Electronic Companion to "A Graphical Point Process Framework for Understanding Removal Effects in Multi-Touch Attribution"

In this Electronic Companion, we first list the important notations in Table EC.1. Then, we present the details of the proposed ADMM algorithm in Section EC.1, the assumptions for consistency in Section EC.2, and the proofs of Theorem 3, Theorem 4, and Theorem 5 in Sections EC.3-EC.5.

Notation	Type	Description
p	Integer	The number of event types (conversion and touchpoints) in the attribution problem.
$\mathcal{E}$	Set of labels	$\{1,\dots,p\}$, the set of all event types, used as the node set for the Granger causality graph.
e	Label in $\mathcal{E}$	An event type, standing for a node on the Granger causality graph.
i	Integer	The index of an event on a path.
t_i	Number in $\mathbb{R}_{\geq 0}$	The relative timestamp of the i-th event on a given path.
e_i	Label in $\mathcal{E}$	The event label of the i-th event on a given path.
(t_i, e_i)	Element in $\mathbb{R}_{\geq 0}\times\mathcal{E}$	A detailed representation of an event.
$i^\star$	Integer	The index of a conversion event on a path. $(t_{i^\star}, e_{i^\star})$ is the conversion event.
m	Integer	The number of events on a path.
D	Set of events	$\{(t_i,e_i)\}_{i=1}^m$, a detailed representation of a path.
T	Number in $\mathbb{R}_{>0}$	The time length of observation or the upper bound of t_i.
n	Integer	The number of observed paths, i.e. sample size.
j	Integer	The index of a path among all the observations. When emphasized, the notations in the above rows are adapted correspondingly as in $D_j=\{(t_i^j,e_i^j)\}_{i=1}^{m_j}$ with length T_j.
$N_e(t)$	Counting function $\mathbb{R}_{\geq 0}\to\mathbb{N}$	$\sum_{i:e_i=e}\mathbb{1}_{\{t_i\leq t\}}$, the point process of event type e.
$\mathbf{N}(t)$	Vector of counting functions	$(N_1(t),\dots,N_p(t))^\top$, the multivariate point process of event type $1,\dots,p$.
$\mathbf{N}_W(t)$	Vector of counting functions	$(N_e(t))_{e\in W}$ the subprocess of $\mathbf{N}(t)$ for event label subset $W\subseteq\mathcal{E}$.
$\mathbf{N}_W^D(t)$	Vector of counting functions	$(N_e^D(t))_{e\in W}$ with $N_e^D(t)=\sum_{(t_i,e_i)\in D:e_i=e}\mathbb{1}_{\{t_i\leq t\}}$, the observed subprocess for path D.
$\lambda_e(t\,\vert\,\mathcal{H}_t)$	Non-negative function	The conditional intensity function for event type e.
$(e'\to e)$	Element in $\mathcal{E}\times\mathcal{E}$	The directed edge from node e' to node e, representing Granger causality.
$\mathcal{A}$	Set of edges	The set of all edges where the Granger causality relations exist.
q	Integer	$1\leq q<p$, the number of customer-initiated event types.
$\mathcal{E}_{\mathrm{c}}$	Set of labels	The set of customer-initiated event types, assumed to be $\{1,\dots,q\}$.
$\mathcal{E}_{\mathrm{f}}$	Set of labels	The set of firm-initiated event types, assumed to be $\{q+1,\dots,p\}$.
μ_e	Number in $\mathbb{R}_{\geq 0}$	The baseline intensity of event type e.
$\psi_{e'e}(\cdot)$	Non-negative function	The kernel/transfer function describing the shape of the Granger causality impact $(e'\to e)$.
$\alpha_{e'e}$	Number in $\mathbb{R}_{\geq 0}$	The coefficient describing the scale of the Granger causality impact $(e'\to e)$.
$\boldsymbol{\alpha}_e$	Vector in $\mathbb{R}_{\geq 0}^p$	$(\alpha_{1e},\dots,\alpha_{pe})^\top$, the vector of Granger causality coefficients for node e.
A	Matrix in $\mathbb{R}_{\geq 0}^{p\times q}$	$(\alpha_{e'e})_{e'\in\mathcal{E},e\in\mathcal{E}_{\mathrm{c}}}$, the Granger causality coefficient matrix.
R	Set of observed events	A touchpoint event subset of a converted path D.
$F_t^W(D)$	Set of observed events	$\{(t_i,e_i)\in D: t_i<t,\ e_i\in W\}$, a subset of D filtered by timestamp t and event type set W.
$\lambda_1(t_{i^\star}\,\vert\,\mathcal{H}_{t_{i^\star}}^D)$	Number in $\mathbb{R}_{\geq 0}$	The conditional intensity value of conversion $(e=1)$ at $t=t_{i^\star}$ given path D.
$\lambda_1(t_{i^\star}\,\vert\,\mathcal{H}_{t_{i^\star}}^{D\setminus R})$	Number in $\mathbb{R}_{\geq 0}$	The conditional intensity value of conversion at $t=t_{i^\star}$ given path $D\setminus R$.
$R^\diamond$	Random set of observed events	A superset of R obtained from thinning. The set of removed events in R and additional events possibly to lose.
γ_e	Number in $\mathbb{R}_{\geq 0}$	The regularization parameters to control the sparsity of $\boldsymbol{\alpha}_e$.
$\boldsymbol{\theta}$	Vector in $\mathbb{R}_{\geq 0}^{p+1}$	$(\mu_e,\alpha_{1e},\dots,\alpha_{pe})^\top$, the vector of parameters to learn for a given node e.
S	Set of indices	$\{0\}\cup\{e'\in\mathcal{E}:\alpha_{e'e}>0\}$, the active set of signals corresponding to nonzero values of $\boldsymbol{\theta}$.
S^{c}	Set of indices	$\{e'\in\mathcal{E}:\alpha_{e'e}=0\}$, the set of signals corresponding to zero values of $\boldsymbol{\theta}$.
s	Integer	The cardinality of the active set S.
r	Integer	The cardinality of the removal set R.
$\hat{\boldsymbol{\theta}}$	Vector in $\mathbb{R}_{\geq 0}^{p+1}$	The regularized estimator of $\boldsymbol{\theta}$.
z	Label in $\{1,\dots,Z\}$	The channel label for a touchpoint event type, where Z is the number of channel types.
$\mathcal{C}_z$	Set of labels	The set of touchpoint event types belonging to channel z.
p_z	Number in $[0,1]$	The proportion of channel-level conversion count (CCC) for channel z.
$\hat{\mathsf{p}}_z$	Number in $[0,1]$	The proportion of channel-level aggregated score (CAS) for channel z.
$\mathbf{p}$	Vector in $[0,1]^Z$	$(\mathsf{p}_1,\dots,\mathsf{p}_Z)^\top$, the vector of proportions of CCC.
$\hat{\mathbf{p}}$	Vector in $[0,1]^Z$	$(\hat{\mathsf{p}}_1,\dots,\hat{\mathsf{p}}_Z)^\top$, the vector of proportions of CAS.

