu_i\cdot 0$$

Therefore, the terms μ_i, $\alpha_{i,t}$, and ϵ_t contribute 0 to the expectation, hence we only compute:

$$\frac{1}{P_\ell}\frac{1}{N(T-\ell')}E\left[(-1)^{W_{i,t-\ell-1}}\left(\mathbb{I}_{i,t}(\ell)+\mathbb{O}_{i,t}(\ell)\right)\sum_{j=0}^{\min(t-1,m)}\delta^{(j)}W_{i,t-j}\right] \tag{6}$$

We analyze this by cases based on the value of j:

Case 2.1: $j\le\ell$

For $j\le\ell$, $W_{i,t-j}$ is part of the sequence $W_{i,t-\ell:t}$:

$$E\left[(-1)^{W_{i,t-\ell-1}}(\mathbb{I}_{i,t}(\ell)+\mathbb{O}_{i,t}(\ell))W_{i,t-j}\delta^{(j)}\right]=\frac{1}{2^{\ell+2}}\cdot\delta^{(j)}+\frac{1}{2^{\ell+2}}\cdot(-\delta^{(j)})+0+0=0 \tag{7}$$

Case 2.2: $j=\ell+1$

Here $W_{i,t-j}=W_{i,t-\ell-1}$ is independent of $W_{i,t-\ell:t}$ and determines the sign:

$$E\left[(-1)^{W_{i,t-\ell-1}}(\mathbb{I}_{i,t}(\ell)+\mathbb{O}_{i,t}(\ell))W_{i,t-j}\delta^{(\ell+1)}\right]=-2\frac{\delta^{(\ell+1)}}{2^{\ell+2}}=-\frac{\delta^{(\ell+1)}}{2^{\ell+1}} \tag{8}$$

Case 2.3: $j>\ell+1$

For $j>\ell+1$, $W_{i,t-j}$, $W_{i,t-\ell-1}$, and $W_{i,t-\ell:t}$ are all independent under randomization:

$$E\left[(-1)^{W_{i,t-\ell-1}}(\mathbb{I}_{i,t}(\ell)+\mathbb{O}_{i,t}(\ell))W_{i,t-j}\delta^{(j)}\right]=0 \tag{9}$$

Combining cases

$$\begin{aligned}E[\hat{\tau}_{\ell,\ell'}(W)-\hat{\tau}_{\ell+1,\ell'}(W)]&=\frac{1}{P_\ell}\frac{1}{N(T-\ell')}\sum_{i=1}^{N}\sum_{t=\ell'+1}^{T}\left(-\frac{\delta^{(\ell+1)}}{2^{\ell+1}}\right)\\&=\frac{1}{P_\ell}\frac{1}{T-\ell'}\sum_{t=\ell'+1}^{T}\left(-\frac{\delta^{(\ell+1)}}{2^{\ell+1}}\right)\\&=\frac{2^{\ell+1}}{T-\ell'}\cdot\frac{1}{2^{\ell+1}}\sum_{t=\ell'+1}^{T}\left(-\delta^{(\ell+1)}\right)\\&=-\delta^{(\ell+1)}\end{aligned} \tag{10}$$

Therefore,

$$\mathbb{E}[\hat{\tau}_{\ell,\ell'}(W)-\hat{\tau}_{\ell+1,\ell'}(W)]=-\delta^{(\ell+1)}.$$

□

A.2 Proof of Generalized Theorem 1

Proof by Induction. We prove by mathematical induction.

Base Case ($J = 1$): For any $\ell' \geq \ell + 1$ we note $\mathbb{E}[\hat{\tau}_{\ell,\ell'}(W) - \hat{\tau}_{\ell+1,\ell'}(W)] = -\delta^{(\ell+1)}$ is proven in theorem 1.

Inductive Step: Assume the statement holds for some positive integer J and any $\ell' \geq \ell + J$, i.e.,

$$\mathbb{E}[\hat{\tau}_{\ell,\ell'}(W) - \hat{\tau}_{\ell+J,\ell'}(W)] = \sum_{j=1}^{J} -\delta^{(\ell+j)}.$$

We need to prove it holds for $J + 1$ and any $\ell'' \geq \ell + J + 1$:

$$\mathbb{E}[\hat{\tau}_{\ell,\ell''}(W) - \hat{\tau}_{\ell+J+1,\ell''}(W)] = \sum_{j=1}^{J+1} -\delta^{(\ell+j)}.$$

We rewrite the left-hand side:

$$\begin{aligned}\mathbb{E}[\hat{\tau}_{\ell,\ell''}(W) - \hat{\tau}_{\ell+J+1,\ell''}(W)] &= \mathbb{E}[\hat{\tau}_{\ell,\ell''}(W) \boldsymbol{- \hat{\tau}_{\ell+J,\ell''}(W) + \hat{\tau}_{\ell+J,\ell''}(W)} - \hat{\tau}_{\ell+J+1,\ell''}(W)] \\ &= \mathbb{E}[\hat{\tau}_{\ell,\ell''}(W) - \hat{\tau}_{\ell+J,\ell''}(W)] + \mathbb{E}[\hat{\tau}_{\ell+J,\ell''}(W) - \hat{\tau}_{\ell+J+1,\ell''}(W)].\end{aligned}$$

For the first term, since $\ell'' > \ell'$ by the inductive step, we have:

$$\mathbb{E}[\hat{\tau}_{\ell,\ell''}(W) - \hat{\tau}_{\ell+J,\ell''}(W)] = \sum_{j=1}^{J} -\delta^{(\ell+j)}.$$

For the second term, the difference are 1 apart, and thus theorem 1 applies:

$$\mathbb{E}[\hat{\tau}_{\ell+J,\ell''}(W) - \hat{\tau}_{\ell+J+1,\ell''}(W)] = -\delta^{(\ell+J+1)}.$$

Therefore:

$$\begin{aligned}\mathbb{E}[\hat{\tau}_{\ell,\ell''}(W) - \hat{\tau}_{\ell+J+1,\ell''}(W)] &= \sum_{j=1}^{J} -\delta^{(\ell+j)} + (-\delta^{(\ell+J+1)}) \\ &= \sum_{j=1}^{J+1} -\delta^{(\ell+j)}.\end{aligned}$$

This completes the inductive step and proves the result. □