Supplementary Materials of “On Robust Clustering of Event Stream Data"

Supporting materials are collected in this file. In Section A, we provide more details about the Catoni's influence function. Extensive discussions on the intensity-based metrics are provided in Section B. The construction of basis functions is given in Section C. Section D includes some explanations on the remarks in Section 3.2. More simulation results including ablation studies are summarized in Section E. Additional figures and tables for numerical sections are listed in Section F. Section H to Section L collect all technical proofs for the main theorems, propositions, and supporting lemma.

A Supporting Information of Catoni's Influence Function

We specifically choose the following Catoni's influence function,

$$\phi(x) = \begin{cases} \log(1 + x + 0.5 \cdot x^2) & x \le 2, \\ 0.032/9 \cdot (x - 9.5)^3 + 1.5 + \log(5) & 2 < x \le 9.5, \\ 1.5 + \log(5) & x > 9.5, \end{cases} \tag{1}$$

for $x \in \mathbb{R}^+$ and $\phi(0) = 0$. When $x < 0$, define $\phi(x) := -\phi(-x)$. It is not hard to see that the constructed $\phi(x)$ has the continuous second derivative, which facilitates the theoretical analyses.

Remark 1 *The constant (e.g. 9.5) in* (1) *could be modified. Here the only principle in choosing ϕ is that it satisfies* (2.1) *and is sufficiently smooth, that is, the second derivative is continuous.*

We provide the graphical illustrations of Catoni's influence function $\phi(x)$ and its derivative $\phi'(x)$ in Figure 1.

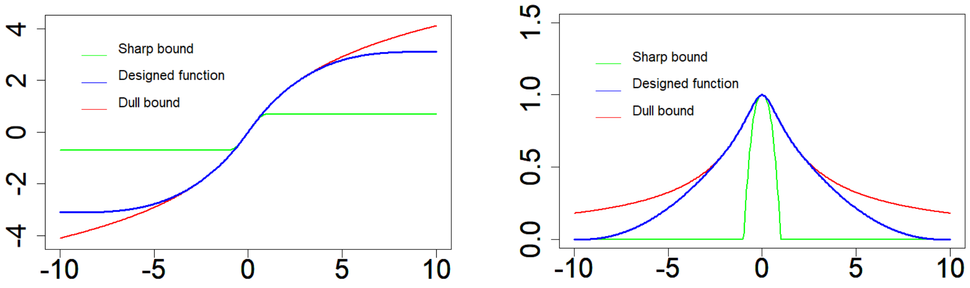

Figure 1: Left figure: Catoni influence function ϕ and the widest influence function ϕ_{dull} and the narrowest influence function ϕ_{sharp}. Right figure: First-order derivatives of ϕ, ϕ_{dull} and ϕ_{sharp}. For the definitions of ϕ_{dull} and ϕ_{sharp}, please refer to (2) and (3).

The first-order derivative and second-order derivative of the function can be derived as

$$\phi'(x) = \begin{cases} \dfrac{1+x}{1+x+0.5x^2} & x \le 2; \\ 0.032/3 \cdot (x - 9.5)^2 & 2 < x \le 9.5; \\ 0 & x \ge 9.5 \end{cases}$$

and

$$\phi''(x) = \begin{cases} -\dfrac{x+0.5x^2}{(1+x+0.5x^2)^2} & x \le 2; \\ 0.064/3 \cdot (x - 9.5) & 2 < x \le 9.5; \\ 0 & x \ge 9.5. \end{cases}$$

The formula of ϕ_{dull} and ϕ_{sharp} plotted in Figure 1 are given as follows.

$$\phi_{dull}(x) = \begin{cases} \log(1+x+\frac{1}{2}|x|^2) & x \geq 0 \\ -\log(1-x+\frac{1}{2}|x|^2) & x < 0, \end{cases} \tag{2}$$

and

$$\phi_{sharp}(x) = \begin{cases} -\log 2 & \text{if } x \leq -1 \\ -\log(1-x+\frac{1}{2}|x|^2) & \text{if } -1 \leq x \leq 0, \\ \log(1+x+\frac{1}{2}|x|^2) & \text{if } 0 < x \leq 1, \\ \log 2 & \text{if } x \geq 1. \end{cases} \tag{3}$$

B Literature on Intensity-based Distance

For the ease of discussion, throughout this section, we suppose all events are observed within time interval $[0, T]$, where T is a fixed real number. Most existing distances for TPPs are based on the random time change theorem [Brown et al., 2002]. That is, an event stream $S = (t_1, \ldots, t_N)$ is distributed according to a TPP with intensity $\lambda^*(t)$ on the time interval $[0, T]$ if and only if the transformed sequence $Z := (v_1, \ldots, v_N) = (\Lambda^*(t_1), \ldots, \Lambda^*(t_N))$ is distributed according to a standard Poisson process on $[0, \Lambda^*(T)]$, where $\Lambda^*(t) := \int_0^t \lambda^*(u)du$ is the cumulative intensity function.

Barnard [1953] proposed a Kolmogorov-Smirnov (KS) statistic-based metric, which quantifies the distance between observed event stream S and the theoretical intensity $\lambda^*(t)$. The idea is to check whether the transformed arrival times $v_1, \ldots, v_N$ are uniformly distributed within interval $[0, T]$. To do so, it compares $\hat{F}_{\text{arr}}$, the empirical cumulative distribution function (CDF) of the arrival times, with $F_{\text{arr}}(u) = u/\Lambda^*(T)$, the CDF of the uniform random variable. Specifically, the distance is defined as

$$\kappa_{\text{arr}}(S, \lambda^*(\cdot)) := \sqrt{N} \cdot \sup_{u \in [0,V]} \left| \hat{F}_{\text{arr}}(u) - F_{\text{arr}}(u) \right|,$$

where $\hat{F}_{\text{arr}}(u) = N^{-1} \sum_{i=1}^{N} \mathbf{1}(v_i \leq u)$.

Another possible metric relies on the fact that the inter-event time $w_i := v_{i+1} - v_i$ follows the standard exponential distribution (Cox and Lewis [1966]). It then compares $\hat{F}_{\text{int}}$, the empirical CDF of w_i's, and $F_{\text{int}}(u) := 1 - \exp(-u)$. This leads to

$$\kappa_{\text{int}}(S, \lambda^*(\cdot)) := \sqrt{N} \cdot \sup_{u \in [0,\infty)} \left| \hat{F}_{\text{int}}(u) - F_{\text{int}}(u) \right|,$$

where $\hat{F}_{\text{int}}(u) = (N+1)^{-1} \sum_{i=1}^{N+1} \mathbf{1}(w_i \leq u)$.

Although metrics κ_{arr} and κ_{int} are popular in testing the goodness-of-fit of various Poisson processes Daley et al. [2003], Gerhard et al. [2011], Alizadeh et al. [2013], Kim and Whitt [2014], Li et al. [2018], Tao et al. [2018], they still have many limitations. They suffer severe non-identifiability issues. Two very different event streams can be very close under such metrics. More failure modes of κ_{arr} and κ_{int} can be found in Pillow [2009].

Taking into account the above problems, Shchur et al. [2021] proposed a sum-of-squared-spacings metric,

$$\kappa_{sss}(S, \lambda^*(\cdot)) := \frac{1}{\Lambda^*(T)} \sum_{i=1}^{N+1} w_i^2 = \frac{1}{\Lambda^*(T)} \sum_{i=1}^{N+1} (v_i - v_{i-1})^2,$$

which extends the idea in Greenwood [1946]. As we can see, the above method can measure the closeness between the sample and the specific distribution well. However, they fail to meet the data requirements in our scenarios. To be more specific, we can only observe the sample data and has no information of model specification, which means that $\lambda^*(\cdot)$ or $\Lambda^*(\cdot)$ is unknown. For any two samples S_1 and S_2, of course, we can consider to estimate $\Lambda_1^*(\cdot)$ ($\Lambda_2^*(\cdot)$) based on sample S_1 (S_2) first, and then calculate the above KS-type distance between sample S_2 (S_1) and the estimated $\Lambda_1^*(\cdot)$ ($\Lambda_2^*(\cdot)$). Unfortunately, this procedure makes it not symmetric about S_1 and S_2 and also fails to satisfy the triangle inequality. As a result, it is not a proper metric distance.

C Construction of Spline Basis

Let $U = (u_0, u_1, \ldots, u_H)$ be a set of $H+1$ non-decreasing numbers satisfying $0 = u_0 < u_1 \cdots < u_H = T$. (We may treat $T = 1$ for the ease of presentation). Points u_i's are called knots and the set U is known as the knot vector, and the half-open interval $[u_i, u_{i+1})$ the i-th knot span. For practical use, the knots are usually equally spaced, i.e., $u_{i+1} - u_i$ is a constant equal to $\Delta u := T/H$ for $0 \leq i \leq H-1$. To construct the cubic spline basis functions, we follow the classical procedure by defining $N_{i,p}(u)$ as the i-th B-spline basis function of degree p. Then its formula can be recursively written as

$$N_{i,0}(u) = \begin{cases} 1 & \text{if } u_i \leq u < u_{i+1} \\ 0 & \text{otherwise} \end{cases},$$

$$N_{i,p}(u) = \frac{u - u_i}{u_{i+p} - u_i} N_{i,p-1}(u) + \frac{u_{i+p+1} - u}{u_{i+p+1} - u_{i+1}} N_{i+1,p-1}(u).$$

The above is usually referred to as the Cox-deBoor recursion formula [De Boor, 1972]. Applying the Cox-deBoor recursion formula, the first cubic spline basis function $\kappa_1(\cdot)$ can be found as follows.

$$\kappa_1(u) = \begin{cases} \frac{1}{6\Delta u^3} u^3, & u \in [0, \Delta u], \\ \frac{1}{6\Delta u^3} \left((2\Delta u - u)u^2 + (u - \Delta u)(4\Delta u - u)(3\Delta u - u) + (4\Delta u - u)(u - \Delta u)^2 \right), & u \in [\Delta u, 2\Delta u], \\ \frac{1}{6\Delta u^3} \left((u - 4\Delta u)^2(u - 2\Delta u) + (u - \Delta u)(4\Delta u - u)(3\Delta u - u) + (u - 3\Delta u)^2 u \right), & u \in [2\Delta u, 3\Delta u], \\ \frac{1}{6\Delta u^3} (4\Delta u - u)^3, & u \in [3\Delta u, 4\Delta u]. \end{cases}$$

For $h \in \{2, ..., H\}$, we can define h-th basis $\kappa_h(u) := \kappa_1(u - h\Delta u)$. (When $u < h\Delta u$, $\kappa_h(u) = \kappa_1(u - h\Delta u + T)$.)

D Additional Comments on Remarks

Explanations of Remark 1. In addition to the non-homogeneous Poisson model given in the main context, we can use other types of TPPs. For example, to capture more temporal dependencies, we can take it to be the self-exciting process (also known as the Hawkes process, Hawkes and Oakes [1974]),

$$\lambda_k(t) := \sum_{h=1}^{H} b_{k,h} \kappa_h(t) + \sum_{t_i < t} \sum_{h'=1}^{H'} \alpha_{k,h'} g_{h'}(t - t_i), \tag{4}$$

where $g_{h'}(t)$'s is another set of basis functions for modelling the triggering part. To account for more individual heterogeneity, we can take it to be the frailty model [Duchateau and Janssen, 2008],

$$\lambda_k(t) := \omega \cdot \left(\sum_{h=1}^{H} b_{k,h} \kappa_h(t) \right), \tag{5}$$

where ω is a positive random variable (e.g. log-normal, gamma, etc.).

Explanations of Remark 3. When using other optimization methods, the main modification lies in the robust objective (3.11). For VI, (3.11) changes to

$$\sum_{n=1}^{N} q_{nk}^{(t)} \left(\mathbb{E}_{q^{(t)}} [\log \mathrm{NHP}\,(S_n \mid \boldsymbol{B}_k)] - \mu \right) = 0, \tag{6}$$

where $q_{nk}^{(t)}$ and $q^{(t)}$'s are the variational distributions. For stochastic EM, (3.11) changes to

$$\sum_{n=1}^{N} \mathbf{1}\{Z_n^{(t)} = k\} \cdot (\log \mathrm{NHP}\,(S_n \mid \boldsymbol{B}_k) - \mu) = 0, \tag{7}$$

where $Z_n^{(t)}$ follows the multinomial$(r_{n1}^{(t)}, ..., r_{nK}^{(t)})$.

E Additional Simulation Results

E.1 Working model: Hakwes Process

In this setting, the inlier event sequences are generated according to the 4-class Hawkes process. The corresponding intensity functions are

$$\lambda^*_{k,hawkes}(t) = \lambda^*_k(t)/2 + \sum_{t_j<t} g^*_k(t-t_j), \ \ k \in \{1,2,3,4\}, \tag{8}$$

where $\lambda^*_k(t)$'s are the same as in the main context and $g^*_k(t)$'s are given as follows.

$$\begin{aligned} g^*_1(t) &= \frac{0.05}{\sqrt{\pi}}\exp(-t^2/4), g^*_2(t) = \frac{0.1}{3\sqrt{\pi}/2}\exp(-t^2/9), \\ g^*_3(t) &= \frac{0.15}{\sqrt{\pi}}\exp(-t^2/4), g^*_4(t) = \frac{0.15}{3\sqrt{\pi}/2}\exp(-t^2/9). \end{aligned}$$

We still consider three types of outlier event sequences. The first two types are the same as in the previous subsection, while the sequences from the third type follow $\lambda_{out,hawkes}(t) = \lambda_{out3}(t) + \sum_{t_j<t} g_{out3}(t-t_j)$ with $g_{out3}(t) = 0.5\exp(-t^2/9)/(1.5\sqrt{\pi})$ and

$$\begin{aligned} \lambda_{out3}(t) = {} & 25/2 \cdot \exp(-(t - 24 \cdot B_1)^2/0.03) + 25/3 \cdot \exp(-(t - 24 \cdot B_2)^2/0.03) \\ & + 25/6 \cdot \exp(-(t - 24 \cdot B_3)^2/0.03), \end{aligned}$$

where $B_i \sim U(0,1) + 0.1, \forall i \in \{1,2,3\}$.

The choices of N, L, M, N', ϵ, and α remain the same. In addition, we set $H' = 6$ and $\rho = 0.8 \cdot \sqrt{\int_0^T \log^2 \lambda_k^{(0)}(t) \cdot \lambda_k^{(0)}(t)dt}$, where H' is selected according to Xu and Zha [2017]. The results are summarized in Table A.

E.2 Working model: Frailty model

In this setting, the inlier sequences follow the four-class frailty model with

$$\begin{aligned} \lambda^*_1(t) &= \omega \cdot (5/3\exp(-(t+4.8)^2/10) + 5/3\exp(-(t-2.4)^2/50)), \\ \lambda^*_2(t) &= \omega \cdot (5/3\exp(-(t-6)^2/4) + 15/4\exp(-(t-21.6)^2/4)), \\ \lambda^*_3(t) &= \omega(15/4\exp(-(t-4.8)^2/1.5) + 35/12\exp(-(t-12)^2) + 15/4\exp(-(t-19.2)^2/1.5)), \\ \lambda^*_4(t) &= \omega \cdot (10/3\exp(-(t-21.6)^2/40) + 5/3\exp(-(t-26.4)^2/10)), \end{aligned}$$

where the frailty $\omega \sim \text{Lognormal}(-0.1, 0.2)$. Outlier generation procedure are almost the same as before except for the third type, which has the following intensity formula $\lambda_{out3}(t) = \omega \cdot 125/3 \cdot \exp(-(t - 24 \cdot B_1)^2/0.01)$, where $\omega \sim \text{Lognormal}(-0.1, 0.2)$ and $B_1 \sim U(0,1)$. We replace EM algorithm by VI method as described in Remark 3. In the experiment, we set tuning parameter ρ for class k to be $2 \cdot \sqrt{\int_0^T \log^2 \lambda_k^{(0)}(t) \cdot \lambda_k^{(0)}(t)dt}$, $H = 10$, $\epsilon = 0.1$, $\alpha = 0.2$, $\beta = 0.3$, $M = 30$, and $N' = 0.6 \cdot N$. The results are summarized in Table B.