Table EC.1 Summary of important notations.

EC.1. Details of the Proposed ADMM Algorithm

For each node $e \in \mathcal{E}_c$, learning its parent nodes yields

$$\min_{\mu_e \geq 0, \boldsymbol{\alpha}_e \in \mathbb{R}^p_{\geq 0}} \frac{1}{n} \sum_{j=1}^{n} \phi_e(D_j; \mu_e, \boldsymbol{\alpha}_e) + \gamma_e \|\boldsymbol{\alpha}_e\|_1, \tag{EC.1}$$

where

$$\phi_e(D_j; \mu_e, \boldsymbol{\alpha}_e) = \frac{1}{2} \int_0^{T_j} [\lambda_e(t \mid \mathcal{H}_t^{D_j})]^2 dt - \int_0^{T_j} \lambda_e(t \mid \mathcal{H}_t^{D_j}) dN_e^{D_j}(t).$$

In the following context, we write $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)^\top = (\mu_e, \boldsymbol{\alpha}_e^\top)^\top \in \mathbb{R}^{p+1}_{\geq 0}$. Let $\boldsymbol{X}_j(t) = (X_{j,0}(t), X_{j,1}(t), \ldots, X_{j,p}(t))^\top$ with

$$X_{j,0}(t) = 1, \quad X_{j,k}(t) = \int_0^t \psi_{ke}(t-u) dN_k^{D_j}(u), \ k = 1, \ldots, p.$$

Now the conditional intensity can be written as $\lambda_e(t \mid \mathcal{H}_t^{D_j}) = \boldsymbol{\theta}^\top \boldsymbol{X}_j(t)$. Let $V = (V_{kk'})_{k,k'=0}^p \in \mathbb{R}^{(p+1)\times(p+1)}$ and $\boldsymbol{b} = (b_0, \cdots, b_p)^\top \in \mathbb{R}^{p+1}$, where for $k = 0, \ldots, p$,

$$V_{kk'} = \frac{1}{n} \sum_{j=1}^{n} \int_0^{T_j} X_{j,k}(t) X_{j,k'}(t) dt, \quad b_k = \frac{1}{n} \sum_{j=1}^{n} \int_0^{T_j} X_{j,k}(t) dN_e^{D_j}(t). \tag{EC.2}$$

Then the regularized solution satisfies

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}_{\geq 0}} \frac{1}{2} \boldsymbol{\theta}^\top V \boldsymbol{\theta} - \boldsymbol{b}^\top \boldsymbol{\theta} + \gamma_e \|\boldsymbol{\alpha}_e\|_1. \tag{EC.3}$$

The above problem is equivalent to a linearly constrained one

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}_{\geq 0}, \boldsymbol{\alpha}_e = \boldsymbol{\alpha}'_e} \frac{1}{2} \boldsymbol{\theta}^\top V \boldsymbol{\theta} - \boldsymbol{b}^\top \boldsymbol{\theta} + \gamma_e \|\boldsymbol{\alpha}'_e\|_1.$$

For any p-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$, let $\|\boldsymbol{x}\|_2 := \sqrt{\sum_{k=1}^p x_k^2}$ denote its L_2-norm. We design an alternating direction method of multipliers (ADMM). The corresponding augmented Lagrangian function is

$$\mathcal{L}_\eta(\boldsymbol{\theta}, \boldsymbol{\alpha}'_e, \boldsymbol{\omega}) = \frac{1}{2} \boldsymbol{\theta}^\top V \boldsymbol{\theta} - \boldsymbol{b}^\top \boldsymbol{\theta} + \gamma_e \|\boldsymbol{\alpha}'_e\|_1 + \boldsymbol{\omega}^\top (\boldsymbol{\alpha}_e - \boldsymbol{\alpha}'_e) + \frac{1}{2} \eta \|\boldsymbol{\alpha}_e - \boldsymbol{\alpha}'_e\|_2^2,$$

where $\|\boldsymbol{\alpha}_e - \boldsymbol{\alpha}'_e\|_2^2 = (\boldsymbol{\alpha}_e - \boldsymbol{\alpha}'_e)^\top (\boldsymbol{\alpha}_e - \boldsymbol{\alpha}'_e)$ represents the squared L_2-distance between $\boldsymbol{\alpha}_e$ and $\boldsymbol{\alpha}'_e$. The learning algorithm is shown in Algorithm 3. It consists of two steps, pre-computation and optimization, corresponding to (EC.2) and (EC.3) respectively.

Algorithm 3 Graphical point process learning by ADMM

Input: Paths $D_1, \ldots, D_n$ and the regularization parameter γ_e.

Pre-compute: V, $\boldsymbol{b}$.

Initialize: Set $\eta > 0$ and proper initial values of μ_e, $\boldsymbol{\alpha}_e$, $\boldsymbol{\alpha}'_e$, and $\boldsymbol{\omega}$.

while not converge **do**

$$\begin{pmatrix} \mu_e \\ \boldsymbol{\alpha}_e \end{pmatrix} = \left[\left(V + \begin{pmatrix} 0 & \\ & \eta I_p \end{pmatrix} \right)^{-1} \left(\boldsymbol{b} + \begin{pmatrix} 0 \\ \eta \boldsymbol{\alpha}'_e - \boldsymbol{\omega} \end{pmatrix} \right) \right]_+ .$$

$\boldsymbol{\alpha}'_e = (\boldsymbol{\alpha}_e + \eta^{-1}\boldsymbol{\omega} - \eta^{-1}\gamma_e \mathbb{1}_p)_+$.

