9 Appendices

9.1 Proofs of Theorems

9.1.1 Proof of Theorem 4.1

Uniqueness of the solution The elements of the Hessian for $\hat{\mathcal{R}}(\boldsymbol{\beta};t)$ in (6) are:

$$\begin{aligned} H(\hat{\mathcal{R}})_{ii} &= \sum_{j:j\neq i} (\tilde{X}_{ij}(t)+\tilde{X}_{ji}(t))\frac{\exp\beta_i\exp\beta_j}{(\exp\beta_i+\exp\beta_j)^2} \\ H(\hat{\mathcal{R}})_{ij} &= -(\tilde{X}_{ij}(t)+\tilde{X}_{ji}(t))\frac{\exp\beta_i\exp\beta_j}{(\exp\beta_i+\exp\beta_j)^2} \end{aligned} \tag{22}$$

Note that the Hessian has positive diagonal elements, non-positive off-diagonal elements, and zero column sums. With Condition 4.1, this implies that the Hessian can be regarded as a graph Laplacian for a connected graph. Following the classical proof of the property of graph Laplacian (von Luxburg, 2007),

$$v^T H(\hat{\mathcal{R}})v = \sum_{i<j} \frac{|\tilde{X}_{ij}(t)+\tilde{X}_{ji}(t)|}{2}(v_i-v_j)^2 \geq 0 \tag{23}$$

Then, Condition 4.1 guarantees that "=" is achieved if and only if $v = c\mathbf{1}$. This proves the uniqueness up to constant shifts.

Existence of solution Plugging in $\boldsymbol{\beta}=\mathbf{0}$, we get an upperbound for the minimum loss function $\hat{\mathcal{R}}^\star(t) := \hat{\mathcal{R}}(\hat{\boldsymbol{\beta}};t)$:

$$\hat{\mathcal{R}}^\star(t) \leq \log 2 \tag{24}$$

As $\hat{\mathcal{R}}(\boldsymbol{\beta};t)$ is continuous with respect to $\boldsymbol{\beta}$, it suffices to show that the level set of $\hat{\mathcal{R}}(\cdot;t)$ at $\log 2$ within $\{\boldsymbol{\beta}:\sum_{i=1}^N\beta_i=0\}$ is bounded so that it is compact.

Suppose that $\boldsymbol{\beta}$ is in the intersection between the levelset and $\{\boldsymbol{\beta}:\sum_{i=1}^N\beta_i=0\}$. Since each summand of $\hat{\mathcal{R}}(\boldsymbol{\beta})$ in (6) is non-negative, i.e.,

$$\frac{\tilde{X}_{ij}(t)}{\sum_{i',j':i'\neq j'}\tilde{X}_{i'j'}(t)}\log(1+\exp(\beta_j-\beta_i)) \geq 0 \tag{25}$$

if i and j satisfies $\tilde{X}_{ij}(t)>0$ then the corresponding summand should be smaller than $\log 2$ so that:

$$\begin{aligned} \beta_j-\beta_i &\leq \log(1+\exp(\beta_j-\beta_i)) \\ &\leq \log 2\frac{\sum_{i',j':i'\neq j'}\tilde{X}_{i'j'}(t)}{\tilde{X}_{ij}(t)} \end{aligned} \tag{26}$$

By Condition 4.1, for any distinct i and j, there exists an index sequence $(i=i_0,i_1,\ldots,i_n=j$ such that $X_{i_{k-1}i_k}>0$ for $k=1,2,\ldots,n$. Hence,

$$\begin{aligned} \beta_j-\beta_i &\leq \log 2\sum_{k=1}^n\frac{\sum_{i',j':i'\neq j'}\tilde{X}_{i'j'}(t)}{\tilde{X}_{i_{k-1}i_k}(t)} \\ &\leq \log 2\sum_{i',j':i'\neq j'}\tilde{X}_{i'j'}(t)\sum_{i',j':i'\neq j'}\frac{1}{\tilde{X}_{i'j'}(t)} =: B \end{aligned} \tag{27}$$

where $B\in(0,\infty)$.