E.3 Comments on the Baseline

To end the simulation section, we explain the reason why we do not include another baseline, Algorithm 1 with proposed initialization but without robust influence function, in our simulation. Such baseline method may have obvious defects. Consider a case that the inlier event streams are from homogeneous Poisson process of four classes, whose intensities are 1, 2, 3, and 4, respectively. There are 30 event sequences for each class and one outlier event sequence which follows a Poisson process with intensity 100. In this case, even if we start from the true values, it still leads to bad classification result if ϕ_ρ is not used. To see this, after the first iteration, the

Time	Algorithm	type 1			type 2			type 3		
		$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$
$L=1$	Standard	0.5061	0.6355	0.7748	0.4479	0.5467	0.6415	0.7371	0.8247	0.8507
	Robust	0.5681	0.6764	0.7896	0.6248	0.7438	0.8291	0.8370	0.8725	0.8869
	Robust & Initialization	**0.7514**	**0.8901**	**0.8913**	**0.7262**	**0.8853**	**0.8868**	**0.8565**	**0.8848**	**0.8868**
$L=2$	Standard	0.6835	0.9064	0.9673	0.5134	0.6909	0.8146	0.8745	0.9314	0.9634
	Robust	0.7754	0.9725	0.9752	0.9145	0.9790	0.9780	0.9076	0.9581	0.9763
	Robust & Initialization	**0.8709**	0.9644	0.9742	0.8940	0.9495	**0.9780**	**0.9517**	**0.9763**	**0.9757**
$L=4$	Standard	0.6993	0.7993	0.8690	0.6596	0.7748	0.8994	0.8127	0.8888	0.9232
	Robust	0.8393	0.8693	0.9144	0.7894	0.9041	0.9387	0.8283	0.9186	0.9233
	Robust& Initialization	**0.9340**	**0.9986**	**0.9986**	**0.8943**	**0.9493**	**0.9993**	**0.9477**	**0.9681**	**0.9979**

Table A: Purity indices returned by three algorithms under the setting of outlier type 1, 2, and 3 with Hawkes process working model.

Time	Algorithm	type 1			type 2			type 3		
		$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$
$L=1$	Standard	0.5797	0.7000	0.8715	0.3546	0.3410	0.3967	0.8598	0.9310	0.9357
	Robust	0.7303	0.8543	0.9011	0.6850	0.7891	0.8574	0.8036	0.8883	0.9258
	Robust & Initialization	**0.8938**	**0.9399**	**0.9364**	**0.9318**	**0.9459**	**0.9433**	**0.8683**	**0.9318**	**0.9492**
$L=2$	Standard	0.6367	0.7983	0.9418	0.3002	0.3589	0.4243	0.7506	0.7795	0.8447
	Robust	0.8423	0.9534	0.9824	0.9292	0.9667	0.9867	0.7007	0.7350	0.8232
	Robust & Initialization	**0.8708**	**0.9601**	**0.9927**	0.9055	**0.9683**	0.9817	**0.9026**	**0.9327**	**0.9774**
$L=4$	Standard	0.7085	0.8594	0.9348	0.4451	0.6141	0.7428	0.5348	0.5596	0.6198
	Robust	0.8199	0.8537	0.9014	0.8188	0.8843	0.9147	0.5450	0.5649	0.6258
	Robust& Initialization	**0.8127**	**0.9433**	**0.9681**	**0.9191**	**0.9940**	**0.9840**	**0.8949**	**0.9598**	**0.9894**

Table B: Purity indices returned by three algorithms under the setting of outlier type 1, 2, and 3 with Frailty working model.

outlier will be classified into class 4 and the intensity parameter of this class will be updated to approximately $(30 \times 4 + 100)/31 \approx 7.10$. After the second iteration, event streams from class 3 and 4 will be mixed together and the intensity parameter of four classes will be approximately 1, 2, 3.5, and 100, respectively. Then the algorithm converges in the next iteration. Therefore, outlier is classified into a single class and purity index is no larger than 0.75. This indicates the usefulness of ϕ_ρ.

E.4 Ablation study

Here we provide additional ablation studies to show the effectiveness of each component in the proposed algorithms 1 and 2. In particular, we answer the following questions.

a Is the robust estimation part useful? In other words, can $w_{nk}^{(t)}$ be removed by being replaced with 1?

b Is the proposed distance induced by the cubic splines useful? What is the performance like when other common metrics replace it?

c In the inlier weighting part, we adopt the K-means++ algorithm. Does the performance change when we use other clustering algorithms?

d Can the role of normalizers M_A and M_B in distance 3.2 be replaced by the corresponding L_p norm of the intensity functions?

e Whether the results are sensitive to the choices of the hyperparameters (e.g. ρ, N') in the proposed algorithm?

f What will the results look like if we do not take into account the time shift when there truly exists such phenomena in the data?

To address all the above issues, we compare the proposed method with several baselines under the same NHP setting as described in Section 5. “Standard & Initialization" denotes the algorithm with $w_{nk}^{(t)}$ being replaced with 1. “Frechet distance" denotes the algorithm by replacing the proposed distance with Frechet distance. “Spectral clustering" denotes the algorithm by replacing K-means++ with the spectral clustering algorithm. In addition to the purity index, we also report the performance of different algorithms under the other two metrics, the Adjusted Rand Index (ARI) and the silhouette scores (Silhouette), whose definitions are given Supplementary G. The comparative results are given in Table C. We additionally consider using L_1 and L_2 norms to replace the normalizers M_A and M_B. The results are given in Table D. We also choose different $\rho = 1.2$, 0.3, or $N' = 0.65$, 0.83. The results are reported in Table E. Moreover, we compare the results by using or not using the shift-invariant distance for the time-shifted data. The results are summarized in Table F.

Results from Table C suggest that the proposed algorithm can achieve much better performance than the other baselines when the outlier type is 1 or 2, and can have competitive performance when the outlier type is 3. As a result, all components designed in the proposed algorithm play a useful role. Results from Table D suggest that the original normalization consistently yields better clustering performance. It indicates that the square-root normalization maintains superior adaptability to density fluctuations compared to L_1/L_2-normalized counterparts in our framework. From Table E, we can see that the algorithm with modified hyperparameters ($\rho = 0.3, 1.2$ vs original 0.6; $N' = 0.65, 0.83$ vs original 0.75) shows no significant performance difference compared to the default settings, demonstrating a relatively wide selection range for these parameters. The results in Table F demonstrate that applying the non-shift invariant distance to the shifted data significantly underperformed compared to the shift-based approach, thereby validating the effectiveness of the shift-invariant metric in handling temporal misalignments during event stream clustering.

Outlier Type	Algorithm	Purity			ARI			Silhouette		
		$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$
Type 1	Standard & Initialization	0.7172	0.9000	0.9599	0.6692	0.8808	0.9376	0.4622	0.5438	0.5384
	Robust & Initialization	**0.9300**	**0.9925**	**1**	**0.9152**	**0.9814**	**0.9609**	**0.5782**	**0.5855**	0.5192
	Frechet distance	0.4875	0.6025	0.7075	0.3387	0.4933	0.6300	0.3555	0.4027	0.4084
	Spectral clustering	0.8425	0.8975	0.9525	0.7869	0.8391	0.8834	0.4886	0.4879	0.4648
Type 2	Standard & Initialization	0.5048	0.6550	0.7900	0.4142	0.5988	0.7564	0.3977	0.4581	0.5108
	Robust & Initialization	**0.9200**	**0.9825**	**0.9975**	**0.9047**	**0.9655**	**0.9510**	**0.5754**	**0.5732**	**0.5198**
	Frechet distance	0.5850	0.7575	0.8625	0.4793	0.6986	0.8112	0.3826	0.4500	0.4538
	Spectral clustering	0.7700	0.8450	0.9075	0.7100	0.7894	0.8444	0.4607	0.4844	0.4752
Type 3	Standard & Initialization	0.9795	0.9873	1	0.9749	0.9676	0.9635	0.6118	0.5692	0.5340
	Robust & Initialization	0.9772	0.9924	**1**	0.9717	0.9678	0.9464	0.6108	0.5642	0.5051
	Frechet distance	0.9598	1	1	0.9526	0.9868	0.9677	0.5990	0.5896	0.5211
	Spectral clustering	0.8127	0.9175	0.9250	0.7694	0.8865	0.8768	0.5177	0.5447	0.4920

Table C: Performance metrics (Purity, ARI, Silhouette) returned by four algorithms under three different outlier types with non-homogeneous Poisson working model. "**Robust & Initialization**" stands for the proposed method.

Outlier Type	Lp-norm	Purity			AIRI			Silhouette		
		$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$
Type 1	**Default**	**0.9300**	**0.9925**	**1**	**0.9152**	**0.9814**	**0.9609**	**0.5782**	**0.5855**	**0.5192**
	$p=1$	0.4700	0.6000	0.7000	0.3113	0.4868	0.6285	0.3419	0.4069	0.4067
	$p=2$	0.4650	0.6100	0.6826	0.3048	0.5067	0.6012	0.3479	0.4082	0.4023
Type 2	**Default**	**0.9200**	**0.9825**	**0.9975**	**0.9047**	**0.9655**	**0.9510**	**0.5754**	**0.5732**	**0.5198**
	$p=1$	0.5950	0.7200	0.8250	0.5005	0.6356	0.7651	0.4065	0.4186	0.4492
	$p=2$	0.5974	0.7125	0.8300	0.4987	0.6406	0.7704	0.3974	0.4490	0.4535
Type 3	**Default**	**0.9772**	0.9924	**1**	**0.9717**	0.9678	0.9464	**0.6108**	0.5642	0.5051
	$p=1$	0.9500	0.9975	1	0.9415	0.9847	0.9702	0.5903	0.5870	0.5291
	$p=2$	0.9399	0.9950	1	0.9301	0.9828	0.9720	0.5833	0.5820	0.5404

Table D: Performance metrics (Purity, ARI, Silhouette) returned by different L_p-norm normalizations under three outlier types with non-homogeneous Poisson working model. "**Default**" stands for the original proposed method.

An important question for the practitioner is how to choose the optimal number of clusters. Here we propose a criterion, the adjusted Bayesian Information Criterion (adjusted BIC),

$$\mathrm{BIC}_{adj} = -2 \sum_{n:S_n \in \mathcal{S}_{in}} \log \mathrm{NHP}\left(S_n \mid \hat{\boldsymbol{B}}_{k(n)}\right) + C_{adj} \cdot \Theta_{all} \cdot \log(|\mathcal{S}_{in}|), \tag{9}$$

Outlier Type	Algorithm	Purity			AIRI			Silhouette		
		$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$	$K=4$	$K=5$	$K=6$
Type 1	**Default**	0.9300	0.9925	**1**	0.9152	0.9814	**0.9609**	0.5782	0.5855	0.5192
	$\rho = 1.2$	0.9325	0.9875	0.9974	0.9183	0.9734	0.9568	0.5795	0.5753	0.5080
	$\rho = 0.3$	0.9275	0.9925	0.9925	0.9134	0.9742	0.9378	0.5793	0.5820	0.4965
	$N' = 0.65$	0.9474	0.9950	1	0.9378	0.9816	0.9544	0.5944	0.5840	0.5071
	$N' = 0.83$	0.8950	0.9825	0.9950	0.8762	0.9732	0.9543	0.5693	0.5995	0.6193
Type 2	**Default**	0.9200	0.9825	0.9975	0.9047	0.9655	0.9510	0.5754	0.5732	0.5198
	$\rho = 1.2$	0.9450	0.9850	1	0.9356	0.9663	0.9583	0.5932	0.5696	0.5284
	$\rho = 0.3$	0.9425	0.9850	1	0.9329	0.9676	0.9609	0.5938	0.5823	0.5304
	$N' = 0.65$	0.9300	0.9850	0.9950	0.9176	0.9622	0.9458	0.5847	0.5653	0.5171
	$N' = 0.83$	0.8225	0.9425	0.9875	0.7920	0.9260	0.9568	0.5314	0.5798	0.5524
Type 3	**Default**	**0.9772**	0.9924	**1**	**0.9717**	0.9678	0.9464	**0.6108**	0.5642	0.5051
	$\rho = 1.2$	0.9675	1	1	0.9612	0.9817	0.9518	0.6049	0.5742	0.5181
	$\rho = 0.3$	0.9675	1	1	0.9606	0.9706	0.9670	0.6067	0.5558	0.4939
	$N' = 0.65$	0.9673	0.9925	1	0.9611	0.9706	0.9577	0.4762	0.6193	0.5490
	$N' = 0.83$	0.9350	0.9900	0.9975	0.9250	0.9726	0.9502	0.5867	0.5752	0.5215

Table E: Performance metrics (Purity, ARI, Silhouette) with different hyperparameters under three outlier types

Outlier Type	Algorithm	Shift Data		
		$K=4$	$K=5$	$K=6$
Type 1	no-shift	0.5134	0.5300	0.5732
	shift	0.7610	0.7728	0.7910
Type 2	no-shift	0.5308	0.5553	0.5903
	shift	0.7969	0.8026	0.8233
Type 3	no-shift	0.5167	0.5368	0.5708
	shift	0.6735	0.7356	0.8083

Table F: Purity metrics for the algorithms using shift-invariant distance or no-shift version on the shift data.

where $k(n) := \arg\max_k \hat{r}_{nk}$ is the estimated label of sample n, Θ_{all} represents the number of parameters, i.e., $\Theta_{all} = KH + K - 1$ for K clusters and C_{adj} is a constant that can be tuned by the user. We suggest choosing $C_{adj} = 25$. We apply the adjusted BIC to the proposed algorithm. The results are given in Table G It is clear that this criterion can consistently choose the underlying number of clusters, i.e., $K^* = 4$.

Lastly, we include the additional experiments with reduced outlier proportion ($\eta = 5\%$) and outlier-free scenarios ($\eta = 0\%$) in Table H, confirming that our method maintains effectiveness across these configurations. As a result, the proposed method works well under both realistic sparse-outlier conditions ($< 5\%$) and the ideal outlier-absent environments.

Outlier Type	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
Type 1	-46588.07	-49121.01	**-50203.39**	-50065.27	-49303.61
Type 2	-45723.89	-47818.59	**-48774.43**	-48228.51	-47525.22
Type 3	-45326.24	-47061.89	**-48394.52**	-48047.19	-47242.43

Table G: The adjusted BIC values for K=2 to 6 under the non-homogeneous Poisson working model, evaluated with three different outlier types.

Outlier Type	Algorithm	Purity			ARI			Silhouette		
		$K = 4$	$K = 5$	$K = 6$	$K = 4$	$K = 5$	$K = 6$	$K = 4$	$K = 5$	$K = 6$
Type 1 (5%)	Standard & Initialization	0.7508	0.9352	0.9826	0.7115	0.9180	0.9628	0.4721	0.5544	0.5450
	Robust & Initialization	**0.9552**	**1**	**0.9975**	**0.9396**	**0.9709**	0.9321	**0.5802**	0.5262	0.4497
	Standard	0.7509	0.7683	0.8306	0.7104	0.7307	0.8030	0.4485	0.4612	0.5052
	Robust	0.7608	0.7808	0.8356	0.7220	0.7452	0.8088	0.4556	0.4700	0.5089
Type 2 (5%)	Standard & Initialization	0.8966	0.9829	0.9924	0.8762	0.9727	0.9718	0.5643	0.5931	0.5618
	Robust & Initialization	**0.9701**	**0.9950**	**0.9999**	**0.9621**	0.9631	0.9318	**0.6026**	0.5477	0.4650
	Standard	0.7528	0.7708	0.8231	0.7118	0.7334	0.7942	0.4504	0.4639	0.5012
	Robust	0.8729	0.9278	0.9551	0.8522	0.9159	0.9472	0.5370	0.5765	0.5899
Type 3 (5%)	Standard & Initialization	0.9676	1	1	0.9611	0.9855	0.9578	0.6037	0.5752	0.5069
	Robust & Initialization	0.9751	**1**	**1**	0.9684	0.9631	0.9124	0.6072	0.5439	0.4384
	Standard	1	1	1	0.9999	0.9986	0.9976	0.6265	0.6155	0.5981
	Robust	1	1	1	0.9999	0.9992	0.9977	0.6265	0.6175	0.5996
No outlier (0%)	Standard & Initialization	0.9700	0.9975	0.9975	0.9605	0.9782	0.9561	0.5974	0.5555	0.4750
	Robust & Initialization	0.9850	**1**	**1**	0.9804	0.9668	0.9248	0.6129	0.5115	0.4077
	Standard	1	1	1	1	0.9997	0.9987	0.6284	0.6249	0.6151
	Robust	1	1	1	1	0.9997	0.9990	0.6284	0.6238	0.6171

Table H: The performance indices (Purity, ARI, Silhouette) of four algorithms at outlier proportions ranging from 0% to 5% under the non-homogeneous Poisson model framework.

F Additional Figures and Tables in Numerical Studies

To help readers to gain more intuitions, the curves of intensity function considered in simulation studies are shown in Figure 2.