$\boldsymbol{\omega} = \boldsymbol{\omega} + \eta(\boldsymbol{\alpha}_e - \boldsymbol{\alpha}'_e)$.

end while

Return: $\boldsymbol{\theta} = (\mu_e, \boldsymbol{\alpha}_e^\top)^\top$.

For typical attribution use cases with sufficient observations, the main computational cost of Algorithm 3 is the pre-computation step, which requires traversing all the paths to collect the first-order and second-order statistics. Recall the assumption that the number of events m_j on each path D_j satisfies $m_j < \overline{m},\ j = 1, \ldots, n$. The time cost for this step is $O(\overline{m}^2 n)$ for each node.

According to Hong and Luo (2017), ADMM converges linearly in this problem. During each iteration, the computation cost depends on $(p+1)$-dimensional matrix operations. Then the complexity of the optimization step is $O(p^3)$ for each node.

DRE requires the neighborhood of conversion only with $p+1$ parameters and thus the learning complexity is $O(\overline{m}^2 n + p^3)$. As a comparison, TRE needs the information of the entire graph. There are q unique nodes to be learned, with a total of $(p+1)q$ parameters. Therefore, the overall graph learning complexity is $O(\overline{m}^2 nq + p^3 q)$. It is worth pointing out that these q nodes can be learned parallelly.

EC.2. Assumptions

In this section, we provide the assumptions for establishing the rates of convergence of the proposed estimators and consistency results.

Assumption EC.1. *(Independence) The process of the customer-initiated event types* $\mathbf{N}_{\mathcal{E}_c}^{D_j}(t)$, $j = 1, \ldots, n$ *are independent and follow model* (1).

This assumption basically says the prospective customers behave independently of the ads, which is generally true due to interactions between them being sparse. The convergence properties of the estimator rely on this independent distribution of the observations. We do not need to assume the external (firm-initiated) events are *I.I.D.* across paths.

ASSUMPTION EC.2. *(Boundedness) There exist constants* $\overline{T}$ *and* $\overline{X}$ *such that* $T_j < \overline{T}$ *and* $\sup_{t\in(0,T_j)} \left\|\boldsymbol{X}_j(t)\right\|_\infty < \overline{X}$ *a.s. for* $j = 1,\ldots,n$.

The first condition in this assumption requires a limit to the length of observation. Assuming $\psi_{ke}(\cdot),\ k = 1,\ldots,p$ are universally bounded, the second condition reduces to the overall frequency of each type of event is bounded.

For a matrix $M = (M_{ij})_{i,j}$, the matrix L_1-norm is defined as $\left\|M\right\|_{1,\infty} := \max_i(\sum_j \left|M_{ij}\right|)$. Let $G = (G_{kk'})_{k,k'=0}^p = \mathbb{E}V$ be the population version of V. Assume the sub-matrix $G_{SS} = (G_{kk'})_{k,k'\in S}$ is non-singular and define $\kappa := \left\|G_{SS}^{-1}\right\|_{1,\infty}$.

ASSUMPTION EC.3. *(Irrepresentability) There exists a constant* $\xi \in (0,1)$ *such that*

$$\left\|G_{S^cS}G_{SS}^{-1}\right\|_{1,\infty} < 1-\xi.$$

This condition is similar to Condition 3 for sparse survival models in Lin and Lv (2013). It is a generalization of the *irrepresentable* condition (Zhao and Yu 2006) that is almost necessary and sufficient for the LASSO to achieve the model selection consistency. For example, Wainwright (2009) used their Condition (15) to analyze the recovery of sparsity pattern using the LASSO.

Below we repeat Theorem 3 with constants given in detail.

THEOREM 3. *Under Assumptions EC.1-EC.3, there exist* $C_3 > 0$ *and* $C_4 > 0$ *such that the regularized estimator* $\hat{\boldsymbol{\theta}}$ *satisfies the following properties:*

(i) If p *is fixed, then for any constant* $0 < \nu < 1$*, when*

$$\begin{cases} n > \max(C_3^{-1}, 8\xi^{-2}\kappa^2\overline{X}^4\overline{T}^2 s^2)^{\frac{1}{1-\nu}} \\ \gamma_e = 4\xi^{-1}C_3^{-\frac{1}{2}}C_4 n^{-\frac{1-\nu}{2}} \end{cases},$$

with probability at least $1-2(p+1)(p+2)\exp(-n^\nu)$,

- *(Edge selection)* $\hat{\boldsymbol{\theta}}_{S^c} = 0$
- *(*L_∞*-error)* $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_\infty \le 10\xi^{-1}\kappa C_3^{-\frac{1}{2}}C_4 n^{-\frac{1-\nu}{2}}$;

(ii) If p *diverges, then for any constant* $\zeta > 2$*, when*

$$\begin{cases} n > \max(C_3^{-1}, 8\xi^{-2}\kappa^2\overline{X}^4\overline{T}^2 s^2)\cdot\zeta\log(p+1) \\ \gamma_e = 4\xi^{-1}C_4\sqrt{\frac{\zeta\log(p+1)}{C_3 n}} \end{cases},$$

with probability at least $1-3/(p+1)^{\zeta-2}$,

- *(Edge selection)* $\hat{\boldsymbol{\theta}}_{S^c} = 0$
- *(*L_∞*-error)* $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_\infty \le 10\xi^{-1}\kappa C_3^{-\frac{1}{2}}C_4\sqrt{\frac{\zeta\log(p+1)}{n}}$.

EC.3. Proof of Theorem 3

In this proof, let $\boldsymbol{\theta}^*$ denote the ground truth of $\boldsymbol{\theta}$ to distinguish from the variable $\boldsymbol{\theta}$ in the optimization problem. In the sequel, we first prove part (i) and then prove part (ii) of Theorem 3.

EC.3.1. Proof of (i) of Theorem 3.