In sum,

$$\|\boldsymbol{\beta}\|_\infty \leq \max_{i,j:i\neq j}|\beta_i-\beta_j| \leq B \tag{28}$$

and this proves the existence part of the theorem.

9.1.2 Proof of Theorem 5.1

The proof of this theorem is based on the proof of Lemma 1 in Simons and Yao (1999).

Since the kernel function W in Assumption 5.3 has support $(-\infty,\infty)$, $\tilde{X}_{ij}(t)>0$ if and only if team i defeated team j at least once any time. In other words, if Condition 4.1 holds for at least one time point, then so it does for every time point. Here, we prove that the probability of Condition 4.1 to hold at at least one time point converge to 1 as $N,T\to\infty$.

Given $p_{\min}$ instead of $\max_{i,j:i\neq j}\exp(\beta_i^*-\beta_j^*)$, the probability of the event $\mathcal{S}$ that no team in a subset S loses against a team not of S is no larger than

$$(1-p_{\min})^{|S|(N-|S|-1)T} \tag{29}$$

Hence, we bound the probability that data does not meet Condition 4.1 by a union bound

$$\begin{aligned} &\mathbb{P}(\text{Condition 4.1 fails}) \leq \sum_{S\subset[N]:S\neq\emptyset}\mathbb{P}(\mathcal{S}) \\ &\leq \sum_{n=1}^{N-1}\binom{N}{n}(1-p_{\min})^{n(N-n-1)T} \\ &\leq 2\sum_{n=1}^{\lceil N/2\rceil}\binom{N}{n}(1-p_{\min})^{n(N-n-1)T} \\ &\leq 2\sum_{n=1}^{\lceil N/2\rceil}\binom{N}{n}(1-p_{\min})^{nNT/2} \\ &\leq 2\left[(1+(1-p_{\min})^{NT/2})^N-1\right] \\ &\leq 2\left[(1+e^{-\frac{NTp_{\min}}{2}})^N-1\right] \\ &\leq 4Ne^{-\frac{NTp_{\min}}{2}} \end{aligned} \tag{30}$$

as long as $e^{-\frac{NTp_{\min}}{2}}\leq\frac{\log 2}{N}$. We note that $(1+\frac{\log 2}{N})^N \leq e^{\log 2}=2\leq 2\log 2+1=2N\frac{\log 2}{N}+1$. Hence,

$$\mathbb{P}(\text{Condition 4.1 fails}) \leq 4Ne^{-\frac{NTp_{\min}}{2}} \tag{31}$$

as long as $Ne^{-\frac{NTp_{\min}}{2}} \leq \log 2$.

Since $Ne^{-\frac{NTp_{\min}}{2}} \geq \log 2$ implies $4Ne^{-\frac{NTp_{\min}}{2}}$ to be larger than 1, the probability bound holds for any N, T, and $p_{\min}$.

9.1.3 Proof of Theorem 5.2

For readability, in our notation we will omit the dependence on the time point t in the expressions for $\hat{\boldsymbol{\beta}}(t)$ and $\boldsymbol{\beta}^*(t)$, unless required for clarity.

In our proofs we rely on the results and arguments of Simons and Yao (1999) to demonstrate consistency for the maximum likelihood estimator in the static Bradley-Terry model with an increasing number of parameters. In that setting, the authors parametrize the winning probabilities as $p_{i,j} = \frac{u_i^*}{u_i^*+u_j^*}$, where $u_i^* \equiv \exp(\beta_i^*)$, with $\boldsymbol{\beta}^* \in \mathbb{R}^N$ such that $\beta_1^* = 0$. Then, setting $\Delta u_i = \frac{\hat{u}_i - u_i^*}{u_i^*}$, where $\hat{u}_i$ is the MLE of u_i^* (with $\hat{u}_1 = 0$ by convention), it follows from the proof of Lemma 3 of Simons and Yao (1999) that

$$\max_i \frac{|\Delta u_i|}{|\Delta u_i| + 1} \leq \frac{8}{N-1} \max_{i,j} \frac{u_i^*}{u_j^*} \max_i \sum_{j:j\neq i} \left\{ \frac{\hat{u}_i}{\hat{u}_i + \hat{u}_j} - \frac{u_i^*}{u_i^* + u_j^*} \right\} \tag{32}$$