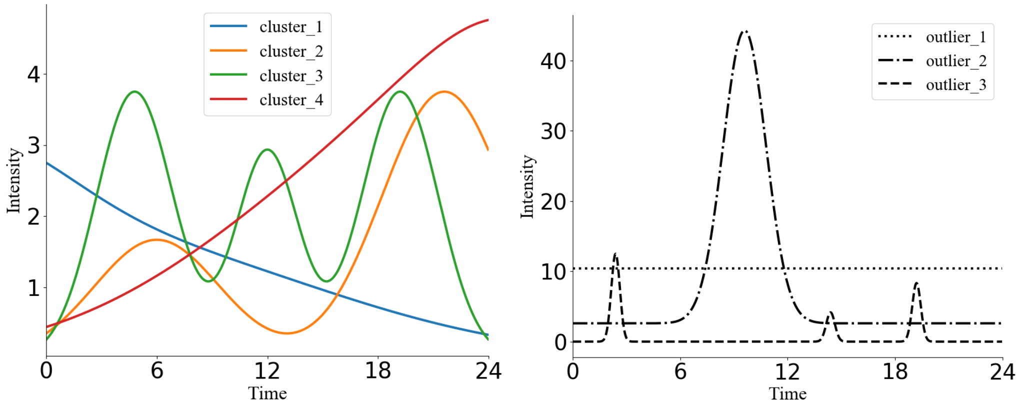

Figure 2: Left: Intensity functions of inlier event streams from 4 classes. Right: Intensity functions of outlier event streams of three types. Due to the randomness of λ_{out1} - λ_{out3}, curves are shown with one random realization of u.

The following Table I and Table J give a quick overview of two real data sets, IPTV and Last.FM, in our numerical studies. We can see that users have a sequence of events of watching TV or listening to music in their daily lives.

	id	time
1	55357201	2012/01/01 18:33:15
2	55357201	2012/01/01 18:34:55
...	...	...
4145	55357201	2012/11/28 02:01:42
4146	55357201	2012/11/28 02:04:01

Table I: IPTV dataset. "id": user identifier. "time": the time stamp when the user started to watch a TV program.

	user_id	time
1	user000685	2005/12/10 06:23:10
2	user000685	2005/12/10 06:26:35
...	...	...
84441	user000685	2009/05/22 06:44:01
84442	user000685	2009/05/23 11:12:10

Table J: Last.FM 1K Dataset. "user_id": user identifier. "time": the time stamp when the user played a song track.

The frequency plots of two real data sets are given in Figure 3 and Figure 4. It empirically indicates the existence of daily effect in user behaviors, i.e., the period of event sequences can be viewed as 24 hours.

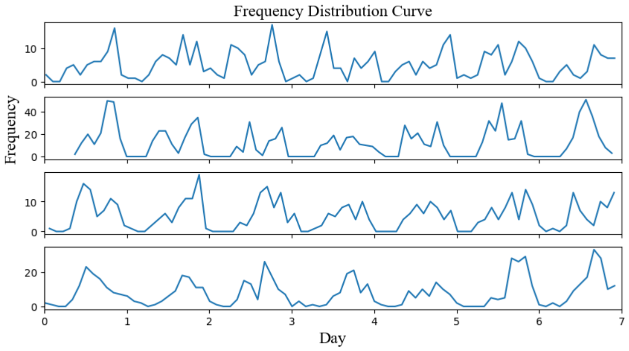

Figure 3: IPTV data: the frequency plot of four randomly selected households.

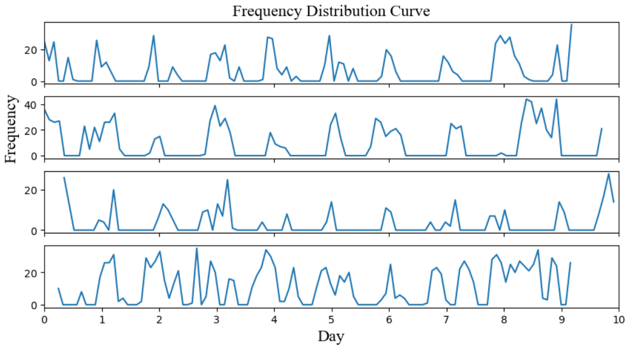

Figure 4: Last.FM 1K User Dataset: the frequency plot of four randomly selected users.

G Details of Performance Metrics

To be self-complete, the details of two metrics, ARI and Silhouette, in our setting are given as follows. The Adjusted Rand Index Halkidi et al. [2002] is defined by

$$\text{ARI} = \frac{\sum_{k,k'} \binom{N_{kk'}}{2} - \left[\sum_k \binom{N_k}{2} \sum_{k'} \binom{N_{k'}}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_k \binom{N_k}{2} + \sum_{k'} \binom{N_{k'}}{2}\right] - \left[\sum_k \binom{N_k}{2} \sum_{k'} \binom{N_{k'}}{2}\right] / \binom{N}{2}}$$

Here, N is the number of data points in a given data set and $N_{kk'} = |\hat{\mathcal{S}}_k \cap \mathcal{S}^*_{k'}|$, $N_k = |\hat{\mathcal{S}}_k|$ and $N_{k'} = |\mathcal{S}^*_{k'}|$.

The silhouette scores Kaufman and Rousseeuw [2009], Rousseeuw [1987] is defined as follows: for any pair of event sequences S_i and S_j, we define the following d-index,

$$d(S_i, S_j) = \int_0^T \left|\hat{\lambda}_i(t) - \hat{\lambda}_j(t)\right| dt, \tag{10}$$

where $\hat{\lambda}_i(t)$ is the estimated intensity function of sample S_i via cubic spline approximation.

For sequences S_i in class k, let

$$a(S_i) = \frac{1}{|\hat{\mathcal{S}}_k| - 1} \sum_{S_j \in \hat{\mathcal{S}}_k, i \neq j} d(S_i, S_j)$$

be the mean distance between S_i and all other sequences in the same cluster, where $|\hat{\mathcal{S}}_k|$ is the number of points belonging to cluster $\hat{\mathcal{S}}_k$.

We then define the mean dissimilarity of sequences S_i to some cluster $k' \neq k$ as the mean of the distance from S_i to all sequence in $\hat{\mathcal{S}}_{k'}$. For each sequence $S_i \in \hat{\mathcal{S}}_k$, we now define

$$b(S_i) = \min_{k' \neq k} \frac{1}{|\hat{\mathcal{S}}_{k'}|} \sum_{j \in \hat{\mathcal{S}}_{k'}} d(S_i, S_j)$$

to be the smallest mean distance of S_i to all sequences in any other cluster. The cluster with the smallest mean dissimilarity is said to be the neighboring cluster of S_i because it is the next best fit cluster for point S_i.

We now define the silhouette value of sequence S_i

$$s(S_i) = \frac{b(S_i) - a(S_i)}{\max\{a(S_i), b(S_i)\}}, \text{ if } |\hat{\mathcal{S}}_k| > 1$$

and

$$s(S_i) = 0, \text{ if } |\hat{\mathcal{S}}_k| = 1.$$

From the above definition, it is clear that $-1 \leq s(S_i) \leq 1$. Then the final silhouette coefficient can be calculated as $\sum_{i=1}^{N} s(S_i)/N$. The values closer to 1 indicate better performance.

H Proof of Propositions

Proof of Proposition 1 First, we consider the case where f is a constant value function, such as f being always equal to 1. If X follows a Poisson distribution with parameter λ, we prove that the variance of $\sqrt{X}$ is approximately $1/4 + O(1/\lambda)$. In general, for a smooth $g(X)$, we can do a Taylor expansion around the mean $\lambda = \mathbb{E}(X)$, so we have

$$g(X) = g(\lambda) + g'(\lambda)(X-\lambda) + \frac{g''(\lambda)}{2!}(X-\lambda)^2 + \frac{g'''(\lambda)}{3!}(X-\lambda)^3 + o(g'''(\lambda)(X-\lambda)^3).$$

Therefore $\mathbb{E}[g(X)] = g(\lambda) + \frac{g''(\lambda)}{2!}m_2 + \frac{g'''(\lambda)}{3!}m_3 + o(g'''(\lambda)m^3)$, where m_i is the i-th centered moment. In our case $m_2 = m_3 = \lambda$, thus

$$\mathbb{E}[\sqrt{X}] = \sqrt{\lambda} - \frac{\lambda^{-1/2}}{8} + \frac{\lambda^{-3/2}}{16} + o(\lambda^{-3/2}),$$

which indicates that the expected value is approximately $\sqrt{\lambda}$. Taking square of it, it gives

$$\left(\mathbb{E}[\sqrt{X}]\right)^2 = \lambda - \frac{1}{4} + \frac{9}{64\lambda} + o\left(\frac{1}{\lambda}\right).$$

Then $\mathrm{Var}(\sqrt{X}) = 1/4 - 9/(64\lambda) + o(1/\lambda)$, which is approximately $1/4$ for large λ.

Next, we divide the interval $[0,T]$ into n segments, each of which is $0 = a_0 < a_1 < \cdots < a_{n-1} < a_n = T$. Write $X_i := N(T)^{-1/2}\sum_{t_j \in (a_{i-1},a_i)} f(t_j)$, then $\mathrm{var}(X_i) = T^{-1}\int_{a_{i-1}}^{a_i} f^2(t)dt \cdot (1/4 - 9/(64\lambda) + o(1/\lambda))$. So the variance of $N(T)^{-1/2}\sum_{t_j} f(t_j)$ is $\sum_i \mathrm{var}(X_i) = T^{-1}\int_0^T f^2(t)dt \cdot O(1/\lambda)$.

Proof of Proposition 2: By the definition of $\hat{\mu}_\phi^{(t)}(\boldsymbol{B}_k)$, we know that

$$\frac{\partial}{\partial \boldsymbol{B}_k}\left\{\sum_{n=1}^N r_{nk}^{(t)} \cdot L(S_n) \cdot \phi_\rho\left(\log \mathrm{NHP}\left(S_n \mid \boldsymbol{B}_k\right)/L(S_n) - \hat{\mu}_\phi^{(t)}(\boldsymbol{B}_k)\right)\right\} = 0,$$

which implies

$$\begin{aligned}&\frac{\partial \hat{\mu}_\phi^{(t)}(\boldsymbol{B}_k)}{\partial \boldsymbol{B}_k}\\ &= \sum_{n=1}^N \frac{r_{nk}^{(t)}\phi'_\rho\left(\log \mathrm{NHP}\left(S_n \mid \boldsymbol{B}_k\right)/L(S_n) - \hat{\mu}_\phi(\boldsymbol{B}_k^{(t-1)})\right)}{\sum_{n=1}^N r_{nk}^{(t)}\phi'_\rho\left(\log \mathrm{NHP}\left(S_n \mid \boldsymbol{B}_k\right)/L(S_n) - \hat{\mu}_\phi(\boldsymbol{B}_k^{(t-1)})\right)L(S_n)} \cdot \frac{\partial \log \mathrm{NHP}\left(S_n \mid \boldsymbol{B}_k\right)}{\partial \boldsymbol{B}_k}.\end{aligned}$$

Plugging $B_k = B_k^{(t-1)}$ into the above formula, we get the desired gradient $\varrho_k^{(t)}$. This completes the proof.

I Proof of Theorem 1

Here we would like to point out that we say the event sequence S is different from S' if their induced intensity $\hat{\lambda}_S/\sqrt{M}$'s are different. Otherwise, we treat them as the same event sequence.

Proof of Theorem 1 It is easy to know that the distance between an object and itself is always zero and the distance between distinct objects is always positive. Moreover, the distance from S_A to S_B is always the same as the distance from S_B to S_A. We only need to prove that $d(S_A, S_B)$ satisfies the triangle inequality.

By definition we know that $d(S_A, S_B) = \int_0^T \left|\hat{\lambda}_A(t)/\sqrt{M_A} - \hat{\lambda}_B(t+\delta_B)/\sqrt{M_B}\right| dt$, where $\delta_B = \arg min_{\delta_B}\int_0^T \left|\hat{\lambda}_A(t)/\sqrt{M_A} - \hat{\lambda}_B(t+\delta_B)/\sqrt{M_B}\right| dt$. In the same way we define δ_C. Then

$$\begin{aligned}d(S_B, S_C) &\le \int_0^T \left|\hat{\lambda}_C(t+\delta_C)/\sqrt{M_C} - \hat{\lambda}_B(t+\delta_B)/\sqrt{M_B}\right| dt\\ &\le \int_0^T \left|\hat{\lambda}_C(t+\delta_C)/\sqrt{M_C} - \hat{\lambda}_A(t)/\sqrt{M_A}\right| dt + \int_0^T \left|\hat{\lambda}_A(t)/\sqrt{M_A} - \hat{\lambda}_B(t+\delta_B)/\sqrt{M_B}\right| dt\\ &= d(S_A, S_B) + d(S_A, S_C).\end{aligned}$$

J Proof of Theorem 2 and Theorem 3

We first provide a lemma showing that the “outlier screening" procedure can eliminate all outliers with high probability.

Lemma 1 *Under Assumption 1 and 2, steps 3-5 in Algorithm 2 eliminate all outliers with high probability.*

Proof of Lemma 1 Without loss of generality, we consider Cluster 1. Assume that Cluster 1 accounts for π_1 proportion of the set $\mathcal{S}$. When we select M samples from an N-element set, it is easy to know that the amount of Cluster 1 obey the binomial distribution $B(M, \pi_1)$. Then the probability of α-quantile being smaller than $r_{max} := \max_{S_{n_1}, S_{n_2} \in \mathcal{S}_1} d(S_{n_1}, S_{n_2})$ is

$$p(\alpha) := \sum_k \mathbb{P}(X \in \mathcal{C}_k) \cdot \mathbb{P}(X_{dis} \geq M \cdot \alpha) = \sum_k \left(\pi_k \sum_{i \geq \alpha \cdot M} \tbinom{M}{i} \pi_k^i \cdot (1 - \pi_k)^{M-i} \right),$$

where $X_{dis} \sim B(M, \pi_k)$.

We choose a suitable α such that $p(\alpha) \geq 1 - \delta_1$, and then choose β such that $\sum_{i \geq \beta \cdot N'} \binom{N'}{i} p(\alpha)^i (1 - p(\alpha))^{N'-i} > 1 - \delta_2$, where δ_1, δ_2 are small enough positive numbers. Repeat it until we choose enough samples, and we avoid selecting outliers with a high probability.

Next we show that the proposed “inlier weighting" procedure can produce a set of good initial centers. In the following proof, we consider an arbitrary pseudo-metric d which has quasi-triangular properties, that is, $d(x, z) \leq M(d(x, y) + d(y, z))$ for all $x, y, z \in \mathcal{S}$. For our proposed distance function, it holds $M \equiv 1$.

Overview of Proof of Theorem 2. In order to find the upper bound of the Υ, we use mathematical induction to prove that the upper bound of the objective function Υ can be controlled after adding several centers. Lemma 3 proves the case of one-step addition and Lemma 4 generalizes to the general case. As defined previously, we know that under the optimal center set $\mathcal{C}_{\mathrm{OPT}}$, each sequence will be classified into the same class of an element in $\mathcal{C}_{\mathrm{OPT}}$, so we can divide $\mathcal{S}_{in}$ into K sub-sets.

Lemma 2 *Let $\mathcal{S}$ be a set of sequences, and let s_1 and s_2 be two arbitrary sequences. Then $\sum_{x \in S} d(x, s_1)^2 - 2M^2 \sum_{x \in S} d(x, s_2)^2 \leq 2M^2 |S| \cdot d(s_1, s_2)^2$.*

Lemma 3 *Let $\mathcal{A}$ be an arbitrary data set, and let $\mathcal{C}$ be an arbitrary set of centers. Define $\Upsilon_{\mathcal{C}}(\mathcal{A}) := \sum_{a \in \mathcal{A}} \min_{c \in \mathcal{C}} d(a, c)^2$, $\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{A}) := \sum_{a \in \mathcal{A}} \min_{c \in \mathcal{C}_{\mathrm{OPT}}} d(a, c)^2$. If we add a random center to $\mathcal{C}$ from $\mathcal{A}$, chosen with D^2 weighting (i.e. step (b)), then $\mathbb{E}[\Upsilon_{\mathcal{C}}(\mathcal{A})] \leq 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{A})$.*

Proof of Lemma 3 The probability that we choose some fixed a_0 as our center is precisely $D(a_0)^2 / \sum_{a \in \mathcal{A}} D(a)^2$. Furthermore, after choosing the center a_0, a sequence a will contribute precisely $\min(D(a), d(a, a_0))^2$ to the potential. Therefore,

$$\mathbb{E}[\Upsilon_{\mathcal{C}}(\mathcal{A})] = \sum_{a_0 \in \mathcal{A}} \frac{D(a_0)^2}{\sum_{a \in \mathcal{A}} D(a)^2} \sum_{a \in \mathcal{A}} \min(D(a), d(a, a_0))^2 .$$

Note by the triangle inequality that $D(a_0) \leq M(D(a) + d(a, a_0))$ for all a, a_0. From this, the powermean inequality implies that $D(a_0)^2 \leq 2M^2(D(a)^2 + d(a, a_0)^2)$. Summing over all a, we then have that $D(a_0)^2 \leq \frac{2M^2}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} D(a)^2 + \frac{2M^2}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} d(a, a_0)^2$. Then $\mathbb{E}[\Upsilon_{\mathcal{C}}(\mathcal{A})]$ is at most

$$r \frac{2M^2}{|\mathcal{A}|} \cdot \sum_{a_0 \in \mathcal{A}} \frac{\sum_{a \in \mathcal{A}} D(a)^2}{\sum_{a \in \mathcal{A}} D(a)^2} \cdot \sum_{a \in \mathcal{A}} \min(D(a), d(a, a_0))^2$$
$$+ \frac{2M^2}{|\mathcal{A}|} \cdot \sum_{a_0 \in \mathcal{A}} \frac{\sum_{a \in \mathcal{A}} d(a, a_0)^2}{\sum_{a \in \mathcal{A}} D(a)^2} \cdot \sum_{a \in \mathcal{A}} \min(D(a), d(a, a_0))^2 .$$

In the first expression, we substitute $\min\left(D(a), d(a,a_0)\right)^2 \leq d(a,a_0)^2$, and in the second expression, we substitute $\min\left(D(a), d(a,a_0)\right)^2 \leq D(a)^2$. Simplifying, we then have,

$$\mathbb{E}[\Upsilon_{\mathcal{C}}(\mathcal{A})] \leq \frac{4M^2}{|\mathcal{A}|} \cdot \sum_{a_0 \in \mathcal{A}} \sum_{a \in \mathcal{A}} d(a,a_0)^2 \leq 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{A}).$$

The last step here follows from Lemma 2.