Proof of Part (i) of Theorem 3. Let $U(\boldsymbol{\theta})$ be the negative derivative of the loss function, that is

$$U(\boldsymbol{\theta}) = \boldsymbol{b} - V\boldsymbol{\theta} = \frac{1}{n}\sum_{j=1}^{n}\int_0^{T_j} \boldsymbol{X}_j(t)\left(dN_e^{D_j}(t) - \boldsymbol{\theta}^\top \boldsymbol{X}_j(t)dt\right).$$

For any K-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_K)^\top \in \mathbb{R}^K$, where $K \geq 1$ is an arbitrary integer, let $\text{sign}(\boldsymbol{x})$ denote the vector $(\text{sign}(x_1), \ldots, \text{sign}(x_K))^\top$. Then the vector $\hat{\boldsymbol{\theta}} = (\hat{\mu}_e, \hat{\boldsymbol{\alpha}}_e) \in \mathbb{R}^{p+1}_{\geq 0}$ is a strict local minimizer of problem (EC.3), if the following conditions hold:

$$\begin{cases} U_{\hat{S}}(\hat{\boldsymbol{\theta}}) - \gamma_e \boldsymbol{\delta}_{\hat{S}} \circ \text{sign}(\hat{\boldsymbol{\theta}}_{\hat{S}}) = 0 \\ \left\|U_{\hat{S}^c}(\hat{\boldsymbol{\theta}})\right\|_\infty < \gamma_e \end{cases}, \tag{EC.4}$$

where $\boldsymbol{\delta} = (0, \mathbf{1}_p^\top)^\top$ and $\circ$ denotes the Hadamard product.

We study the oracle solution:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}_{\geq 0}, \boldsymbol{\theta}_{S^c} = 0}{\arg\min} \frac{1}{2}\boldsymbol{\theta}^\top V \boldsymbol{\theta} - \boldsymbol{b}^\top \boldsymbol{\theta} + \gamma_e \left\|\boldsymbol{\theta}_{S \setminus \{0\}}\right\|_1. \tag{EC.5}$$

Note that $\tilde{\boldsymbol{\theta}}$ will be a solution to the original regularized problem (EC.3) as long as $\left\|U_{S^c}(\tilde{\boldsymbol{\theta}})\right\|_\infty < \gamma_e$ is satisfied.

For $\tilde{\boldsymbol{\theta}}$ and the true $\boldsymbol{\theta}^*$, we have $U_S(\tilde{\boldsymbol{\theta}}) = \boldsymbol{b}_S - V_{SS}\tilde{\boldsymbol{\theta}}_S$ and $U_S(\boldsymbol{\theta}^*) = \boldsymbol{b}_S - V_{SS}\boldsymbol{\theta}^*_S$, and thus

$$U_S(\tilde{\boldsymbol{\theta}}) - U_S(\boldsymbol{\theta}^*) = -V_{SS}(\tilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}^*_S).$$

Then by (EC.4),

$$\tilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}^*_S = V_{SS}^{-1}[U_S(\boldsymbol{\theta}^*) - \gamma_e \boldsymbol{\delta}_S \circ \text{sign}(\tilde{\boldsymbol{\theta}}_S)]. \tag{EC.6}$$

Need to guarantee V_{SS} is strictly positive definite and thus invertible, which requires that the least eigenvalue of V_{SS}, denoted by $\Lambda_{\min}(V_{SS})$, is positive. Since $0 \leq \int_0^{T_j} X_{j,k}(t)X_{j,k'}(t)dt \leq \overline{X}^2\overline{T}$ is a bounded random variable, by Hoeffding's inequality, for any $\epsilon_1 > 0$,

$$\mathbb{P}(\left|V_{kk'} - G_{kk'}\right| > \epsilon_1) \leq 2\exp(-\frac{2n\epsilon_1^2}{\overline{X}^4\overline{T}^2}).$$

Now, we use the union bound to obtain that

$$\mathbb{P}(\left\|V - G\right\|_\infty > \epsilon_1) \leq 2(p+1)^2\exp(-\frac{2n\epsilon_1^2}{\overline{X}^4\overline{T}^2}).$$

Let $\epsilon_1 = \frac{1}{\sqrt{2}}\overline{X}^2\overline{T}n^{-\frac{1-\nu}{2}}$, then with probability at least $1 - 2(p+1)^2\exp(-n^\nu)$, where $0 < \nu < 1$ is a constant,

$$\left\|V - G\right\|_\infty \leq \frac{1}{\sqrt{2}}\overline{X}^2\overline{T}n^{-\frac{1-\nu}{2}}.$$

We let $\epsilon_1 < \frac{\xi}{4s\kappa}$, which suggests the choice of n:

$$n > (8\xi^{-2}\kappa^2\overline{X}^4\overline{T}^2 s^2)^{\frac{1}{1-\nu}}. \tag{EC.7}$$

Let $\|\cdot\|_{\text{op}}$ denote matrix operator norm. Given the fact that

$$\Lambda_{\min}(G_{SS})^{-1} = \left\|G_{SS}^{-1}\right\|_{\text{op}} \le \left\|G_{SS}^{-1}\right\|_{1,\infty} = \kappa,$$

$$\left\|V_{SS} - G_{SS}\right\|_{\text{op}} \le \left\|V_{SS} - G_{SS}\right\|_{1,\infty} \le s\epsilon_1 < \frac{\xi}{4\kappa},$$

we have

$$\Lambda_{\min}(V_{SS}) \ge \Lambda_{\min}(G_{SS}) - \left\|V_{SS} - G_{SS}\right\|_{\text{op}} \ge (1 - \frac{\xi}{4})\kappa^{-1} > 0.$$

Then we give a bound of $\left\|U(\boldsymbol{\theta}^*)\right\|_\infty$, the gradient of the loss function at $\boldsymbol{\theta}^*$. By the definition of $U(\boldsymbol{\theta})$,

$$U(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{j=1}^{n}\int_0^{T_j} \boldsymbol{X}_j(t)\left(dN_e^{D_j}(t) - (\boldsymbol{\theta}^*)^\top \boldsymbol{X}_j(t)dt\right).$$

Note that $N_e^{D_j}(t) - \int_0^t (\boldsymbol{\theta}^*)^\top \boldsymbol{X}_j(u)du$ is a counting process martingale with respect to $\mathcal{H}_t$. On the other hand, the component

$$u_{j,k}(\boldsymbol{\theta}^*) := \int_0^{T_j} X_{j,k}(t)\left(dN_e^{D_j}(t) - (\boldsymbol{\theta}^*)^\top \boldsymbol{X}_j(t)dt\right)$$

for $k = 0, \ldots, p$ is a martingale as well. By Theorem 3 in Hansen et al. (2015), the random variable $u_{j,k}(\boldsymbol{\theta}^*)$ is sub-exponential. For any $\epsilon_2 > 0$,

$$\mathbb{P}(\left|u_{j,k}(\boldsymbol{\theta}^*)\right| \ge \sqrt{2\left\|\theta^*\right\|_1 \overline{X}^2\overline{T}\epsilon_2} + \frac{1}{3}\epsilon_2) \le 2\exp\left(-\frac{\epsilon_2}{\overline{X}}\right).$$