where $u_i^* = \exp(\beta_i^*)$. Next, the authors derived a high probability upper bound on

$$\max_i \sum_{j:j\neq i} \left\{ \frac{\hat{u}_i}{\hat{u}_i + \hat{u}_j} - \frac{u_i^*}{u_i^* + u_j^*} \right\} \tag{33}$$

using the facts that

$$\sum_{j:j\neq i} p_{ij} = \sum_{j:j\neq i} \frac{u_i^*}{u_i^* + u_j^*} \tag{34}$$

and

$$\sum_{j:j\neq i} \frac{X_{ij}}{T} = \sum_{j:j\neq i} \frac{\hat{u}_i}{\hat{u}_i + \hat{u}_j}, \tag{35}$$

where X_{ij} is the number of matches in which i defeated j. The second identity comes from the first order optimality condition of $\hat{\boldsymbol{\beta}}$.

In our time-varying setting, however, the subgradient optimality of $\hat{\boldsymbol{\beta}}(t)$ for $\hat{\mathcal{R}}(\boldsymbol{\beta}; t)$ only imply that, for each j,

$$\sum_{j:j\neq i} \tilde{X}_{ij}(t) = \sum_{j:j\neq i} \tilde{T}_{ij}(t) \frac{e^{\hat{\beta}_i}}{e^{\hat{\beta}_j} + e^{\hat{\beta}_i}}. \tag{36}$$

Thus, Eq. (35) does not hold in the dynamic setting, due to different $\tilde{X}_{ij}(t) + \tilde{X}_{ji}(t)$ across all $j \neq i$. Instead, we have that

$$\begin{aligned} &\frac{1}{N-1}\left(\sum_{j:j\neq i} \frac{\tilde{X}_{ij}(t)}{\tilde{T}_{ij}(t)} - \sum_{j:j\neq i} \frac{e^{\hat{\beta}_i}}{e^{\hat{\beta}_j} + e^{\hat{\beta}_i}}\right) \\ &= \begin{pmatrix} \sum_{j:j\neq i} \left(\frac{1}{N-1} - \frac{\tilde{T}_{ij}(t)}{\tilde{T}_i(t)}\right) \frac{\tilde{X}_{ij}(t)}{\tilde{T}_{ij}(t)} \\ + \sum_{j:j\neq i} \left(\frac{\tilde{T}_{ij}(t)}{\tilde{T}_i(t)} - \frac{1}{N-1}\right) \frac{e^{\hat{\beta}_i}}{e^{\hat{\beta}_j} + e^{\hat{\beta}_i}} \end{pmatrix} \end{aligned} \tag{37}$$

Since $\frac{\tilde{X}_{ij}(t)}{\tilde{T}_{ij}(t)}, \frac{e^{\hat{\beta}_i}}{e^{\hat{\beta}_j}+e^{\hat{\beta}_i}} < 1$,

$$\begin{aligned} &\left|\frac{1}{N-1}\left(\sum_{j:j\neq i} \frac{\tilde{X}_{ij}(t)}{\tilde{T}_{ij}(t)} - \sum_{j:j\neq i} \frac{e^{\hat{\beta}_i}}{e^{\hat{\beta}_j} + e^{\hat{\beta}_i}}\right)\right| \\ &\leq 2\delta_h(t) \end{aligned} \tag{38}$$

and

$$\begin{aligned} &\frac{1}{N-1} \sum_{j:j\neq i} \left\{ \frac{e^{\hat{\beta}_i}}{e^{\hat{\beta}_i} + e^{\hat{\beta}_j}} - \frac{e^{\beta_i^*}}{e^{\beta_i^*} + e^{\beta_j^*}} \right\} \\ &\leq 2\delta_h(t) + \frac{1}{N-1} \sum_{j:j\neq i} \left\{ \frac{\tilde{X}_{ij}(t)}{\tilde{T}_{ij}(t)} - p_{ij}(t) \right\}. \end{aligned} \tag{39}$$