Lemma 4 *Let $\mathcal{C}$ be the current center set, when we choose $u > 0$ "uncovered" class, and let $\mathcal{S}_u$ denote the set of sequences in these class. Also let $\mathcal{S}_c = \mathcal{S} - \mathcal{S}_u$. Now suppose we add $t \leq u$ random centers to $\mathcal{C}$, chosen with D^2 weighting. Let $\mathcal{C}'$ denote the new center set, and let $\Upsilon_{\mathcal{C}'}(\mathcal{S})$ denote the corresponding potential. Then $\mathbb{E}\left[\Upsilon_{\mathcal{C}'}(\mathcal{S})\right]$ is at most*

$$\left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\right) \cdot (1 + H_t) + \frac{u-t}{u} \cdot \Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right).$$

Here H_t denotes the harmonic sum, $1 + 1/2 + \cdots + 1/t$.

Proof of Lemma 4

We prove the conclusion by induction, showing that if the result holds for $(t-1, u)$ and $(t-1, u-1)$, then it also holds for (t, u). Therefore, it suffices to check $t = 0, u > 0$ and $t = u = 1$ as our base cases.

If $t = 0$ and $u > 0$, the result follows from the fact that $1 + H_t = (u-t)/u = 1$. Next, suppose $t = u = 1$. We choose our one new center from one uncovered class with probability exactly $\Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)/\Upsilon_{\mathcal{C}}(\mathcal{S})$. In this case, Lemma 3 guarantees that $\mathbb{E}\left[\Upsilon_{\mathcal{C}'}(\mathcal{S})\right] \leq \Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)$. Since $\Upsilon_{\mathcal{C}'}(\mathcal{S}) \leq \Upsilon_{\mathcal{C}}(\mathcal{S})$, even if we choose a center from a covered class, we have

$$\begin{aligned}\mathbb{E}\left[\Upsilon_{\mathcal{C}'}(\mathcal{S})\right] &\leq \frac{\Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)}{\Upsilon} \cdot \left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\right) + \frac{\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right)}{\Upsilon_{\mathcal{C}}(\mathcal{S})} \cdot \Upsilon_{\mathcal{C}}(\mathcal{S}) \\ &\leq 2\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\end{aligned}$$

Since $1 + H_t = 2$ here, we have shown the result holds for both base cases.

We now proceed to prove the inductive step. It is convenient here to consider two cases. First, suppose we choose our first center from a covered class. As above, this happens with probability exactly $\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right)/\Upsilon_{\mathcal{C}}(\mathcal{S})$. Note that this new center can only decrease $\Upsilon_{\mathcal{C}}(\mathcal{S})$. We apply the inductive hypothesis with the same choice of covered class, but with t decreased by 1. It follows that our contribution to $\mathbb{E}\left[\Upsilon_{\mathcal{C}'}(\mathcal{S})\right]$ in this case is at most,

$$\frac{\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right)}{\Upsilon_{\mathcal{C}}(\mathcal{S})} \cdot \left(\left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\right) \cdot (1 + H_{t-1}) + \frac{u-t+1}{u} \cdot \Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)\right).$$

On the other hand, suppose we choose our first center from some uncovered class $\mathcal{A}$. This happens with probability $\Upsilon_{\mathcal{C}}\left(\mathcal{A}\right)/\Upsilon_{\mathcal{C}}(\mathcal{S})$. Let p_a denote the probability that we choose $a \in \mathcal{A}$ as our center, given the center is somewhere in $\mathcal{A}$, and let Υ_a denote $\Upsilon_{\mathcal{C}}\left(\mathcal{A}\right)$ after we choose a as our center. Once again we apply our inductive hypothesis, as well as decrease both t and u by 1. It follows that our contribution to $\mathbb{E}\left[\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\right]$ in this case is at most,

$$\begin{aligned}&\frac{\Upsilon_{\mathcal{C}}\left(\mathcal{A}\right)}{\Upsilon_{\mathcal{C}}(\mathcal{S})} \cdot \sum_{a \in \mathcal{A}} p_a \{\left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + \Upsilon_a + 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right) - 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{A})\right) \cdot (1 + H_{t-1}) \\ &+ \frac{u-t}{u-1} \cdot \left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right) - \Upsilon_{\mathcal{C}}\left(\mathcal{A}\right)\right)\} \\ &\leq \frac{\Upsilon_{\mathcal{C}}\left(\mathcal{A}\right)}{\Upsilon_{\mathcal{C}}(\mathcal{S})} \cdot \left(\left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\right) \cdot (1 + H_{t-1}) + \frac{u-t}{u-1} \cdot \left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right) - \Upsilon_{\mathcal{C}}\left(\mathcal{A}\right)\right)\right).\end{aligned}$$

The last step here follows from the fact that $\sum_{a \in \mathcal{A}} p_a \Upsilon_a \leq 16M^4 \Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{A})$, which is implied by Lemma 3.

Now, the power-mean inequality implies that $\sum_{\mathcal{A}\subset\mathcal{S}_u} \Upsilon_{\mathcal{C}}\left(\mathcal{A}\right)^2 \geq \Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)^2/u$. Therefore, if we sum over all uncovered class $\mathcal{A}$, we obtain a contribution at most,

$$\begin{aligned}
&\frac{\Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)}{\Upsilon_{\mathcal{C}}(\mathcal{S})} \cdot \left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\right) \cdot \left(1 + H_{t-1}\right) + \frac{1}{\Upsilon} \cdot \frac{u-t}{u-1} \cdot \left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)^2 - \frac{1}{u} \cdot \Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)^2\right) \\
&= \frac{\Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)}{\Upsilon_{\mathcal{C}}(\mathcal{S})} \cdot \left(\left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\right) \cdot \left(1 + H_{t-1}\right) + \frac{u-t}{u} \cdot \Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)\right).
\end{aligned}$$

Combining the potential contribution to $\mathbb{E}\left[\Upsilon_{\mathcal{C}'}(\mathcal{S})\right]$ from both cases, we now obtain the desired bound:

$$\begin{aligned}
\mathbb{E}\left[\Upsilon_{\mathcal{C}'}(\mathcal{S})\right] &\leq \left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\right) \cdot \left(1 + H_{t-1}\right) + \frac{u-t}{u} \cdot \Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right) + \frac{\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right)}{\Upsilon_{\mathcal{C}}(\mathcal{S})} \cdot \frac{\Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right)}{u} \\
&\leq \left(\Upsilon_{\mathcal{C}}\left(\mathcal{S}_c\right) + 16M^4\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}\left(\mathcal{S}_u\right)\right) \cdot \left(1 + H_{t-1} + \frac{1}{u}\right) + \frac{u-t}{u} \cdot \Upsilon_{\mathcal{C}}\left(\mathcal{S}_u\right).
\end{aligned}$$

The inductive step now follows from the fact that $1/n \leq 1/t$.

Proof of Theorem 2 Consider the clustering $\mathcal{C}^{ini}$ after we have completed Step 1. Applying Lemma 4 with $\mathcal{S} = \mathcal{S}_{in}$, $t = u = k-1$ and with $\mathcal{S}_c$ being the first covered class, we have,

$$\begin{aligned}
\mathbb{E}[\Upsilon_{\mathcal{C}^{ini}}(\mathcal{S}_{in}) \mid \mathcal{S}_{in}] &= \mathbb{E}[\mathbb{E}[\Upsilon_{\mathcal{C}^{ini}}(\mathcal{S}_{in}) \mid \mathcal{S}_c] \mid \mathcal{S}_{in}] \\
&\leq \mathbb{E}[\left(\Upsilon(\mathcal{S}_c) + 16M^4\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{S}_u)\right) \cdot (1 + H_{k-1}) \mid \mathcal{S}_{in}] \\
&= \mathbb{E}[\left(\Upsilon(\mathcal{S}_c) + 16M^4\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{S}_{in}) - 16M^4\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{S}_c)\right) \cdot (1 + H_{k-1}) \mid \mathcal{S}_{in}] \\
&\leq 16M^4(\ln K + 2)\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{S}_{in}).
\end{aligned}$$

The result now follows from Lemma 3, and from the fact that $H_{k-1} \leq 1 + \ln k$.

Proof of Theorem 3 By Assumption 2, we know that there are at least α proportion of samples here that are not classified into the correct class. Denote the correctly classified set as $\mathcal{S}_{right}$, and the incorrectly classified set as $\mathcal{S}_{wrong}$. Then

$$\Upsilon_{\mathcal{C}_{lack}}(\mathcal{S}_{in}) = \sum_{x\in\mathcal{S}_{wrong}} \min_{c\in\mathcal{C}_{lack}} d\left(x,c\right)^2 + \sum_{x\in\mathcal{S}_{right}} \min_{c\in\mathcal{C}_{lack}} d\left(x,c\right)^2. \tag{11}$$

We consider the part $\mathcal{S}_{right}$ first, we know that for each sample, there is an estimated function of cubic spline approximation, which is $\hat{\lambda}(t) = \sum_{h=1}^{H} b_h\kappa_h(t)$. When sequences x and c are generated from the same class, the distance between them is $d(x,c) = \int_0^T \left|\hat{\lambda_x}(t)/\sqrt{M_x} - \hat{\lambda_c}(t)/\sqrt{M_c}\right| dt \leq \sum_{h=1}^{H} \left|b_h^x/\sqrt{M_x} - b_h^c/\sqrt{M_c}\right| \int_0^T \kappa_h(t)dt$. Thus we know $d(x,c) \sim O(L^{-1/2})$. As $L \to \infty$, we get that $\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{S}_{in})/\Upsilon_{\mathcal{C}_{lack}}(\mathcal{S}_{in}) \sim O(L^{-1/2})$. Therefore, $\Upsilon_{\mathcal{C}_{lack}}(\mathcal{S}_{in}) > 16(\log K + 2)\Upsilon_{\mathcal{C}_{\mathrm{OPT}}}(\mathcal{S}_{in})$ with high probability whenever L is sufficiently larger than $(\log K)^2$.

K Proof of Theorem 4 and Theorem 5

We first provide several supporting results regarding the properties of Poisson random variables and Poisson processes.

Let $h : [-1,\infty) \to \mathbb{R}$ be the function defined by $h(u) := 2[(1+u)\ln(1+u) - u]/u^2$.

Lemma 5 *Let* $X \sim \mathrm{Poisson}(\lambda)$ *with* $\lambda > 0$. *Then, for any* $x > 0$, *we have*

$$\mathbb{P}\left(X \geq \lambda + x\right) \leq \exp\left(-\frac{x^2}{2\lambda}h\left(\frac{x}{\lambda}\right)\right)$$

and, for any $0 < x < \lambda$,

$$\mathbb{P}\left(X \leq \lambda - x\right) \leq \exp\left(-\frac{x^2}{2\lambda}h\left(-\frac{x}{\lambda}\right)\right).$$

In particular, this implies that $\mathbb{P}\left(X \geq \lambda + x\right)$ *and* $\mathbb{P}\left(X \leq \lambda - x\right) \leq \exp\left(-x^2/(2(\lambda + x))\right)$, *for* $x > 0$; *from which*

$$\mathbb{P}\left(|X - \lambda| \geq x\right) \leq 2\exp\left(-\frac{x^2}{2(\lambda + x)}\right), \quad x > 0.$$

Proof of Lemma 5 Recall that if $\left(Y^{(n)}\right)_{n\geq 1}$ is a sequence of independent random variables such that $Y^{(n)}$ follows a Binomial $(n, \lambda/n)$ distribution, then $\left(Y^{(n)}\right)_{n\geq 1}$ converges in law to X, a random variable with Poisson (λ) distribution. In particular, since convergence in law corresponds to pointwise convergence of distribution functions, this implies that, for any $t \in \mathbb{R}$,

$$\mathbb{P}\left(Y^{(n)} \geq t\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(X \geq t\right).$$

For any fixed $n \geq 1$, by the definition, we can write $Y^{(n)}$ as $Y^{(n)} = \sum_{k=1}^{n} Y_k^{(n)}$, where $Y_1^{(n)}, \ldots, Y_n^{(n)}$ are i.i.d. random variables with Bernoulli (λ/n) distribution. Note that $\mathbb{E}\left[Y^{(n)}\right] = \lambda$ and $\operatorname{Var}\left[Y^{(n)}\right] = \lambda\left(1 - \lambda/n\right) \leq \lambda$. As $\mathbb{E}\left[Y_k^{(n)}\right] = \lambda/n$ and $\left|Y_k^{(n)}\right| \leq 1$ for all $1 \leq k \leq n$, we can apply Bennett's inequality [Boucheron et al., 2013], to obtain, for any $t \geq 0$,

$$\mathbb{P}\left(Y^{(n)} \geq \lambda + x\right) = \mathbb{P}\left(Y^{(n)} \geq \mathbb{E}\left[Y^{(n)}\right] + x\right) \leq \exp\left(-\frac{x^2}{2\lambda} h\left(\frac{x}{\lambda}\right)\right).$$

Taking the limit as n goes to infinity, we obtain that $\mathbb{P}\left(X \geq \lambda + x\right) \leq \exp\left(-x^2 h\left(x/\lambda\right)/(2\lambda)\right)$.

Lemma 6 (Bernstein's inequality [Vershynin, 2018]) *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right],$$

where $K = \max_i \|X_i\|_{\psi_1}$ and $\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \leq 2\}$.

Lemma 7 *When event sequence S is sampled from the NHP process with parameter λ_*, its log-likelihood function $\log \mathrm{NHP}(S \mid \boldsymbol{B}_i)$ follows a sub-exponential distribution.*

Proof of Lemma 7 Divide the interval $[0, T]$ into $\mathcal{M}$ small intervals $[a_0, a_1], \cdots, [a_{\mathcal{M}-1}, a_{\mathcal{M}}]$, where $0 = a_0 < a_1 < \cdots < a_{\mathcal{M}} = T$. Within the small interval $[a_i, a_{i+1}]$, there is approximately a homogeneous Poisson process with intensity $\lambda(a_i + \eta)$, where $\eta < a_{i+1} - a_i$. At this point we can divide the log-likelihood function into $\mathcal{M}$ parts $F_1, \cdots, F_{\mathcal{M}}$, where $F_\ell := \sum_{t_i \in [a_{\ell-1}, a_\ell]} \log(t_i)$. At this time $F_\ell / log(a_i + \eta)$ approximately obeys the homogeneous Poisson process with the parameter $\lambda(a_i + \eta) \cdot (a_{i+1} - a_i)$, so its variance is $\lambda(a_i + \eta)(a_{i+1} - a_i) \cdot \log(\lambda(a_i + \eta))^2$. According to Lemma 5, each of F_ℓ follows a sub-exponential distribution. Using Lemma 6, we know that

$$\mathbb{P}\left(|\log \mathrm{NHP}(S \mid \boldsymbol{B}_i)/L(S) - \mu_{avg}| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{L(S)^2 t^2}{C^2 \max\log(\lambda_*)^2}, \frac{L(S)t}{C\max\log(\lambda_*)}\right)\right],$$

where C is a finite constant depend on $\boldsymbol{B}_i$ and $\mu_{avg} := \mathbb{E}_{S\sim\lambda_*} \log \mathrm{NHP}(S \mid \boldsymbol{B}_i)/L(S)$.