By $U_k(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{j=1}^n u_{j,k}(\boldsymbol{\theta}^*)$ and Bernstein's inequality (Vershynin 2018), there exist constant $C_3 > 0$ and $C_4 > 0$ depending on $\overline{X}$ and $\overline{T}$ such that

$$\mathbb{P}(\left|U_k(\boldsymbol{\theta}^*)\right| \ge \epsilon_2) \le 2\exp\left(-C_3 n \min(\frac{\epsilon_2^2}{C_4^2}, \frac{\epsilon_2}{C_4})\right).$$

And thus

$$\mathbb{P}(\left\|U(\boldsymbol{\theta}^*)\right\|_\infty \ge \epsilon_2) \le 2(p+1)\exp\left(-C_3 n \min(\frac{\epsilon_2^2}{C_4^2}, \frac{\epsilon_2}{C_4})\right).$$

When $\epsilon_2 < C_4$, with probability at least $1 - 2(p+1)\exp(-n^\nu)$,

$$\left\|U(\boldsymbol{\theta}^*)\right\|_\infty \le C_3^{-\frac{1}{2}} C_4 n^{-\frac{1-\nu}{2}}. \tag{EC.8}$$

Now we check the strict dual feasibility condition (EC.4). According to (EC.6),

$$\begin{aligned} U_{S^c}(\tilde{\boldsymbol{\theta}}) =& U_{S^c}(\boldsymbol{\theta}^*) - V_{S^cS}(\tilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^*) \\ =& U_{S^c}(\boldsymbol{\theta}^*) - V_{S^cS}V_{SS}^{-1}[U_S(\boldsymbol{\theta}^*) - \gamma_e \boldsymbol{\delta}_S \circ \text{sign}(\tilde{\boldsymbol{\theta}}_S)]. \end{aligned}$$

If $\left\|U(\boldsymbol{\theta}^*)\right\|_\infty \le \epsilon_2$, then

$$\begin{aligned}\left\|U_{S^c}(\tilde{\boldsymbol{\theta}})\right\|_\infty &\le \left\|U_{S^c}(\boldsymbol{\theta}^*)\right\|_\infty + \left\|V_{S^cS}V_{SS}^{-1}\right\|_{1,\infty}\left\|U_S(\boldsymbol{\theta}^*) - \gamma_e\boldsymbol{\delta}_S \circ \operatorname{sign}(\tilde{\boldsymbol{\theta}}_S)\right\|_\infty \\ &\le \left\|U_{S^c}(\boldsymbol{\theta}^*)\right\|_\infty + \left\|V_{S^cS}V_{SS}^{-1}\right\|_{1,\infty}(\left\|U_S(\boldsymbol{\theta}^*)\right\|_\infty + \gamma_e) \\ &\le \epsilon_2 + \left\|V_{S^cS}V_{SS}^{-1}\right\|_{1,\infty}(\epsilon_2 + \gamma_e).\end{aligned}$$

Let $\epsilon_2 \le \xi\gamma_e/4 < C_4$, which suggests the choice of n and γ_e by (EC.8):

$$\begin{cases} n > C_3^{-\frac{1}{1-\nu}} \\ \gamma_e = 4\xi^{-1}C_3^{-\frac{1}{2}}C_4 n^{-\frac{1-\nu}{2}} \end{cases}. \tag{EC.9}$$

Recall the choice of n in (EC.7), which implies $\kappa s\epsilon_1 < \xi/4 < 1/2$. Then

$$\begin{aligned}\left\|V_{SS}^{-1} - G_{SS}^{-1}\right\|_{1,\infty} &= \left\|G_{SS}^{-1}(G_{SS} - V_{SS})V_{SS}^{-1}\right\|_{1,\infty} \\ &\le \left\|G_{SS}^{-1}\right\|_{1,\infty}\left\|G_{SS} - V_{SS}\right\|_{1,\infty}\left\|V_{SS}^{-1}\right\|_{1,\infty} \\ &\le \left\|G_{SS}^{-1}\right\|_{1,\infty}\left\|G_{SS} - V_{SS}\right\|_{1,\infty}(\left\|G_{SS}^{-1}\right\|_{1,\infty} + \left\|V_{SS}^{-1} - G_{SS}^{-1}\right\|_{1,\infty}) \\ &\le \frac{\left\|G_{SS}^{-1}\right\|_{1,\infty}^2\left\|G_{SS} - V_{SS}\right\|_{1,\infty}}{1 - \left\|G_{SS}^{-1}\right\|_{1,\infty}\left\|G_{SS} - V_{SS}\right\|_{1,\infty}} \\ &\le \frac{\kappa^2 s\epsilon_1}{1 - \kappa s\epsilon_1} \\ &< \kappa.\end{aligned}$$

Therefore,

$$\left\|V_{SS}^{-1}\right\|_{1,\infty} \le \left\|G_{SS}^{-1}\right\|_{1,\infty} + \left\|V_{SS}^{-1} - G_{SS}^{-1}\right\|_{1,\infty} < 2\kappa. \tag{EC.10}$$

As a result,

$$\begin{aligned}&\left\|V_{S^cS}V_{SS}^{-1} - G_{S^cS}G_{SS}^{-1}\right\|_{1,\infty} \\ \le&\left\|(V_{S^cS} - G_{S^cS})V_{SS}^{-1}\right\|_{1,\infty} + \left\|G_{S^cS}G_{SS}^{-1}(V_{SS} - G_{SS})V_{SS}^{-1}\right\|_{1,\infty} \\ \le&(\left\|V_{S^cS} - G_{S^cS}\right\|_{1,\infty} + \left\|G_{S^cS}G_{SS}^{-1}\right\|_{1,\infty}\left\|V_{SS} - G_{SS}\right\|_{1,\infty})\left\|V_{SS}^{-1}\right\|_{1,\infty} \\ <&2s\epsilon_1\kappa.\end{aligned}$$