To make the bias-variance trade-off due to kernel smoothing more explicit, we decompose the term

$$\sum_{j:j\neq i} \left\{ \frac{\tilde{X}_{ij}(t)}{\tilde{T}_{ij}(t)} - p_{ij}(t) \right\} \tag{40}$$

as

$$\begin{aligned} &\sum_{j:j\neq i} \left(\frac{\sum_k W_h(t_k, t)(\mathbf{1}_{ij}(t_k) - p_{ij}(t_k))}{\sum_k W_h(t_k, t)} \right) \\ &+ \sum_{j:j\neq i} \left(\frac{\sum_k W_h(t_k, t) p_{ij}(t_k)}{\sum_k W_h(t_k, t)} - p_{ij}(t) \right) \\ &=: \Delta_i^{(var)} + \Delta_i^{(bias)} \end{aligned} \tag{41}$$

where, for brevity, t_k and $\mathbf{1}_{ij}(t_k)$ here stand for $t_k^{(i,j)}$ and $\mathbf{1}(i \text{ defeats } j \text{ at } t_k)$, respectively.

For the first term, we have that

$$\begin{aligned} &\mathbb{P}\left(\left|\Delta_i^{(var)}\right| \geq \epsilon\right) \\ &= \mathbb{P}\left(\left|\sum_{j:j\neq i} \frac{\sum_k W_h(t_k, t)(\mathbf{1}_{ij}(t_k) - p_{ij}(t_k))}{\sum_k W_h(t_k, t)}\right| \geq \epsilon\right) \\ &= \mathbb{P}\left(\begin{array}{l} \left|\sum_{j,k} h W_h(t_k, t) \frac{s_{\min}}{s_j} (\mathbf{1}_{ij}(t_k) - p_{ij}(t_k))\right| \\ \geq \epsilon \cdot h \cdot s_{\min} \end{array}\right) \end{aligned} \tag{42}$$

where $s_j = \sum_k W_h(t_k, t)$ and $s_{\min} = \min_{j:j\neq i}\ s_j$.

Next, $hW_h(t_k,t)\frac{s_{\min}}{s_j} = W\left(\frac{t_k-t}{h}\right)\frac{s_{\min}}{s_j} \leq 1$ and hence that multiplicative Chernoff bound (see, e.g. Raghavan, 1988) yields that

$$\begin{aligned}&\mathbb{P}\Big(\Big|\Delta_i^{(var)}\Big| \geq \epsilon\Big)\\ &\leq 2\exp\left(-\frac{(\epsilon\cdot h\cdot s_{\min})^2}{3\sum_{j,k} hW_h(t_k,t)\frac{s_{\min}}{s_j}p_{ij}(t_k)}\right)\\ &\leq 2\exp\left(-\frac{\epsilon^2 hD_mT}{18(N-1)(1-p_{\min})}\right)\end{aligned} \tag{43}$$

for each i as long as

$$\frac{\epsilon}{\sum_{j,k}\frac{W_h(t_k,t)}{\sum_{k'}W_h(t_{k'},t)}p_{ij}(t_k)} \leq 1. \tag{44}$$

This condition holds for $\epsilon \leq p_{\min}$.

We note that we have also used the bounds

$$\frac{1}{6}D_mT \leq \sum_k W_h(t_k,t) \leq D_MT \tag{45}$$

for any i, j and sufficiently small h, which were shown in Section 9.1.4.

Then using the union bound,

$$\begin{aligned}&\mathbb{P}\Big(\max_i\Big|\Delta_i^{(var)}\Big| \geq \epsilon\Big)\\ &\leq 2N\exp\left(-\frac{\epsilon^2 hD_mT}{18(1-p_{\min})(N-1)}\right)\end{aligned} \tag{46}$$

Hence, plugging in $\epsilon = \sqrt{\frac{36(1-p_{\min})(N-1)\log N}{hD_mT}}$, we get that, with probability at least $1-\frac{2}{N}$,

$$\max_i|\Delta_i^{(var)}| \leq \sqrt{\frac{36(1-p_{\min})(N-1)\log N}{hD_mT}} \tag{47}$$

To handle the deterministic bias terms $\Delta_i^{(bias)}$, we rely on the following bound, whose proof is given below in Section 9.1.4.