Similar to the derivative function of $\log \mathrm{NHP}(S \mid \boldsymbol{B}_i)$, there is

$$\mathbb{P}\left(\left|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_i)}{\partial \boldsymbol{B}_i} / L(S) - \mu_{avg}\right| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{L(S)^2 t^2}{C^2(\max\frac{\kappa_{\max}}{\lambda_*(t)})^2}, \frac{L(S)t}{C\max\frac{\kappa_{\max}}{\lambda_*(t)}}\right)\right].$$

Corollary 1 *According to proposition 2.7.1 from [Vershynin, 2018], $m(S)$ follow a sub-exponential distribution. From Lemma 6, we know that for $m(S)$ with L periods, it follows a sub-exponential distribution as well, and $\mathbb{P}\left(\left|m(S)/L - \int_0^T \lambda_*(t)dt\right| > t\right) \leq 2\exp(-K_0 L t)$. Take a small enough $\delta > 0$, we have $\mathbb{P}\left(m(S)/L > m_c\right) < \delta$ when $m_c \geq \int_0^T \lambda_*(t)dt + \log(2/\delta)/(L \cdot K_0)$. Define $C_0 := m_c \cdot L$, which can be viewed as the high probability bound of number of events in event sequence S.*

Overview of Proof Theorem 4. In order to prove the local convergence property of the proposed algorithm, we need to check the following three key important aspects. (i) What is the difference $|\mu(\boldsymbol{B}_k|\boldsymbol{B}'_k) - \mu(\boldsymbol{B}_k|\boldsymbol{B}''_k)|$

when $\boldsymbol{B}_k'$ and $\boldsymbol{B}_k''$ are close; see Lemma 13. (ii) What is the difference between sample gradient $\varrho_k^{(t)}$ and population gradient $\nabla\mu(\boldsymbol{B}_k|\boldsymbol{B}_k^{(t-1)})$ (“∇" stands for the derivative with respect to parameter $\boldsymbol{B}_k$); see Lemma 14. (iii) The local concavity of $\mu(\boldsymbol{B}_k|\boldsymbol{B}_k^{(t)})$ holds around $\boldsymbol{B}_k = \boldsymbol{B}_k^*$; see Lemma 12.

Define the weight $w_k(S;\mathbf{B}) = \pi_k \,\mathrm{NHP}(S \mid \boldsymbol{B}_k)/\sum_j \pi_j \,\mathrm{NHP}(S \mid \boldsymbol{B}_j)$ for $k \in [K]$.

Lemma 8 *If* $\|\boldsymbol{B}_k - \boldsymbol{B}_k^*\|_2 < a/(T \cdot \kappa_{\max})$ *for* $\forall k \in [K]$*, there exists a constant* $G > 0$ *such that*

$$\mathbb{E}_S \left[w_k(S;\mathbf{B})\left(1 - w_k(S;\mathbf{B})\right) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_k)}{\partial \boldsymbol{B}_k} \right\|^p \right] \sim O(L(S)^p \exp(-G \cdot L(S)))$$

for $p = 1, 2$.

Proof of Lemma 8 Without loss of generality, we prove the claim for $k = 1$. Taking the expectation of S, we get

$$\begin{aligned}
&\mathbb{E}_S \left[w_1(S;\mathbf{B})\left(1 - w_1(S;\mathbf{B})\right) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1} \right\|^p \right] \\
=& \sum_{k\in[K]} \pi_k \mathbb{E}_{s\sim\mathcal{POI}(\boldsymbol{B}_k^*)} \left[w_1(S;\mathbf{B})\left(1 - w_1(S;\mathbf{B})\right) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1} \right\|^p \right] \\
\leq& \pi_1 \mathbb{E}_{s\sim\mathcal{POI}(\boldsymbol{B}_1^*)} \left[w_1(S;\mathbf{B})\left(1 - w_1(S;\mathbf{B})\right) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1} \right\|^p \right] \\
&+ \sum_{k\neq 1} \pi_k \mathbb{E}_{s\sim\mathcal{POI}(\boldsymbol{B}_k^*)} \left[w_1(S;\mathbf{B})\left(1 - w_1(S;\mathbf{B})\right) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1} \right\|^p \right].
\end{aligned}$$

For the the first term, we define event $\mathcal{E}_r^{(1)} = \{S : S \sim \mathcal{POI}\left(\boldsymbol{B}_1^*\right); \|\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1^*)/\partial \boldsymbol{B}_1\| \leq r \cdot L(S)\}$ for some $r > 0$. According to the assumption that $\|\boldsymbol{B}_1 - \boldsymbol{B}_1^*\| \leq a/(T * \kappa_{\max})$, we know that $\max\left|\lambda_{\boldsymbol{B}_1}(s) - \lambda_{\boldsymbol{B}_1^*}(s)\right| \leq a/T$. Then for $S \in \mathcal{E}_r^{(1)}$, using triangle inequality, we have

$$\begin{aligned}
&\left| \sum_t^{m(S)} \frac{\kappa_h(s_t)}{\lambda_{\boldsymbol{B}_1}(s_t)} - \int_0^T \kappa_h(x)dx \right| \\
\leq& \left| \sum_t^{m(S)} \frac{\kappa_h(s_t)}{\lambda_{\boldsymbol{B}_1^*}(s_t)} - \int_0^T \kappa_h(x)dx \right| + \left| \sum_t^{m(S)} \kappa_h(s_t)\left(\frac{1}{\lambda_{\boldsymbol{B}_1}(s_t)} - \frac{1}{\lambda_{\boldsymbol{B}_1^*}(s_t)}\right) \right| \\
\leq& L(S) \cdot r + \frac{m(S)a}{T\tau^2},\ \forall h \in \{1, \cdots, H\}.
\end{aligned}$$

Because $\left|\lambda_{\boldsymbol{B}_k}(t) - \lambda_{\boldsymbol{B}_k^*}(t)\right| < a/T$ for $k = 1, 2, \ldots, K$, then we have $\log \mathrm{NHP}(S \mid \boldsymbol{B}_1) = \sum_i \log \lambda_{\boldsymbol{B}_1}(t_i) - \int \lambda_{\boldsymbol{B}_1}(s)ds \geq \log \mathrm{NHP}(S \mid \boldsymbol{B}_1^*) - m(S)\log\left((\tau + a/T)/\tau\right) - a \cdot L(S)$.

For $k \neq 1$, $\log \mathrm{NHP}(S \mid \boldsymbol{B}_k) - \log \mathrm{NHP}(S \mid \boldsymbol{B}_k^*) = \sum_k \log\left(\lambda_{\boldsymbol{B}_k}(t_i)/\lambda_{\boldsymbol{B}_k^*}(t_i)\right) - \int(\lambda_{\boldsymbol{B}_k}(s) - \lambda_{\boldsymbol{B}_k^*}(s))ds \leq a \cdot L(S) + m(S)\log\left((\tau + a/T)/\tau\right)$. By Assumption 5, we know that $\log \mathrm{NHP}(S \mid \boldsymbol{B}_k) \leq \log \mathrm{NHP}(S \mid \boldsymbol{B}_1^*) - C \cdot L(S) + a \cdot L(S) + m(S)\log\left((\tau + a/T(S))/\tau\right)$. Then we get that

$$\begin{aligned}
&\mathbb{E}_S \left[w_1(S;\mathbf{B})\left(1 - w_1(S;\mathbf{B})\right) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1} \right\|^p \mid \mathcal{E}_r^{(1)} \right] \\
\leq& \frac{1-\pi_1}{\pi_1} \exp\left(2a \cdot L(S) + 2m(S)\log\left(\frac{\tau + a/T(S)}{\tau}\right) - C \cdot L(S) \right) \cdot \left(rL(S) + \frac{a}{\tau^2}\frac{m(S)}{T(S)} \right)^p.
\end{aligned}$$

For $\mathcal{E}_r^c$ part, we now have $\|\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)/\partial \boldsymbol{B}_1\| > r \cdot L(S)$. We define

$$\begin{aligned}
M_h :=& \int_0^{L(S)*T} \frac{\kappa_h(t)}{\lambda_{\boldsymbol{B}_1}(t)} dN(t) - \int_0^{L(S)*T} \kappa_h(x)dx \\
=& \sum_{l=1}^{L(S)} \int_{(l-1)*T}^{l*T} \frac{\kappa_h(t)}{\lambda_{\boldsymbol{B}_1}(t)} dN(t) - \int_{(l-1)*T}^{l*T(S)} \kappa_h(x)dx \\
=& \sum_{l=1}^{L(S)} X_l,
\end{aligned}$$

where X_l's are independent. According to Lemma 7, there exists $c_0 > 0$ such that

$$\mathbb{P}\left(|M_h/L(S)| \geq t\right) \leq 2\exp\left(-\frac{tL(S)}{c_0}\right).$$

Obviously we have $w_1(S;\mathbf{B})\left(1-w_1(S;\mathbf{B})\right) \leq 1/4$. Then

$$\begin{aligned}
&\mathbb{E}_S\left[w_1(S;\mathbf{B})\left(1-w_1(S;\mathbf{B})\right)\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\|^p \mid \mathcal{E}_r^c\right] \\
&\leq \frac{1}{4}\int_r^\infty t^p d\mathbb{P}\left(\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\| \geq t \cdot L(S)\right) \\
&= \frac{1}{4}\big(r^p \cdot L(S)\mathbb{P}\left(\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\| \geq r \cdot L(S)\right) \\
&+ \int_r^\infty pt^{p-1}\mathbb{P}\left(\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\| \geq t \cdot L(S)\right)dt\big) \\
&\leq \frac{1}{2}\left(r^p L(S)\exp\left(-\frac{rL(S)}{c_0}\right) + \int_r^\infty pt^{p-1}\exp\left(-\frac{tL(S)}{c_0}\right)dt\right).
\end{aligned}$$

For fixed $r \geq 0$, when $L(S) \to \infty$, it is easy to know that

$$\frac{1}{2}\left(r^p L(S)\exp\left(-\frac{rL(S)}{c_0}\right) + \int_r^\infty pt^{p-1}\exp\left(-\frac{tL(S)}{c_0}\right)dt\right) \to 0.$$

Next we consider the remainder of the gradient. For $k \neq 1$,

$$\begin{aligned}
&\pi_k\mathbb{E}_{s\sim\mathcal{POI}(\boldsymbol{B}_k^*)}\left[w_1(S;\mathbf{B})\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\|^p\right] \\
&= \underbrace{\int_{\left\|\frac{\partial \log \mathrm{NHP}(\mathbf{S}|\boldsymbol{B}_k^*)}{\partial \boldsymbol{B}_k}\right\| < r\cdot L(S)} \frac{\pi_1 \mathrm{NHP}(S \mid \boldsymbol{B}_1)\pi_k \mathrm{NHP}(S \mid \boldsymbol{B}_k^*)}{\sum_k \pi_k \mathrm{NHP}(S \mid \boldsymbol{B}_k)}\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\|^p dS}_{I_1} \\
&+ \underbrace{\int_{\left\|\frac{\partial \log \mathrm{NHP}(\mathbf{S}|\boldsymbol{B}_i^*)}{\partial \boldsymbol{B}_i}\right\| > r\cdot L(S)} \frac{\pi_1 \mathrm{NHP}(S \mid \boldsymbol{B}_1)\pi_k \mathrm{NHP}(S \mid \boldsymbol{B}_k^*)}{\sum_k \pi_k \mathrm{NHP}(S \mid \boldsymbol{B}_k)}\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\|^p dS}_{I_2}.
\end{aligned}$$

When $\|\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_k^*)/\partial \boldsymbol{B}_k\| < r \cdot L(S)$, we have

$$\frac{\mathrm{NHP}(S \mid \boldsymbol{B}_k)}{\mathrm{NHP}(S \mid \boldsymbol{B}_k^*)} \leq \exp\left(a \cdot L(S) + m(S)\log\left(\frac{\tau + a/T}{\tau}\right)\right)$$

and

$$\frac{\mathrm{NHP}(S \mid \boldsymbol{B}_k^*)}{\mathrm{NHP}(S \mid \boldsymbol{B}_k)} \leq \exp\left(a \cdot L(S) + m(S)\log\left(\frac{\tau + a/T}{\tau}\right)\right).$$

Then it holds

$$\begin{aligned}
&I_1 \leq \frac{\pi_k \mathrm{NHP}(S \mid \boldsymbol{B}_k^*)}{\pi_i \mathrm{NHP}(S \mid \boldsymbol{B}_k)} \cdot \int_{\left\|\frac{\partial \log \mathrm{NHP}(\mathbf{S}|\boldsymbol{B}_k^*)}{\partial \boldsymbol{B}_k}\right\| < rL(S)} \pi_1 \mathrm{NHP}(S \mid \boldsymbol{B}_1)\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\|^p dS \\
&\leq \pi_1 \exp\left(aL(S) + m(S)\log(\frac{\tau + a/T}{\tau})\right)\int_{\left\|\frac{\partial \log \mathrm{NHP}(\mathbf{S}|\boldsymbol{B}_k^*)}{\partial \boldsymbol{B}_k}\right\| < rL(S)} \mathrm{NHP}(S \mid \boldsymbol{B}_1)\left\|\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1}\right\|^p dS \\
&\leq \pi_1 \exp\left(aL(S) + m(S)\log(\frac{\tau + a/T}{\tau})\right) \\
&\cdot \int_{\left\|\frac{\partial \log \mathrm{NHP}(\mathbf{S}|\boldsymbol{B}_k^*)}{\partial \boldsymbol{B}_k}\right\| < rL(S)} \mathrm{NHP}(S \mid \boldsymbol{B}_k^*)\exp\left(-CL(S) + 2aL(S) + 2m(S)\log(\frac{\tau + a/T}{\tau})\right)(C_0L(S))^p dS \\
&\leq \pi_1 \exp\left(-CL(S) + 2aL(S) + 2m(S)\log(\frac{\tau + a/T}{\tau})\right) \cdot (C_0L(S))^p,
\end{aligned}$$

where C_0 is the upper bound of $\|\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_k)/\partial \boldsymbol{B}_k\|, \forall k = 1, \cdots, K$ with probability of $1 - \delta$.

When $\|\partial \log \mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_k^*)/\partial \boldsymbol{B}_k\| > r \cdot L(S)$ and $L(S) \to \infty$, it holds

$$\begin{aligned}
I_2 &= \frac{\pi_1 \, \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\sum_k \pi_k \, \mathrm{NHP}(S \mid \boldsymbol{B}_k)} \cdot \int_{\left\| \frac{\partial \log \mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_k^*)}{\partial \boldsymbol{B}_k} \right\| > r \cdot L(S)} \pi_k \, \mathrm{NHP}(S \mid \boldsymbol{B}_k^*) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1} \right\|^p dS \\
&\le \int_{\left\| \frac{\partial \log \mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_k^*)}{\partial \boldsymbol{B}_k} \right\| > r \cdot L(S)} \pi_k \, \mathrm{NHP}(S \mid \boldsymbol{B}_k^*) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1} \right\|^p dS \\
&\le \pi_i (C_0 L(S))^p \int_{\left\| \frac{\partial \log \mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_k^*)}{\partial \boldsymbol{B}_k} \right\| > r \cdot L(S)} \mathrm{NHP}(S \mid \boldsymbol{B}_k^*) dS \\
&\le 2\pi_i (C_0 L(S))^p \exp\left(-\frac{tL(S)}{c_0} \right) dS,
\end{aligned}$$

where we use the same conclusion obtained above that $\mathbb{P}\left(\|\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_k^*)/\partial \boldsymbol{B}_k\| / L(S) \ge t \right) \le 2\exp\left(-tL(S)/c_0\right)$. We take $G = \min\{C_{gap} - 2a - 2m_c \log((\tau + a/T(S))/\tau), t/c_0\}$, where $\mathbb{P}\left(|M(S)/L(S)| \ge m_c\right) < \delta$ for small enough $\delta > 0$. Thus we get the result.

Lemma 9 *If $\|\boldsymbol{B}_k - \boldsymbol{B}_k^*\| < a/(T \cdot \kappa_{\max})$ for $\forall k \in [K]$, then it holds*

$$\|\nabla w_k(S, \mathbf{B})\| \sim O(\sqrt{H} L(S) \exp(-G \cdot L(S))).$$

Proof of Lemma 9 Without loss of generality, we prove the lemma for $k = 1$. Recall the definition of $w_1(S; \mathbf{B})$, for any given S, consider the function $\mathbf{B} \to w_1(S; \mathbf{B})$, it is easy to know that

$$\nabla w_1(S; \mathbf{B}) = \begin{pmatrix} -w_1(S; \mathbf{B}) \left(1 - w_1(S; \mathbf{B})\right) \dfrac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_1)}{\partial \boldsymbol{B}_1} \\ w_1(S; \mathbf{B}) w_2(S; \mathbf{B}) \dfrac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_2)}{\partial \boldsymbol{B}_2} \\ \vdots \\ w_1(S; \mathbf{B}) w_K(S; \mathbf{B}) \dfrac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_K)}{\partial \boldsymbol{B}_K} \end{pmatrix},$$

where

$$\frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_k)}{\partial \boldsymbol{B}_k} = \begin{pmatrix} \displaystyle\sum_t^{m(S)} \frac{\kappa_1(s_t)}{\lambda_{\boldsymbol{B}_k}(s_t)} - \int_0^T \kappa_1(x) dx \\ \vdots \\ \displaystyle\sum_t^{m(S)} \frac{\kappa_H(s_t)}{\lambda_{\boldsymbol{B}_k}(s_t)} - \int_0^T \kappa_H(x) dx \end{pmatrix}^\top.$$

To calculate the upper bound of $\|\nabla w_i(S, \mathbf{B})\|$, we start by considering the first line. By Lemma 8, it is easy to know that the first line is of order $O(L(S) \exp(-G \cdot L(S)))$. Then we turn to other lines. Note that

$$\mathbb{E}_S \left[w_1(S; \mathbf{B}) w_i(S; \mathbf{B}) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_i)}{\partial \boldsymbol{B}_i} \right\| \right] \le \mathbb{E}_S \left[w_i(S; \mathbf{B}) \left(1 - w_i(S; \mathbf{B})\right) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_i)}{\partial \boldsymbol{B}_i} \right\| \right]$$

for $\forall i \neq 1$. Therefore the upper bound of line i has the same order as that of line 1.