Use the fact that $\kappa s\epsilon_1 < \xi/4$ again, we have

$$\left\|V_{S^cS}V_{SS}^{-1}\right\|_{1,\infty} \le \left\|V_{S^cS}V_{SS}^{-1} - G_{S^cS}G_{SS}^{-1}\right\|_{1,\infty} + \left\|G_{S^cS}G_{SS}^{-1}\right\|_{1,\infty} < 1 - \xi/2.$$

Therefore

$$\begin{aligned}\left\|U_{S^c}(\tilde{\boldsymbol{\theta}})\right\|_\infty &\le \epsilon_2 + \left\|V_{S^cS}V_{SS}^{-1}\right\|_{1,\infty}(\epsilon_2 + \gamma_e) \\ &< \xi\gamma_e/4 + (1 - \xi/2)(\xi\gamma_e/4 + \gamma_e) \\ &< \gamma_e.\end{aligned}$$

Thus $\tilde{\boldsymbol{\theta}}$ satisfies the strict dual feasibility condition (EC.4). Therefore, the oracle solution $\tilde{\boldsymbol{\theta}}$ is a finite solution to the original problem and any solution has a support $\hat{S}$ as a subset of S. Thus $\hat{\boldsymbol{\theta}}$ is also a solution to the oracle problem (EC.5) and we have $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ since the objective function of the oracle problem (EC.5) is strictly convex.

Finally, by (EC.6), (EC.7), (EC.8), (EC.9), and (EC.10), if

$$\begin{cases} n > \max(C_3^{-1}, 8\xi^{-2}\kappa^2\overline{X}^4\overline{T}^2 s^2)^{\frac{1}{1-\nu}} \\ \gamma_e = 4\xi^{-1}C_3^{-\frac{1}{2}}C_4 n^{-\frac{1-\nu}{2}} \end{cases},$$

then with probability at least $1-2(p+1)(p+2)\exp(-n^\nu)$,

$$\begin{aligned} \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty =& \left\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty \\ =& \left\|\tilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}^*_S\right\|_\infty \\ =& \left\|V_{SS}^{-1}[U_S(\boldsymbol{\theta}^*) - \gamma_e \boldsymbol{\delta}_S \circ \operatorname{sign}(\tilde{\boldsymbol{\theta}}_S)]\right\|_\infty \\ \le& \left\|V_{SS}^{-1}\right\|_{1,\infty}(\left\|U_S(\boldsymbol{\theta}^*)\right\|_\infty + \gamma_e) \\ \le& 2\kappa(\epsilon_2 + \gamma_e) \\ \le& 10\xi^{-1}\kappa C_3^{-\frac{1}{2}}C_4 n^{-\frac{1-\nu}{2}}. \end{aligned}$$

Now, we complete the proof of part (i). □

EC.3.2. Proof of (ii) of Theorem 3.

Proof of Part (ii) of Theorem 3. Compared with the fixed p setting, the case where p diverges does not have a significant difference except for the tail probability.

Recall the L_∞-error of the Hessian matrix obtained by the union bound, which is

$$\mathbb{P}(\left\|V - G\right\|_\infty > \epsilon_1) \le 2(p+1)^2 \exp(-\frac{2n\epsilon_1^2}{\overline{X}^4\overline{T}^2}).$$

Let $\epsilon_1 = \overline{X}^2\overline{T}\sqrt{\frac{\zeta\log(p+1)}{2n}}$, then with probability at least $1-2/(p+1)^{\zeta-2}$, where $\zeta > 2$ is a constant,

$$\left\|V - G\right\|_\infty \le \overline{X}^2\overline{T}\sqrt{\frac{\zeta\log(p+1)}{2n}}.$$

We let $\epsilon_1 < \frac{\xi}{4s\kappa}$, which requires

$$n > 8\xi^{-2}\zeta\kappa^2\overline{X}^4\overline{T}^2 s^2\log(p+1). \tag{EC.11}$$

The martingale concentration gives

$$\mathbb{P}(\left\|U(\boldsymbol{\theta}^*)\right\|_\infty \ge \epsilon_2) \le 2(p+1)\exp\left(-C_3 n\min(\frac{\epsilon_2^2}{C_4^2}, \frac{\epsilon_2}{C_4})\right).$$

When $\epsilon_2 < C_4$, with probability at least $1-2/(p+1)^{\zeta-1}$, where $\zeta>1$ is a constant,

$$\left\|U(\boldsymbol{\theta}^*)\right\|_\infty \le C_4\sqrt{\frac{\zeta\log(p+1)}{C_3 n}}. \tag{EC.12}$$

For the strict dual feasibility condition (EC.4), we let $\epsilon_2 \le \xi\gamma_e/4 < C_4$, which suggests the choice of n and γ_e by (EC.12):

$$\begin{cases} n > C_3^{-1}\zeta\log(p+1) \\ \gamma_e = 4\xi^{-1}C_4\sqrt{\frac{\zeta\log(p+1)}{C_3 n}} \end{cases}. \tag{EC.13}$$

Finally, combining (EC.6), (EC.10), (EC.11), (EC.12), (EC.13), if

$$\begin{cases} n > \max(C_3^{-1}, 8\xi^{-2}\kappa^2\overline{X}^4\overline{T}^2 s^2)\cdot\zeta\log(p+1) \\ \gamma_e = 4\xi^{-1}C_4\sqrt{\frac{\zeta\log(p+1)}{C_3 n}} \end{cases},$$

then with probability at least $1-3/(p+1)^{\zeta-2}$,

$$\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty \le \left\|V_{SS}^{-1}\right\|_{1,\infty}(\left\|U_S(\boldsymbol{\theta}^*)\right\|_\infty+\gamma_e) \le 10\xi^{-1}\kappa C_4\sqrt{\frac{\zeta\log(p+1)}{C_3 n}}.$$

Now, we have completed the proof of part (ii). □

Therefore, the proof of Theorem 3 is complete.