Lemma 9.1. *Suppose that*

1. *$t_1, t_2, \ldots, t_T$ satisfies Eq. (8) and*
2. *$\frac{1}{T} = o(h)$ as $T\to\infty$.*

Then, for a L_f-Lipschitz function $f:[0,1]\to\mathbb{R}$,

$$\sup_{t\in[0,1]}\left|\sum_{k=1}^T \frac{W_h(t_k,t)}{\sum_{k'}W_h(t_{k'},t)}f(t_k) - f(t)\right| \leq C_sh \tag{48}$$

with a universal constant C_s depending only on D_m, D_M, W and L_f.

Accordingly,

$$\begin{aligned}&\max_i|\Delta_i^{(bias)}|\\ &\leq \max_i\sum_{j:j\neq i}\left\{\frac{\sum_k W_h(t_k,t)p_{ij}(t_k)}{\sum_k W_h(t_k,t)} - p_{ij}(t)\right\}\\ &\leq C_s(N-1)h\end{aligned} \tag{49}$$

for some constant C_s depending only on D_m, D_M, W and L_p.

Thus, combining all the pieces,

$$\begin{aligned}&\max_i\frac{|e^{\hat\beta_i-\beta_i^*}-1|}{|e^{\hat\beta_i-\beta_i^*}-1|+1}\\ &\leq 8M(t)\left(2\delta_h(t) + \max_i\frac{|\Delta_i^{(var)}(t)| + |\Delta_i^{(bias)}(t)|}{N-1}\right)\\ &\leq 8M(t)\left(2\delta_h(t) + \sqrt{\frac{36(1-p_{\min})\log N}{hD_m(N-1)T}} + C_sh\right)\end{aligned} \tag{50}$$

with probability at least $1-\frac{2}{N}$ as long as $\epsilon \leq p_{\min}$.

Plugging in $h = \max\left\{\left(\frac{1}{T}\right)^{1+\eta}, \left(\frac{36(1-p_{\min})\log N}{C_s^2D_m(N-1)T}\right)^{\frac{1}{3}}\right\}$ leads to the bound

$$\begin{aligned}&\max_i\frac{|e^{\hat\beta_i-\beta_i^*}-1|}{|e^{\hat\beta_i-\beta_i^*}-1|+1}\\ &\leq 8M(t)\left(2\delta_h(t) + \left(\frac{36C_s(1-p_{\min})\log N}{D_m(N-1)T}\right)^{\frac{1}{3}} + C_sh\right)\\ &\leq 16M(t)\left(\delta_h(t) + C_sh\right)\end{aligned} \tag{51}$$

with probability at least $1-\frac{2}{N}$ when $\epsilon \leq p_{\min}$. We note that, given our choice for h, $\epsilon = \sqrt{\frac{36(1-p_{\min})(N-1)\log N}{hD_mT}} \leq C_sh$. Hence, for a sufficiently small h, if the right hand side is smaller than, say, $\frac{1}{3}$ then

$$\begin{aligned}\|\hat{\boldsymbol\beta} - \boldsymbol\beta^*\|_\infty &\leq 3\max_i\frac{|e^{\hat\beta_i-\beta_i^*}-1|}{|e^{\hat\beta_i-\beta_i^*}-1|+1}\\ &\leq 48M(t)\left(\delta_h(t)+C_sh\right)\end{aligned} \tag{52}$$

with probability at least $1-\frac{2}{N}$ since $\frac{|e^x-1|}{|e^x-1|+1} \geq \frac{|x|}{3}$ for $|x|\leq 1$.

9.1.4 Proof of Lemma 9.1

Since f is L_f-Lipschitz,

$$\begin{aligned}&\left|\sum_{k=1}^{T}\frac{W_h(t_k,t)}{\sum_{k'}W_h(t_{k'},t)}f(t_k)-f(t)\right|\\&\le\sum_{k=1}^{T}\frac{W_h(t_k,t)}{\sum_{k'}W_h(t_{k'},t)}|f(t_k)-f(t)|\\&\le L_f\sum_{k=1}^{T}\frac{W_h(t_k,t)}{\sum_{k'}W_h(t_{k'},t)}|t_k-t|.\end{aligned}\tag{53}$$