Lemma 10 *If $\|\boldsymbol{B}_k - \boldsymbol{B}_k^*\| < a/(T \cdot \kappa_{\max})$, $\forall k \in [K]$. Then $\forall i, j \in [K]$, we have*

$$\mathbb{E}_S \left[w_i(S; \mathbf{B}) w_j(S; \mathbf{B}) \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_i)}{\partial \boldsymbol{B}_i} \right\| \cdot \left\| \frac{\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_j)}{\partial \boldsymbol{B}_j} \right\| \right] \sim O(L(S)^2 \exp(-G \cdot L(S))).$$

Proof of Lemma 10 Taking the expectation with respect to S, we get

$$
\begin{aligned}
&\mathbb{E}_S\left[w_i(S;\mathbf{B})w_j(S;\mathbf{B})\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i)}{\partial\boldsymbol{B}_i}\right\|\cdot\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_j)}{\partial\boldsymbol{B}_j}\right\|\right]\\
\leq&\mathbb{E}_S\left[w_i(S;\mathbf{B})w_j(S;\mathbf{B})\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i)}{\partial\boldsymbol{B}_i}\right\|\cdot\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_j)}{\partial\boldsymbol{B}_j}\right\|\mid\mathcal{E}_0\right]\mathbb{P}(\mathcal{E}_0)\\
&+\sum_k\pi_k\mathbb{E}_{s\sim\mathcal{POI}(\boldsymbol{B}_k^*)}\left[w_i(S;\mathbf{B})w_j(S;\boldsymbol{B})\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i)}{\partial\boldsymbol{B}_i}\right\|\cdot\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_j)}{\partial\boldsymbol{B}_j}\right\|\mid\right.\\
&\left.\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_k)}{\partial\boldsymbol{B}_k}\right\|\leq r\right]\\
&:=I_0+\sum_k I_k.
\end{aligned}
$$

Next we consider the remainder of the gradient. When $\|\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_k^*)/\partial\boldsymbol{B}_k\|<r\cdot L(S)$, we have $\mathrm{NHP}(S\mid\boldsymbol{B}_k^*)/\mathrm{NHP}(S\mid\boldsymbol{B}_k)\leq\exp\left(a\cdot L(S)+m(S)\log((\tau+a/T)/\tau)\right)$. Then for I_k,

$$
\begin{aligned}
I_k=&\int_S\frac{\pi_i\,\mathrm{NHP}(S\mid\boldsymbol{B}_i)\pi_j\,\mathrm{NHP}(S\mid\boldsymbol{B}_j)\pi_k\,\mathrm{NHP}(S\mid\boldsymbol{B}_k^*)}{(\sum_j\pi_j\,\mathrm{NHP}(S\mid\boldsymbol{B}_j))^2}\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i)}{\partial\boldsymbol{B}_i}\right\|\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_j)}{\partial\boldsymbol{B}_j}\right\|dS\\
\leq&\int_S\frac{\pi_i\,\mathrm{NHP}(S\mid\boldsymbol{B}_i)\pi_j\,\mathrm{NHP}(S\mid\boldsymbol{B}_j)\pi_k\,\mathrm{NHP}(S\mid\boldsymbol{B}_k)\exp\left(aL(S)+m(S)\log(\frac{\tau+a/T}{\tau})\right)}{(\sum_j\pi_j\,\mathrm{NHP}(S\mid\boldsymbol{B}_j))^2}\\
&\cdot\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i)}{\partial\boldsymbol{B}_i}\right\|\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_j)}{\partial\boldsymbol{B}_j}\right\|dS.
\end{aligned}
$$

Because $i\neq j$, it is easy to know that at least one of i,j is not equal to k. Without loss of generality, assume that $i\neq k$, we have

$$
\begin{aligned}
I_k=&\pi_i\frac{\pi_j\,\mathrm{NHP}(S\mid\boldsymbol{B}_j)\pi_k\,\mathrm{NHP}(S\mid\boldsymbol{B}_k)\exp\left(aL(S)+m(S)\log(\frac{\tau+a/T}{\tau})\right)}{(\sum_j\pi_j\,\mathrm{NHP}(S\mid\boldsymbol{B}_j))^2}\\
&\cdot\int_S\mathrm{NHP}(S\mid\boldsymbol{B}_i)\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i)}{\partial\boldsymbol{B}_i}\right\|\cdot\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_j)}{\partial\boldsymbol{B}_j}\right\|dS\\
&\leq\pi_i\exp\left(aL(S)+m(S)\log(\frac{\tau+a/T}{\tau})\right)\int_S\mathrm{NHP}(S\mid\boldsymbol{B}_i)\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i)}{\partial\boldsymbol{B}_i}\right\|\\
&\cdot\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_j)}{\partial\boldsymbol{B}_j}\right\|dS\\
&\leq\pi_i\exp\left(aL(S)+m(S)\log(\frac{\tau+a/T}{\tau})\right)\\
&\cdot\int_S\mathrm{NHP}(S\mid\boldsymbol{B}_k^*)*\exp\left(-CL(S)+aL(S)+m(S)\log(\frac{\tau+a/T}{\tau})\right)(C_0L(S))^2dS\\
&\leq\pi_1\exp\left(-CL(S)+2aL(S)+2m(S)\log(\frac{\tau+a/T}{\tau})\right)\cdot(C_0L(S))^2,
\end{aligned}
$$

where C_0 is the upper bound of $\|\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i^*)/\partial\boldsymbol{B}_i\|,\forall i=1,\cdots,K$ with probability of $1-\delta$.

When $\|\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_i^*)/\partial\boldsymbol{B}_i\|>r\cdot L(S)$, if $L(S)\to\infty$,

$$
\begin{aligned}
I_0=&\frac{\pi_1\,\mathrm{NHP}(S\mid\boldsymbol{B}_1)}{\sum_j\pi_j\,\mathrm{NHP}(S\mid\boldsymbol{B}_j)}\cdot\int_{\left\|\frac{\partial\log\mathrm{NHP}(\mathbf{S}\mid\boldsymbol{B}_i^*)}{\partial\boldsymbol{B}_i}\right\|>r\cdot L(S)}\pi_i\,\mathrm{NHP}(S\mid\boldsymbol{B}_i^*)\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_1)}{\partial\boldsymbol{B}_1}\right\|dS\\
\leq&\int_{\left\|\frac{\partial\log\mathrm{NHP}(\mathbf{S}\mid\boldsymbol{B}_i^*)}{\partial\boldsymbol{B}_i}\right\|>r\cdot L(S)}\pi_i\,\mathrm{NHP}(S\mid\boldsymbol{B}_i^*)\left\|\frac{\partial\log\mathrm{NHP}(S\mid\boldsymbol{B}_1)}{\partial\boldsymbol{B}_1}\right\|dS\\
\leq&\pi_iC_0L(S)\int_{\left\|\frac{\partial\log\mathrm{NHP}(\mathbf{S}\mid\boldsymbol{B}_i^*)}{\partial\boldsymbol{B}_i}\right\|>r\cdot L(S)}\mathrm{NHP}(S\mid\boldsymbol{B}_i^*)dS\\
\leq&2\pi_iC_0L(S)\exp\left(-\frac{tL(S)}{c_0}\right)dS,
\end{aligned}
$$

where we use the same conclusion obtained above that $\mathbb{P}\left(\|\partial \log \mathrm{NHP}(S \mid \boldsymbol{B}_i^*)/\partial \boldsymbol{B}_i\| / L(S) \geq t\right) \leq 2\exp\left(-tL(S)/c_0\right)$. We still take $G = \min\{C_{gap} - 2a - 2m_c \log((\tau + a/T(S))/\tau), t/c_0\}$, where $\mathbb{P}\left(|M(S)/L(S)| \geq m_c\right) < \delta$ for small enough $\delta > 0$.

Lemma 11 (Matrix Chernoff I [Tropp, 2012]) *Consider a finite sequence of independent, random, self-adjoint matrices $\{\mathbf{X}_k\}$ with dimension d. Assume that each random matrix satisfies: $\mathbf{X}_k \succeq \mathbf{0}$ and $\lambda_{max}(\mathbf{X}_k) \leq R$ almost surely. Define*

$$\mu_{min} = \lambda_{min}\left(\sum_k \mathbb{E}\,\mathbf{X}_k\right) \quad \text{and} \quad \mu_{max} = \lambda_{max}\left(\sum_k \mathbb{E}\,\mathbf{X}_k\right).$$

Then we have

$$\mathbb{P}\left(\lambda_{min}\left(\sum_k \mathbf{X}_k\right) \leq (1-\delta)\mu_{min}\right) \leq d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{min}/R} \quad \text{for } \delta \in [0,1)$$
$$\mathbb{P}\left(\lambda_{max}\left(\sum_k \mathbf{X}_k\right) \geq (1+\delta)\mu_{max}\right) \leq d \cdot \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{max}/R} \quad \text{for } \delta \geq 0.$$

Lemma 12 *Function $\mu(\boldsymbol{B}_k \mid \boldsymbol{B}_k^{(t)})$ is a locally concave function with high probability for $k = 1, 2, \ldots, K$.*

Proof of Lemma 12 Without loss of generality, we let $k = 1$. We abuse the notation by treating $\alpha = \rho$ in the following proof. By taking the first derivative of the estimating equation, we have

$$\begin{aligned}
0 &= \nabla_{\boldsymbol{B}_1}\left(\sum_{n=1}^N w_1(S_n; \mathbf{B}^{(t)})\phi_\alpha\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n) - \mu(\boldsymbol{B}_1 \mid \boldsymbol{B}_1^{(t)})\right)\right) \\
&= \sum_{n=1}^N w_1(S_n; \mathbf{B}^{(t)})\phi_\alpha^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n) - \mu(\boldsymbol{B}_1 \mid \boldsymbol{B}_1^{(t)})\right) \\
&\cdot \left(\nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n) - \nabla\mu(\boldsymbol{B}_1 \mid \boldsymbol{B}_1^{(t)})\right).
\end{aligned}$$

By taking the second derivative, we have

$$\begin{aligned}
0 &= \nabla^2_{\boldsymbol{B}_1}\left(\sum_{n=1}^N w_1(S_n; \mathbf{B}^{(t)})\phi_\alpha\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n) - \mu(\boldsymbol{B}_1 \mid \boldsymbol{B}_1^{(t)})\right)\right) \\
&= \sum_{n=1}^N w_1(S_n; \mathbf{B}^{(t)})\phi_\alpha^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n) - \mu(\boldsymbol{B}_1 \mid \boldsymbol{B}_1^{(t)})\right) \\
&\cdot \left(\nabla^2 \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n) - \nabla^2\mu(\boldsymbol{B}_1 \mid \boldsymbol{B}_1^{(t)})\right) \\
&+ \sum_{n=1}^N w_1(S_n; \mathbf{B}^{(t)})\phi_\alpha^{''}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n) - \mu(\boldsymbol{B}_1 \mid \boldsymbol{B}_1^{(t)})\right) \\
&\cdot \alpha\left(\nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n) - \nabla\mu(\boldsymbol{B}_1 \mid \boldsymbol{B}_1^{(t)})\right)^2.
\end{aligned}$$

With a high probability, there exists c_ϕ such that $c_\phi|\phi^{'}(\eta)| > |\phi^{''}(\eta)|$, where $\eta \in (-9.5 + 2/c_\phi, 9.5 - 2/c_\phi)$. By Matrix Chernoff inequalities (Lemma 11), as $L(S) \to \infty$, we claim that $\lambda_{min}\left(\nabla^2 \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n)\right) - c_\phi \alpha \lambda_{max}\left(\nabla(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n))^2\right) \succeq 0$. Next we explain the reasons. Write S_n as $\{S_{n,1}, S_{n,2}, \cdots, S_{n,m(S)}\}$, then

$$\begin{aligned}\left[\nabla\frac{\log\mathrm{NHP}(S_n\mid\boldsymbol{B}_1)}{L(S_n)}\right]^2 \quad &= \quad \begin{bmatrix}\sum_{t=1}^{m(S)}\frac{\kappa_1(S_{n,t})}{\lambda_{\boldsymbol{B}_1}(S_{n,t})\cdot L(S_n)}-\int_0^T\kappa_1(x)dx\\ \vdots\\ \sum_{t=1}^{m(S)}\frac{\kappa_H(S_{n,t})}{\lambda_{\boldsymbol{B}_1}(S_{n,t})\cdot L(S_n)}-\int_0^T\kappa_H(x)dx\end{bmatrix}\\ &\quad\cdot\begin{bmatrix}\sum_{t=1}^{m(S)}\frac{\kappa_1(S_{n,t})}{\lambda_{\boldsymbol{B}_1}(S_{n,t})\cdot L(S_n)}-\int_0^T\kappa_1(x)dx\\ \vdots\\ \sum_{t=1}^{m(S)}\frac{\kappa_H(S_{n,t})}{\lambda_{\boldsymbol{B}_1}(S_{n,t})\cdot L(S_n)}-\int_0^T\kappa_H(x)dx\end{bmatrix}^\top\\ &:= \quad G\cdot G^\top.\end{aligned}$$

Therefore the largest eigenvalue of $\nabla\log\mathrm{NHP}(S_n \mid \boldsymbol{B}_1)/L(S_n)$ is the l2-norm of vector G. For each component of G, we know that $\mathbb{E}\left[\sum_{t=1}^{m(S)}\kappa_h(S_{n,t})/(\lambda_{\boldsymbol{B}_1}(S_{n,t})L(S_n))-\int_0^T\kappa_h(x)dx\right] = \int_0^T\left[(\kappa_h(t)\lambda_{\boldsymbol{B}_1}(t))/\lambda_{\boldsymbol{B}_1}(t)\right]dt/L(S_n)-\int_0^T\kappa_h(x)dx = 0, \forall h=1,\cdots,H$. When S_n is generated from the Poisson process with the intensity function $\lambda_{\boldsymbol{B}_1}(\cdot)$, we know that $\|G\|\sim O(\sqrt{H}\cdot L^{-1/2})$ with high probability. Thus, we get the result that $\alpha c_\phi\lambda_{max}\left(\nabla(\log\mathrm{NHP}(S_n\mid\boldsymbol{B}_1^*)/L(S_n))^2\right)\sim O(\alpha L^{-1/2})\to 0$ as $L\to 0$. For fixed $\boldsymbol{B}_1^{(t)}$, we also know that $\|G\|\sim O(\sqrt{H}L^{-1/2})$, while $\lambda_{min}\left(\nabla^2\log\mathrm{NHP}(S_n\mid\boldsymbol{B}_1)/L(S_n)\right)\sim O(1)$. Because of the continuity of ϕ' and ϕ'', it is easy to confirm the continuity of $\nabla^2\mu(\boldsymbol{B}_1\mid\boldsymbol{B}_1^{(t)})$.