EC.4. Proof of Theorem 4

Proof of Theorem 4. By the triangle inequality,

$$\begin{aligned} &\left|\widehat{\mathrm{att}}^{(\mathrm{direct})}_{t_{i^\star}}(R\mid D) - \mathrm{att}^{(\mathrm{direct})}_{t_{i^\star}}(R\mid D)\right| \\ =&\Big|\sum_{(t_i,e_i)\in R}\widehat{\mathrm{att}}^{(\mathrm{direct})}_{t_{i^\star}}(\{(t_i,e_i)\}\mid D) - \sum_{(t_i,e_i)\in R}\mathrm{att}^{(\mathrm{direct})}_{t_{i^\star}}(\{(t_i,e_i)\}\mid D)\Big| \\ \le&\sum_{(t_i,e_i)\in R}\left|\widehat{\mathrm{att}}^{(\mathrm{direct})}_{t_{i^\star}}(\{(t_i,e_i)\}\mid D) - \mathrm{att}^{(\mathrm{direct})}_{t_{i^\star}}(\{(t_i,e_i)\}\mid D)\right|. \end{aligned} \tag{EC.14}$$

So it is sufficient to show the theorem is true for each touchpoint. Let $\hat{\lambda}_1(t\mid\mathcal{H}_t^D)$ denote the estimated conditional intensity of conversion:

$$\hat{\lambda}_1(t\mid\mathcal{H}_t^D) = \hat{\mu}_1 + \sum_{i:t_i<t}\hat{\alpha}_{e_i1}\psi_{e_i1}(t-t_i).$$

Without loss of generality, assume $\overline{\psi}_1 \ge 1$. Then for the conversion intensity,

$$\begin{aligned} &\left|\hat{\lambda}_1(t_{i^\star}\mid\mathcal{H}^D_{t_{i^\star}}) - \lambda_1(t_{i^\star}\mid\mathcal{H}^D_{t_{i^\star}})\right| \\ =&\left|(\hat{\mu}_1+\sum_{i<i^\star}\hat{\alpha}_{e_i1}\psi_{e_i1}(t_{i^\star}-t_i)) - (\mu_1+\sum_{i<i^\star}\alpha_{e_i1}\psi_{e_i1}(t_{i^\star}-t_i))\right| \\ \le&|\hat{\mu}_1-\mu_1| + \sum_{i<i^\star}|\hat{\alpha}_{e_i1}-\alpha_{e_i1}|\,|\psi_{e_i1}(t_{i^\star}-t_i)| \\ \le&m\overline{\psi}_1\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\|_\infty, \end{aligned}$$

where $\boldsymbol{\theta} = (\mu_1, \boldsymbol{\alpha}_1^\top)^\top$. By Theorem 3, we can take n to be sufficiently large such that $m\overline{\psi}_1 \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_\infty < \lambda_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})/2$. As a result, we have $\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}}) > \lambda_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})/2$. Then the estimated direct removal effect of $\{(t_i, e_i)\}$ given D satisfies

$$
\begin{aligned}
&\left|\widehat{\mathrm{att}}^{(\mathrm{direct})}_{t_{i^\star}}(\{(t_i,e_i)\} \mid D) - \mathrm{att}^{(\mathrm{direct})}_{t_{i^\star}}(\{(t_i,e_i)\} \mid D)\right| \\
=&\left|\frac{\hat{\alpha}_{e_i 1}\psi_{e_i 1}(t_{i^\star} - t_i)}{\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})} - \frac{\alpha_{e_i 1}\psi_{e_i 1}(t_{i^\star} - t_i)}{\lambda_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})}\right| \\
\le&\overline{\psi}_1 \left|\frac{\hat{\alpha}_{e_i 1}}{\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})} - \frac{\alpha_{e_i 1}}{\lambda_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})}\right| \\
\le&\overline{\psi}_1 \left|\frac{\hat{\alpha}_{e_i 1}}{\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})} - \frac{\alpha_{e_i 1}}{\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})}\right| + \overline{\psi}_1 \left|\frac{\alpha_{e_i 1}}{\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})} - \frac{\alpha_{e_i 1}}{\lambda_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})}\right| \\
=&\overline{\psi}_1 \frac{|\hat{\alpha}_{e_i 1} - \alpha_{e_i 1}|}{\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})} + \overline{\psi}_1 |\alpha_{e_i 1}| \frac{|\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}}) - \lambda_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})|}{\hat{\lambda}_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})\lambda_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})} \\
\le&2\overline{\psi}_1 \frac{\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_\infty}{\lambda_1(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})} + 2\overline{\psi}_1 \left\|\boldsymbol{\theta}\right\|_\infty \frac{m\overline{\psi}_1 \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_\infty}{\lambda_1^2(t_{i^\star} \mid \mathcal{H}^D_{t_{i^\star}})}.
\end{aligned} \tag{EC.15}
$$

Combining (EC.14), (EC.15), and Theorem 3, we complete the proof of Theorem 4. □

EC.5. Proof of Theorem 5

Proof of Theorem 5. By the definitions and the triangle inequality,

$$
\begin{aligned}
&\left|\widehat{\mathrm{att}}^{(\mathrm{total})}_{t_{i^\star}}(R \mid D) - \mathrm{att}^{(\mathrm{total})}_{t_{i^\star}}(R \mid D)\right| \\
=&\left|\mathbb{E}[\widehat{\mathrm{att}}^{(\mathrm{direct})}_{t_{i^\star}}(\widehat{R^\diamond} \mid D) \mid D] - \mathbb{E}[\mathrm{att}^{(\mathrm{direct})}_{t_{i^\star}}(R^\diamond \mid D) \mid D]\right| \\
=&\left|\sum_{R \subseteq R' \subseteq \Omega} \widehat{\mathrm{att}}^{(\mathrm{direct})}_{t_{i^\star}}(R' \mid D)\mathbb{P}(\widehat{R^\diamond} = R' \mid D) - \sum_{R \subseteq R' \subseteq \Omega} \mathrm{att}^{(\mathrm{direct})}_{t_{i^\star}}(R' \mid D)\mathbb{P}(R^\diamond = R' \mid D)\right| \\
\le&\sum_{R \subseteq R' \subseteq \Omega} \left|\widehat{\mathrm{att}}^{(\mathrm{direct})}_{t_{i^\star}}(R' \mid D)\mathbb{P}(\widehat{R^\diamond} = R' \mid D) - \mathrm{att}^{(\mathrm{direct})}_{t_{i^\star}}(R' \mid D)\mathbb{P}(R^\diamond = R' \mid D)\right| \\
\le&\sum_{R \subseteq R' \subseteq \Omega} \left|\widehat{\mathrm{att}}^{(\mathrm{direct})}_{t_{i^\star}}(R' \mid D) - \mathrm{att}^{(\mathrm{direct})}_{t_{i^\star}}(R' \mid D)\right| \\
&+ \sum_{R \subseteq R' \subseteq \Omega} \left|\mathbb{P}(\widehat{R^\diamond} = R' \mid D) - \mathbb{P}(R^\diamond = R' \mid D)\right|,
\end{aligned} \tag{EC.16}
$$

where the last inequality is due to the the attribution scores and the probabilities involved being bounded by 1. Each component of the first term is bounded in Theorem 4. For the second term, the estimated version of thinning yields

$$
\mathbb{P}(\widehat{R^\diamond} = R' \mid D) = \prod_{(t_i,e_i) \in \Omega \setminus R} \mathbb{P}(\mathrm{Bernoulli}(1 - \frac{\hat{\lambda}_{e_i}(t_i \mid \mathcal{H}^{D \setminus R'}_{t_i})}{\hat{\lambda}_{e_i}(t_i \mid \mathcal{H}^D_{t_i})}) = \mathbb{1}_{\{(t_i,e_i) \in R'\}}).
$$

Therefore,

$$
\begin{aligned}
&\left|\mathbb{P}(\widehat{R^\diamond}=R'\mid D)-\mathbb{P}(R^\diamond=R'\mid D)\right| \\
=&\Bigg|\prod_{(t_i,e_i)\in\Omega\setminus R}\left(1-\frac{\hat\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D\setminus R'})}{\hat\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D})}\right)^{\mathbb{1}\{(t_i,e_i)\in R'\}}\left(\frac{\hat\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D\setminus R'})}{\hat\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D})}\right)^{1-\mathbb{1}\{(t_i,e_i)\in R'\}} \\
&-\prod_{(t_i,e_i)\in\Omega\setminus R}\left(1-\frac{\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D\setminus R'})}{\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D})}\right)^{\mathbb{1}\{(t_i,e_i)\in R'\}}\left(\frac{\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D\setminus R'})}{\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D})}\right)^{1-\mathbb{1}\{(t_i,e_i)\in R'\}}\Bigg| \\
\le&\sum_{(t_i,e_i)\in\Omega\setminus R}\left|\frac{\hat\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D\setminus R'})}{\hat\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D})}-\frac{\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D\setminus R'})}{\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D})}\right| \\
\le&(m-r)\max_{(t_i,e_i)\in\Omega\setminus R}\left|\frac{\hat\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D\setminus R'})}{\hat\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D})}-\frac{\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D\setminus R'})}{\lambda_{e_i}(t_i\mid\mathcal{H}_{t_i}^{D})}\right|. \qquad \text{(EC.17)}
\end{aligned}
$$

Using similar tricks in Theorem 4, we can show that each component in (EC.17) has the same rate of convergence up to some constants. Combining (EC.16), (EC.17), and Theorem 3, we can derive the desired bound. It suffices to make sure Theorem 3 holds for each $e=1,\dots,q$. The corresponding tail probabilities can be obtained by the union bound. Let $\boldsymbol{\theta}^{(e)}=(\mu_e,\boldsymbol{\alpha}_e^\top)^\top$ denote the true parameter and $\hat{\boldsymbol{\theta}}^{(e)}$ be its regularized estimate for $e=1,\dots,q$ in the following context. Then there exists $C_2>0$ such that

(i) If p is fixed and n is sufficiently large, then for any constant $0<\nu<1$,

$$
\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\mathrm{att}}_{t_{i^\star}}^{(\mathrm{total})}(R\mid D)-\mathrm{att}_{t_{i^\star}}^{(\mathrm{total})}(R\mid D)\right|>C_2n^{-\frac{1-\nu}{2}}\right) \\
\le&\sum_{e=1}^{q}\mathbb{P}\left(\left\|\hat{\boldsymbol{\theta}}^{(e)}-\boldsymbol{\theta}^{(e)}\right\|_\infty>10(\xi^{(e)})^{-1}\kappa^{(e)}(C_3^{(e)})^{-\frac12}C_4^{(e)}n^{-\frac{1-\nu}{2}}\right) \\
\le&q\cdot2(p+1)(p+2)\exp(-n^\nu) \\
\le&2p(p+1)(p+2)\exp(-n^\nu);
\end{aligned}
$$

(ii) If p diverges and n is sufficiently large, then for any constant $\zeta>3$,

$$
\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\mathrm{att}}_{t_{i^\star}}^{(\mathrm{total})}(R\mid D)-\mathrm{att}_{t_{i^\star}}^{(\mathrm{total})}(R\mid D)\right|>C_2\sqrt{\frac{\zeta\log(p+1)}{n}}\right) \\
\le&\sum_{e=1}^{q}\mathbb{P}\left(\left\|\hat{\boldsymbol{\theta}}^{(e)}-\boldsymbol{\theta}^{(e)}\right\|_\infty>10(\xi^{(e)})^{-1}\kappa^{(e)}(C_3^{(e)})^{-\frac12}C_4^{(e)}\sqrt{\frac{\zeta\log(p+1)}{n}}\right) \\
\le&q\cdot3/(p+1)^{\zeta-2} \\
<&3/(p+1)^{\zeta-3}.
\end{aligned}
$$

Therefore, the proof of Theorem 5 is complete. □

References

Hansen, N. R., Reynaud-Bouret, P. and Rivoirard, V. (2015), 'Lasso and probabilistic inequalities for multivariate point processes', *Bernoulli* **21**(1), 83–143.

Hong, M. and Luo, Z.-Q. (2017), 'On the linear convergence of the alternating direction method of multipliers', *Mathematical Programming* **162**(1-2), 165–199.

Lin, W. and Lv, J. (2013), 'High-dimensional sparse additive hazards regression', *Journal of the American Statistical Association* **108**(501), 247–264.

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Vol. 47, Cambridge University Press.

Wainwright, M. J. (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)', *IEEE Transactions on Information Theory* **55**(5), 2183–2202.

Zhao, P. and Yu, B. (2006), 'On model selection consistency of lasso', *The Journal of Machine Learning Research* **7**, 2541–2563.