Lemma 13 *If* $\left\|\boldsymbol{B}_k^{(t)}-\boldsymbol{B}_k^*\right\| < a/(T\cdot\kappa_{\max})$ *for* $k\in[K]$*, then* $\left\|\nabla\mu(\boldsymbol{B}_k^{(t)}\mid\boldsymbol{B}_k^{(t)})-\nabla\mu(\boldsymbol{B}_k^{(t)}\mid\boldsymbol{B}_k^*)\right\|\leq\gamma\left\|\boldsymbol{B}_k^{(t)}-\boldsymbol{B}_k^*\right\|$*. When we take the tuning parameter* α *sufficiently small, we get that* $\gamma\sim O(\sqrt{H}\exp(-GL)\cdot L)\to 0$ *as* $L\to\infty$*.*

Proof of Lemma 13

Let $\mathbf{B}^u=\mathbf{B}^*+u\left(\mathbf{B}^{(t)}-\mathbf{B}^*\right), \forall u\in[0,1]$. Then we know that

$$\begin{aligned}&\nabla_{\mathbf{B}_1^u}\left(w_1(S;\mathbf{B}^u)\phi_\alpha'\left(\log\mathrm{NHP}(S\mid\mathbf{B}_1^{(t)})/L(S)-\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right)\right)\\ =&\nabla_{\mathbf{B}_1^u}w_1(S;\mathbf{B}^u)\cdot\phi_\alpha'\left(\log\mathrm{NHP}(S\mid\mathbf{B}_1^{(t)})/L(S)-\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right)\\ &-w_1(S;\mathbf{B}^u)\phi_\alpha''\left(\log\mathrm{NHP}(S\mid\mathbf{B}_1^{(t)})/L(S)-\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right)\cdot\nabla_{\mathbf{B}_1^u}\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u),\end{aligned}$$

where $\nabla_{\mathbf{B}_1^u}\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)$ satisfy

$$\begin{aligned}0=&\mathbb{E}_S\left[\nabla w_1\left(S;\mathbf{B}^u\right)\phi_\alpha\left(\log\mathrm{NHP}(S\mid\mathbf{B}_1^{(t)})-\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right)\right.\\ &\left.-\ w_1\left(S;\mathbf{B}^u\right)\phi_\alpha'\left(\log\mathrm{NHP}(S\mid\mathbf{B}_1^{(t)})-\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right)\nabla_{\mathbf{B}_1^u}\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right].\end{aligned}$$

With a high probability, there exists c_ϕ such that $c_\phi|\phi^{'}(\eta)|>|\phi^{''}(\eta)|$, where $\eta\in(-9.5+2/c_\phi, 9.5-2/c_\phi)$. So we know that $\mathbb{E}_S[w_1(S;\mathbf{B}^u)\phi_\alpha''\left(\log\mathrm{NHP}(S\mid\mathbf{B}_1^{(t)})/L(S)-\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right)\cdot\nabla_{\mathbf{B}_1^u}\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)] < c_\phi\mathbb{E}_S[w_1(S;\mathbf{B}^u)\phi_\alpha'\left(\log\mathrm{NHP}(S\mid\mathbf{B}_1^{(t)})/L(S)-\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right)\cdot\nabla_{\mathbf{B}_1^u}\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)] = c_\phi\cdot\mathbb{E}_S[\nabla w_1\left(S;\mathbf{B}^u\right)\phi_\alpha\left(\log\mathrm{NHP}(S\mid\mathbf{B}_1^{(t)})-\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^u)\right)]$.

By Taylor's expansion, we have

$$\begin{aligned}
&\left\|\nabla\mu(\boldsymbol{B}_1^{(t)} \mid \boldsymbol{B}_1^{(t)}) - \nabla\mu(\boldsymbol{B}_1^{(t)} \mid \boldsymbol{B}_1^{*})\right\| \\
=&\Big\|\mathbb{E}_S\left(w_1\left(S;\mathbf{B}^{(t)}\right)\phi_\alpha'\left(\log\mathrm{NHP}(S \mid \mathbf{B}_1^{(t)}) - \mu(\boldsymbol{B}_1^{(t)} \mid \boldsymbol{B}_1^{(t)})\right)\right. \\
&\left.-w_1\left(S;\mathbf{B}^{*}\right)\phi_\alpha'\left(\log\mathrm{NHP}(S \mid \mathbf{B}_1^{(t)}) - \mu(\mathbf{B}_1^{(t)} \mid \mathbf{B}_1^{*})\right)\right)\cdot\alpha\nabla\log\mathrm{NHP}(S \mid \boldsymbol{B}_1^{(t)}))/L(S)\Big\| \\
=&\left\|\mathbb{E}\left[\int_{u=0}^{1}\nabla\left(w_1(S;\mathbf{B}^{\mathbf{u}})\phi_\alpha'\left(\log\mathrm{NHP}(S \mid \boldsymbol{B}_1^{(t)}) - \mu(\boldsymbol{B}_1^{(t)} \mid \boldsymbol{B}_1^{u})\right)\right)du\cdot\alpha\nabla\log\mathrm{NHP}(S \mid \boldsymbol{B}_1^{(t)})/L(S)\right]\right\| \\
\leq&\Bigg\|\mathbb{E}\int_{u=0}^{1}w_1(S;\mathbf{B}^{u})\left(1-w_1(S;\mathbf{B}^{u})\right)\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}^{\top}\left(\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^{*}\right)\alpha\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{(t)})}{\partial\boldsymbol{B}_1}/L(S)du \\
&-\sum_{i\neq 1}\mathbb{E}\int_{u=0}^{1}w_1(S;\mathbf{B}^{u})w_i(S;\mathbf{B}^{u})\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_i^{u})}{\partial\boldsymbol{B}_i}^{\top}\left(\boldsymbol{B}_i^{(t)} - \boldsymbol{B}_i^{*}\right)\alpha\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{(t)})}{\partial\boldsymbol{B}_1}/L(S)du\Bigg\|\phi_{max}' \\
&+\Big\|\mathbb{E}\int_{u=0}^{1}w_1(S;\mathbf{B}^{u})\phi_\alpha''\left(\log\mathrm{NHP}(S \mid \mathbf{B}_1^{(t)})/L(S) - \mu(\boldsymbol{B}_1^{(t)} \mid \boldsymbol{B}_1^{u})\right)\nabla_{\mathbf{B}_1^{u}}\mu(\boldsymbol{B}_1^{(t)} \mid \boldsymbol{B}_1^{u})du \\
&\cdot\alpha\nabla\log\mathrm{NHP}(S \mid \boldsymbol{B}_1^{(t)})^{\top}/L(S)\Big\| \\
\leq&U_1\left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^{*}\right\|_2 + \sum_{i\neq 1}U_i\left\|\boldsymbol{B}_i^{(t)} - \boldsymbol{B}_i^{*}\right\|_2 \\
&+\underbrace{\sup_{u\in[0,1]}\left\|\mathbb{E}\nabla w_1(S;\mathbf{B}^{u})\alpha\nabla\log\mathrm{NHP}(S \mid \boldsymbol{B}_1^{(t)})^{\top}/L(S)\right\|_2\cdot c_\phi\phi_{max}}_{I_0}\cdot\left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^{*}\right\|_2,
\end{aligned}$$

where

$$\begin{aligned}
U_1 &= \sup_{u\in[0,1]}\left\|\mathbb{E}w_1(S;\mathbf{B}^{u})\left(1-w_1(S;\mathbf{B}^{u})\right)\alpha/L(S)\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{(t)})}{\partial\boldsymbol{B}_1}\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}^{\top}\right\|_2 \\
U_i &= \sup_{u\in[0,1]}\left\|\mathbb{E}w_1(S;\mathbf{B}^{u})w_i(S;\mathbf{B}^{u})\alpha/L(S)\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{(t)})}{\partial\boldsymbol{B}_1}\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_i^{u})}{\partial\boldsymbol{B}_i}^{\top}\right\|_2.
\end{aligned}$$

For U_1, by triangle inequality, we have

$$\begin{aligned}
U_1 \leq& \sup_{u\in[0,1]}\left\|\mathbb{E}w_1(S;\mathbf{B}^{u})\left(1-w_1(S;\mathbf{B}^{u})\right)\alpha/L(S)\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}^{\top}\right\|_2 \\
&+\sup_{u\in[0,1]}\left\|\mathbb{E}w_1(S;\mathbf{B}^{u})\left(1-w_1(S;\mathbf{B}^{u})\right)\alpha/L(S)\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})^2}{\partial\boldsymbol{B}_1^2}(\boldsymbol{B}_1^{u} - \boldsymbol{B}_1^{(t)})\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}^{\top}\right\|_2 \\
\leq& \sup_{u\in[0,1]}\left\|\mathbb{E}w_1(S;\mathbf{B}^{u})\left(1-w_1(S;\mathbf{B}^{u})\right)\alpha/L(S)\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}^{\top}\right\|_2 \\
&+a\sup_{u\in[0,1]}\left\|\mathbb{E}w_1(S;\mathbf{B}^{u})\left(1-w_1(S;\mathbf{B}^{u})\right)\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}\right\|\cdot\left\|\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})^2}{\partial\boldsymbol{B}_1^2}/L(S)\right\|.
\end{aligned}$$

According to Lemma 8 , we know that $U_1 \sim O(\sqrt{H}\exp(-GL)\cdot L)$. When $L \to \infty$, $U_1 \to 0$. Similarly, for $U_i, i \neq 1$,

$$\begin{aligned}
U_i \leq& \sup_{u\in[0,1]}\left\|\mathbb{E}w_1(S;\mathbf{B}^{u})w_i(S;\mathbf{B}^{u})\alpha/L(S)\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})}{\partial\boldsymbol{B}_1}\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_i^{u})}{\partial\boldsymbol{B}_i}^{\top}\right\|_2 \\
&+a\sup_{u\in[0,1]}\left\|\mathbb{E}w_1(S;\mathbf{B}^{u})w_i(S;\mathbf{B}^{u})\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_i^{u})}{\partial\boldsymbol{B}_i}\right\|\cdot\left\|\frac{\partial\log\mathrm{NHP}(\mathbf{S} \mid \boldsymbol{B}_1^{u})^2}{\partial\boldsymbol{B}_1^2}/L(S)\right\|.
\end{aligned}$$

Refer to Lemma 10, we can get that $U_i \to 0$.

Similarly to U_1 and U_i, we use Lemma 8 and 10 and know that $I_0 \leq \sum_i U_i \cdot c_\phi\phi_{max} = O(\sqrt{H}\exp(-GL)\cdot L) \to 0$ with a high probability when $L(S) \to 0$.

Lemma 14 *For cluster i, we write*

$$
\begin{aligned}
&\nabla\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}} \ \ (\equiv \varrho_i^{(t)})\\
&:= \frac{\frac{1}{N}\sum_{n\in\mathcal{S}} w_1(S_n;\mathbf{B})\phi_\rho^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}}\right)\cdot \nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n)}{\frac{1}{N}\sum_n^N w_1(S_n;\mathbf{B})\phi_\rho^{\prime}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}}\right)},\\
&\nabla\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})\\
&:= \frac{Ew_1(S_n;\mathbf{B})\phi_\rho^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})\right)\cdot \nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n)}{Ew_1(S_n;\mathbf{B})\phi_\rho^{\prime}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})\right)}.
\end{aligned}
$$

Then we have $\left\|\nabla\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}} - \nabla\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})\right\| \le O(\sqrt{H}L\exp(-GL)/\sqrt{N} + (\rho + \sqrt{H})(1/\sqrt{NL} + (\rho v)/L + \log N/(\rho N) + \eta/\rho))$.

Proof of Lemma 14 Recall that $\mathcal{S} = \mathcal{S}_{\text{inlier}} \cup \mathcal{S}_{\text{outlier}}$ with $\mathcal{S}_{\text{inlier}} = \mathcal{S}_1 \cup ... \cup \mathcal{S}_K$. We define

$$
\begin{aligned}
&\nabla\mu(\widetilde{\boldsymbol{B}_i \mid \boldsymbol{B}}_i^{(t)})_{\mathcal{S}_{\text{inlier}}}\\
&:= \frac{N^{-1}\displaystyle\sum_{n\in\mathcal{S}_{\text{inlier}}} w_1(S_n;\mathbf{B})\phi_\rho^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}_{\text{inlier}}}\right)\frac{\nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)}{L(S_n)}}{N^{-1}\sum_{n\in\mathcal{S}_{\text{inlier}}} w_1(S_n;\mathbf{B})\phi_\rho^{\prime}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}_{\text{inlier}}}\right)}\\
&:= \frac{A}{B},
\end{aligned}
$$

which is the gradient based on the inlier samples only. By triangle inequality, we have

$$
\begin{aligned}
&\left\|\nabla\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}} - \nabla\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})\right\|\\
&\le \underbrace{\left\|\nabla\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}} - \nabla\mu(\widetilde{\boldsymbol{B}_i \mid \boldsymbol{B}}_i^{(t)})_{\mathcal{S}_{\text{inlier}}}\right\|}_{I_1} + \underbrace{\left\|\nabla\mu(\widetilde{\boldsymbol{B}_i \mid \boldsymbol{B}}_i^{(t)})_{\mathcal{S}_{\text{inlier}}} - \nabla\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})\right\|}_{I_2}.
\end{aligned}
$$

We consider the part I_2 first. According to Lemma 16 and Lemma 17, the deviation of $\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}_{\text{inlier}}}$ from $\mathbb{E}[\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n)]$ is $O((\rho v)/L + \log N/(\rho N) + \eta/\rho + L^2\exp\{-GL\} + \rho^2/\sqrt{L})$, so $\left|\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}_{\text{inlier}}}\right| \sim O(1/\sqrt{L} + (\rho v)/L + \log N/(\rho N) + \eta/\rho + L^2\exp\{-GL\})$. The standard deviation of $\phi_\rho^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}_{\text{inlier}}}\right)$ is $O(\rho/\sqrt{L} + (\rho^2 v)/L + \log N/N + \eta + \rho L^2\exp\{-L\})$, so the standard deviation of B is $O(\rho/\sqrt{NL} + (\rho^2 v)/L + \log N/N + \eta + \rho L^2\exp\{-L\})$. The standard deviation of part A is similar to part B. Similarly, the standard deviation of $\|\sum_N \nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/NL\|$ is $O(\sqrt{H}/\sqrt{NL})$, then $I_2 \sim O((\rho + \sqrt{H})/\sqrt{NL} + (\rho^2 v)/L + \log N/N + \eta + \rho L^2\exp\{-L\})$.

Next we consider the part I_1. Again by Lemma 16, $\left|\mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}} - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}_{\text{inlier}}}\right| \sim O((\rho v)/L + \log N/(\rho N) + \eta/\rho + L^2\exp\{-GL\})$. Note that

$$
\begin{aligned}
&\frac{1}{N}\sum_{n\in\mathcal{S}} w_1(S_n;\mathbf{B})\phi_\rho^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}}\right)\cdot \nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n)\\
&= \underbrace{\frac{1}{N}\sum_{n\in\mathcal{S}_1} w_1(S_n;\mathbf{B})\phi_\rho^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}}\right)\cdot \nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n)}_{W_1}\\
&+ \underbrace{\frac{1}{N}\sum_{n\in\mathcal{S}_{\text{inlier}}\setminus\mathcal{S}_1} w_1(S_n;\mathbf{B})\phi_\rho^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}}\right)\cdot \nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n)}_{W_2}\\
&+ \underbrace{\frac{1}{N}\sum_{n\in\mathcal{S}_{\text{outlier}}} w_1(S_n;\mathbf{B})\phi_\rho^{'}\left(\log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i \mid \boldsymbol{B}_i^{(t)})_{\mathcal{S}}\right)\cdot \nabla \log \mathrm{NHP}(S_n \mid \boldsymbol{B}_i)/L(S_n)}_{W_3}.
\end{aligned}
$$

According to Lemma 8, $\|W_2\| \leq \left\|N^{-1}\sum_{n\in\mathcal{S}_{\text{inlier}}\setminus\mathcal{S}_1} w_1(S_n;\mathbf{B})\cdot\nabla\log\text{NHP}(S_n\mid\boldsymbol{B}_i)/L(S_n)\right\| \sim O(\sqrt{H}L\exp(-GL))$, so $\|W_2 - EW_2\| \sim O(\sqrt{H}L\exp(-GL)/\sqrt{N})$. Similarly, $\|W_1 - EW_1\| \sim O(\sqrt{H}L\exp(-GL)/\sqrt{N})$. When $\left|\log\text{NHP}(S_n\mid\boldsymbol{B}_i)/L(S_n) - \mu(\boldsymbol{B}_i\mid\boldsymbol{B}_i^{(t)})_{\mathcal{S}}\right| < 9.5$, the gradient of outlier are less than a constant c_{out} with a high probability, then $\|W_3\| \leq O(\eta/\rho)$. Then $\|W_1+W_2+W_3-A\| \leq \|W_1-A\|+\|W_2\|+\|W_3\| \sim O(\sqrt{H}L\exp(-GL)/\sqrt{N}+\eta/\rho)$. The standard deviation of part A is similar to part B. Hence $\|I_1\| \leq O(\sqrt{H}L\exp(-GL)/\sqrt{N}+(\rho v)/L+\log N/(\rho N)+\eta/\rho)$.

In summary, $\left\|\nabla\mu(\boldsymbol{B}_i\mid\boldsymbol{B}_i^{(t)})_{\mathcal{S}} - \nabla\mu(\boldsymbol{B}_i\mid\boldsymbol{B}_i^{(t)})\right\| \leq I_1 + I_2 \leq O(\sqrt{H}L\exp(-GL)/\sqrt{N} + (\rho+\sqrt{H})(1/\sqrt{NL}+(\rho v)/L+\log N/(\rho N)+\eta/\rho))$.

Proof of Theorem 4 Recall the update rule and definition of $\nabla\mu(\boldsymbol{B}_1^{(t)}|\boldsymbol{B}_1^{(t)})$, we know that

$$\boldsymbol{B}_1^{(t+1)} = \boldsymbol{B}_1^{(t)} - \text{lr}\cdot\varrho_1^{(t)} = \boldsymbol{B}_1^{(t)} - \text{lr}\cdot\nabla\mu(\boldsymbol{B}_1^{(t)}|\boldsymbol{B}_1^{(t)})_{\mathcal{S}}.$$

By triangle inequality and Lemma 13 and 14, we have

$$\begin{aligned}
&\left\|\boldsymbol{B}_1^{(t+1)} - \boldsymbol{B}_1^*\right\| = \left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^* + \text{lr}\cdot\nabla\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^{(t)})_{\mathcal{S}}\right\| \\
&\leq \left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^* + \text{lr}\cdot\nabla\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^*)\right\| + \text{lr}\cdot\left\|\nabla\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^{(t)}) - \nabla\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^*)\right\| \\
&+ \text{lr}\cdot\left\|\nabla\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^{(t)}) - \nabla\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^{(t)})_{\mathcal{S}}\right\| \\
&\leq \frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^*\right\| + \frac{2}{\lambda_{\max}+\lambda_{\min}}\gamma\left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^*\right\| + \epsilon^{unif} \\
&\leq \frac{\lambda_{\max}-\lambda_{\min}+2\gamma}{\lambda_{\max}+\lambda_{\min}}\left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^*\right\| + \epsilon^{unif}.
\end{aligned}$$

To see why the second inequality holds, note that, for any $\boldsymbol{B}_1'$ with $\|\boldsymbol{B}_1' - \boldsymbol{B}^*\| \leq a$, $\Delta\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1')$ has the largest eigenvalue $-\lambda_{\min}$ and smallest eigenvalue $-\lambda_{\max}$. Applying the classical result for gradient descent with step size $\text{lr} = 2/(\lambda_{\max}+\lambda_{\min})$, it guarantees (see Nesterov [2003])

$$\left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^* + \text{lr}\cdot\nabla\mu(\boldsymbol{B}_1^{(t)}\mid\boldsymbol{B}_1^*)\right\| \leq \frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\left\|\boldsymbol{B}_1^{(t)} - \boldsymbol{B}_1^*\right\|.$$

This completes the proof.

Lemma 15 *For each sample $S_n \in \mathcal{S}$, when we select robust parameter $\alpha \sim O(L^\beta), 0 < \beta < 1/2$. Then as $L \to \infty$, the weight function $\phi_\alpha'\left(\log\text{NHP}\left(S_n\mid\mathbf{B}\right)/L(S_n) - \hat{\mu}_\phi(\mathbf{B})\right)$ tends to 1 with a high probability. If S_o is not sampled from $\mathbf{B}$, as $L\to\infty$, the weight function $\phi_\alpha'\left(\log\text{NHP}\left(s_o\mid\mathbf{B}\right)/L(S_n) - \hat{\mu}_\phi(\mathbf{B})\right)$ tends to 0 with a high probability.*

Proof of Lemma 15 By Lemma 7, we know that the standard deviation of $\log\text{NHP}\left(S_n\mid\mathbf{B}\right)/L(S_n)$ is $O(L(S_n)^{-1/2})$. From Lemma 16, we know that $\hat{\mu}_\phi(\mathbf{B}) - \mu^*(\mathbf{B}) = O_p((\rho v)/L + logN/(\rho N) + \eta/\rho + L^2\exp\{-GL\})$. So we have

$$\begin{aligned}
&\log\text{NHP}\left(S_n\mid\mathbf{B}\right)/L(S_n) - \hat{\mu}_\phi(\mathbf{B}) \sim O\left(L^{-1/2} + \frac{\rho v}{L} + \frac{\log N}{\rho N} + \frac{\eta}{\rho} + L^2\exp\{-GL\}\right) \\
\Rightarrow &\alpha\left(\log\text{NHP}\left(S_n\mid\mathbf{B}\right)/L(S_n) - \hat{\mu}_\phi(\mathbf{B})\right) \sim O(L^{\beta-1/2} + L^{2+\beta}\exp\{-GL\}) \to 0
\end{aligned}$$

for any $\alpha = O(L^\beta)$ with $0 < \beta < 1/2$, when $L\to\infty$. Looking back at the definition of robust function (1), we can easily know that $\lim_{x\to 0}\phi(x) = 1$. At this time there is $\phi_\alpha'\left(\log\text{NHP}\left(S_n\mid\mathbf{B}\right)/L(S_n) - \hat{\mu}_\phi(\mathbf{B})\right) \to 1$.

For S_o we have

$$\log\text{NHP}\left(S_o\mid\mathbf{B}\right)/L(S_o) - \hat{\mu}_\phi(\mathbf{B}) \sim O(1),$$

which implies

$$\alpha\left(\log\text{NHP}\left(S_o\mid\mathbf{B}\right)/L(S_o) - \hat{\mu}_\phi(\mathbf{B})\right) \sim O(L^\beta) \to \infty$$

when $L \to \infty$. Because of $\lim_{x \to \infty} \phi(x) = 0$, so we have $\phi'_\alpha \left(\log \mathrm{NHP} \left(S_o \mid \mathbf{B} \right) / L(S_o) - \hat{\mu}_\phi(\mathbf{B}) \right) \to 0$.

Proof of Theorem 5 According to Lemma 15, we know that the weight function will tend to 0 for all outliers as $L \to \infty$. Therefore we can distinguish almost all outliers with a high probability by setting the cutoff as 0.1.

Remark 2 *In all the above proofs, we do not take into account the shift parameter. The local convergence result could be still applied, if the algorithm starts with the true shift parameter and* $\left\| \boldsymbol{B}_k^{(0)} - \boldsymbol{B}_k^* \right\|$ *is small enough for* $k \in \{1, 2, \cdots, K\}$.

L Supporting Results of $\hat{\mu}_\phi^{(t)}(\boldsymbol{B}_k)$ and $\mu(\boldsymbol{B}_k | \boldsymbol{B}_k^*)$

In this section, we provide two supporting lemmas to characterize the difference between $\hat{\mu}_\phi^{(t)}(\boldsymbol{B}_k)$ and $\mu(\boldsymbol{B}_k | \boldsymbol{B}_k^*)$.

Lemma 16 *When* $\|\hat{\boldsymbol{B}}_k^{(t)} - \boldsymbol{B}_k^*\| \leq a$ *and* $\eta := |\mathcal{S}_{outlier}|/N < 1/[4(\log 5 + 1.5)]$*, it holds*

$$|\hat{\mu}_\phi^{(t)}(\boldsymbol{B}_k) - \mu^*(\boldsymbol{B}_k)| = O_p\left(\frac{\rho v}{L} + \frac{\log N}{\rho N} + \frac{\eta}{\rho} + L^2 \exp\{-GL\} \right), \tag{12}$$

where $\mu^*(\boldsymbol{B}_k) = \mathbb{E}_{S \sim \lambda_k^*}[\log \mathrm{NHP}(S | \boldsymbol{B}_k)]$ *and* $v := \sup_{\boldsymbol{B}_k} \mathbb{E}[(\log \mathrm{NHP}(S|\boldsymbol{B}_k))^2]$ *(*S *is an event sequence on* $[0, T]$ *generated according to* $\lambda_k^*(t)$*).*

Proof of Lemma 16 First, we define $\bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k)$ to be the solution to

$$\sum_{n=1}^{N} 1/L(S_n) \cdot \phi_\rho \left(\log \mathrm{NHP} \left(S_n \mid \boldsymbol{B}_k \right) / L(S_n) - \mu \right) = 0 \tag{13}$$

with respect to μ. We can show that

$$|\bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k) - \hat{\mu}_\phi^{(t)}(\boldsymbol{B}_k)| = O_p(L^2 \exp\{-GL\}). \tag{14}$$

To see this, we compare the difference between

$$\frac{1}{N} \sum_{n=1}^{N} 1/L(S_n) \cdot \phi_\rho \left(\log \mathrm{NHP} \left(S_n \mid \boldsymbol{B}_k \right) / L(S_n) - \bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k) \right)$$

and

$$\frac{1}{N} \sum_{n=1}^{N} r_{nk}^{(t)} / L(S_n) \cdot \phi_\rho \left(\log \mathrm{NHP} \left(S_n \mid \boldsymbol{B}_k \right) / L(S_n) - \bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k) \right).$$

By the previous analysis, we have already shown that $|r_{nk}^{(t)} - 1| = O_p(L \exp\{-GL\})$. Then such difference is bounded by $CL \exp\{-GL\} \cdot \sum_n L(S_n) \phi_\rho \left(\log \mathrm{NHP} \left(S_n \mid \boldsymbol{B}_k \right) / L(S_n) - \bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k) \right)$ which is order of $\exp\{-GL\}(\eta/\rho + \log L)$ and is less than $L \exp\{-GL\}$. (Here we use the fact that $\eta/\rho \to 0$). By the definition of $\bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k)$, we have

$$|\frac{1}{N} \sum_{n=1}^{N} r_{nk}^{(t)} / L(S_n) \cdot \phi_\rho \left(\log \mathrm{NHP} \left(S_n \mid \boldsymbol{B}_k \right) / L(S_n) - \bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k) \right)| \leq L \exp\{-GL\}.$$

It can be also checked that $\nabla_\mu \left(N^{-1} \sum_{n=1}^{N} r_{nk}^{(t)} / L(S_n) \cdot \phi_\rho \left(\log \mathrm{NHP} \left(S_n \mid \boldsymbol{B}_k \right) / L(S_n) - \mu \right) \right) \geq 1/2L$ for all bounded μ with probability 1. Therefore,

$$\begin{aligned} & \frac{1}{2L} |\bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k) - \hat{\mu}_\phi^{(t)}(\boldsymbol{B}_k)| \\ \leq \quad & |\frac{1}{N} \sum_{n=1}^{N} r_{nk}^{(t)} / L(S_n) \cdot \phi_\rho \left(\log \mathrm{NHP} \left(S_n \mid \boldsymbol{B}_k \right) / L(S_n) - \bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k) \right)| \leq L \exp\{-GL\}, \end{aligned}$$

which gives the desired result (14).

Next, we construct

$$B_{+,\boldsymbol{B}_k}(\mu) = (\mu^*(\boldsymbol{B}_k) - \mu) + \frac{\rho}{2}(\frac{v^*(\boldsymbol{B}_k)}{L} + (\mu^*(\boldsymbol{B}_k) - \mu)^2) + \frac{2\log N}{\pi_k^* N\rho}, \tag{15}$$
$$B_{-,\boldsymbol{B}_k}(\mu) = (\mu^*(\boldsymbol{B}_k) - \mu) - \frac{\rho}{2}(\frac{v^*(\boldsymbol{B}_k)}{L} + (\mu^*(\boldsymbol{B}_k) - \mu)^2) - \frac{2\log N}{\pi_k^* N\rho},$$

where $v^*(\boldsymbol{B}_k) = \mathbb{E}_{S\sim\lambda_k^*}[(\log \mathrm{NHP}(S|\boldsymbol{B}_k))^2]$, to put the upper and lower bounds on ϕ_ρ in (3.13). Following the proof of Theorem 3.1 in Bhatt et al. [2022] and the compactness of parameter space, we can have

$$|\bar{\mu}_\phi^{(t)}(\boldsymbol{B}_k) - \mu^*(\boldsymbol{B}_k)| = O_p\left(\frac{\rho v}{L} + \frac{\log N}{\rho N} + \frac{\eta}{\rho}\right) \tag{16}$$

for all $\boldsymbol{B}_k$, where $v = \max_{\boldsymbol{B}_k} v^*(\boldsymbol{B}_k)$. Combining (14) and (16), we prove the lemma.

Lemma 17 *It holds*

$$|\mu(\boldsymbol{B}_k \mid \boldsymbol{B}_k^*) - \mu^*(\boldsymbol{B}_k)| = O\left(L^2\exp\{-GL\} + \rho^2\sqrt{\frac{1}{L}}\right), \tag{17}$$

where $\mu^*(\boldsymbol{B}_k)$ is defined the same as that in Lemma 16.

Proof of Lemma 17 We first define $\bar{\mu}(\boldsymbol{B}_k \mid \boldsymbol{B}_k^*)$ to be the solution to

$$\mathbb{E}_S[\phi_\rho(\log \mathrm{NHP}(S \mid \boldsymbol{B}_k))/L(S) - \mu)] = 0$$

with respect to μ. By the same procedure as in the first part of proof of Lemma 16, we can show that

$$|\mu(\boldsymbol{B}_k \mid \boldsymbol{B}_k^*) - \bar{\mu}(\boldsymbol{B}_k \mid \boldsymbol{B}_k^*)| \leq L^2\exp\{-GL\}. \tag{18}$$

Next we compute the bound of $|\mathbb{E}_S[\phi_\rho(\log \mathrm{NHP}(S \mid \boldsymbol{B}_k))/L(S) - \mu^*(\boldsymbol{B}_k))]|$. Note that $\phi_\rho(x) = x - \rho^2x^3/6 + o(\rho^2x^3)$ by Taylor expansion. Therefore, for sufficiently small ρ, we have

$$\begin{aligned} & |\mathbb{E}_S[\phi_\rho(\log \mathrm{NHP}(S \mid \boldsymbol{B}_k))/L(S) - \mu^*(\boldsymbol{B}_k))]| \\ \leq\ & \frac{\rho^2}{3}|\mathbb{E}_S[(\log \mathrm{NHP}(S \mid \boldsymbol{B}_k))/L(S) - \mu^*(\boldsymbol{B}_k))^3]| \\ \leq\ & \frac{\rho^2}{3}\left(\mathbb{E}_S[(\log \mathrm{NHP}(S \mid \boldsymbol{B}_k))/L(S) - \mu^*(\boldsymbol{B}_k))^6]\right)^{1/2} \\ =\ & O\left(\rho^2\sqrt{\frac{1}{L}}\right). \end{aligned} \tag{19}$$

Lastly, note that $\nabla_\mu(\mathbb{E}_S[\phi_\rho(\log \mathrm{NHP}(S \mid \boldsymbol{B}_k))/L(S) - \mu)]) \geq 1/2$. Therefore, we have

$$|\bar{\mu}(\boldsymbol{B}_k \mid \boldsymbol{B}_k^*) - \mu^*(\boldsymbol{B}_k)| \leq 2|\mathbb{E}_S[\phi_\rho(\log \mathrm{NHP}(S \mid \boldsymbol{B}_k))/L(S) - \mu^*(\boldsymbol{B}_k))]| = O\left(\rho^2\sqrt{\frac{1}{L}}\right).$$

In summary, we get the desired result

$$\mu(\boldsymbol{B}_k \mid \boldsymbol{B}_k^*) - \mu^*(\boldsymbol{B}_k) = O\left(L^2\exp\{-GL\} + \rho^2\sqrt{\frac{1}{L}}\right).$$

References

Mahnoosh Alizadeh, Anna Scaglione, Jamie Davies, and Kenneth S Kurani. A scalable stochastic model for the electricity demand of electric and plug-in hybrid vehicles. *IEEE Transactions on Smart Grid*, 5(2):848–860, 2013.

GA Barnard. Time intervals between accidents—a note on maguire, pearson and wynn's paper. *Biometrika*, 40(1-2):212–213, 1953.

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Minimax m-estimation under adversarial corruption. In *Proceedings of the 39th International Conference on Machine Learning (ICML), Bartimore, MD*, 2022.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL `https://doi.org/10.1093/acprof:oso/9780199535255.001.0001`.

Emery N Brown, Riccardo Barbieri, Valérie Ventura, Robert E Kass, and Loren M Frank. The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346, 2002.

David Roxbee Cox and Peter AW Lewis. The statistical analysis of series of events. 1966.

Daryl J Daley, David Vere-Jones, et al. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

Carl De Boor. On calculating with b-splines. *Journal of Approximation theory*, 6(1):50–62, 1972.

Luc Duchateau and Paul Janssen. *The frailty model*. Springer, 2008.

Felipe Gerhard, Robert Haslinger, and Gordon Pipa. Applying the multivariate time-rescaling theorem to neural population models. *Neural computation*, 23(6):1452–1483, 2011.

Major Greenwood. The statistical study of infectious diseases. *Journal of the Royal Statistical Society*, 109(2):85–110, 1946.

Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part i. *ACM Sigmod Record*, 31(2): 40–45, 2002.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of applied probability*, 11(3):493–503, 1974.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

Song-Hee Kim and Ward Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480, 2014.

Shuang Li, Shuai Xiao, Shixiang Zhu, Nan Du, Yao Xie, and Le Song. Learning temporal point processes via reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Jonathan Pillow. Time-rescaling methods for the estimation and assessment of non-poisson neural encoding models. *Advances in neural information processing systems*, 22, 2009.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Oleksandr Shchur, Ali Caner Turkmen, Tim Januschowski, Jan Gasthaus, and Stephan Günnemann. Detecting anomalous event sequences with temporal point processes. *Advances in Neural Information Processing Systems*, 34:13419–13431, 2021.

Long Tao, Karoline E Weber, Kensuke Arai, and Uri T Eden. A common goodness-of-fit framework for neural population models using marked point process time-rescaling. *Journal of computational neuroscience*, 45:147–162, 2018.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Hongteng Xu and Hongyuan Zha. A dirichlet mixture model of hawkes processes for event sequence clustering. *Advances in neural information processing systems*, 30, 2017.