Let $I_1=\left[0,\frac{t_1+t_2}{2}\right], I_2=\left[\frac{t_1+t_2}{2},\frac{t_2+t_3}{2}\right],\dots,I_T=\left[\frac{t_{T-1}+t_T}{2},1\right]$ and l_k be the length of I_k. We note that $\frac{1}{D_MT}\le l_k\le\frac{2}{D_mT}$ by Eq. (9). Then,

$$\begin{aligned}&\int_0^1|x-t|W_h(x,t)dx=\sum_k\int_{I_k}|x-t|W_h(x,t)\\&=\begin{pmatrix}\sum_k l_k|t_k-t|W_h(t_k,t)\\+\sum_k\int_{I_k}\begin{pmatrix}\left|\frac{x-t}{h}\right|W\left(\frac{x-t}{h}\right)\\-\left|\frac{t_k-t}{h}\right|W\left(\frac{t_k-t}{h}\right)\end{pmatrix}dx\end{pmatrix}.\end{aligned}\tag{54}$$

Since $|\cdot|W$ has a finite total variation,

$$\begin{aligned}&\sum_k\int_{I_k}\left|\begin{matrix}\left|\frac{x-t}{h}\right|W\left(\frac{x-t}{h}\right)\\-\left|\frac{t_k-t}{h}\right|W\left(\frac{t_k-t}{h}\right)\end{matrix}\right|dx\\&\le\sum_k\frac{1}{D_mT}\sup_{x,y\in I_k}\left|\begin{matrix}\left|\frac{x-t}{h}\right|W\left(\frac{x-t}{h}\right)\\-\left|\frac{y-t}{h}\right|W\left(\frac{t_k-t}{h}\right)\end{matrix}\right|\\&\le\frac{\mathcal{V}(|\cdot|W)}{D_mT}.\end{aligned}\tag{55}$$

Hence,

$$\begin{aligned}&\int_0^1|x-t|W_h(x,t)dx\\&\ge\begin{pmatrix}\frac{1}{D_MT}\sum_k|t_k-t|W_h(t_k,t)\\-\frac{\mathcal{V}(|\cdot|W)}{D_mT}\end{pmatrix}.\end{aligned}\tag{56}$$

As a result,

$$\begin{aligned}&\sum_k|t_k-t|W_h(t_k,t)\\&\le D_MTh\int_{-\infty}^{\infty}|x|W(x)dx+\frac{D_M\mathcal{V}(|\cdot|W)}{D_m}.\end{aligned}\tag{57}$$

On the other hand, with a similar argument,

$$\begin{aligned}&\int_0^1W_h(x,t)dx=\sum_k\int_{I_k}W_h(x,t)\\&=\begin{pmatrix}\sum_k l_kW_h(t_k,t)\\+\sum_k\int_{I_k}\left(\frac{1}{h}W\left(\frac{x-t}{h}\right)-\frac{1}{h}W\left(\frac{t_k-t}{h}\right)\right)dx\end{pmatrix}\\&\le\frac{2}{D_mT}\sum_kW_h(t_k,t)+\frac{\mathcal{V}(W)}{D_mTh},\end{aligned}\tag{58}$$

implying that

$$\sum_kW_h(t_k,t)\ge\frac{D_mT}{2}\int_{-t/h}^{(1-t)/h}W(x)dx-\frac{D_M\mathcal{V}(W)}{2D_mh}.\tag{59}$$

As long as $h\to0$ and $\frac{1}{T}=o(h)$,

$$\inf_{t\in[0,1]}\int_{-t/h}^{(1-t)/h}W(x)dx\tag{60}$$

is bounded below from 0 (in particular, we consider a small enough h so that it is bounded below by, say, $\frac{1}{3}$), and the term $\frac{D_M\mathcal{V}(|\cdot|W)}{D_m}$ and $\frac{D_M\mathcal{V}(W(x)}{2D_mh}$ in Eqs. (57) and (59) become asymptotically negligible. As a result,

$$\sum_{k=1}^{T}\frac{W_h(t_k,t)}{\sum_{k'}W_h(t_{k'},t)}|t_k-t|\le C'h\tag{61}$$

where C' is a universal constant depending only on D_m, D_M, and W, and furthermore

$$\left|\sum_{k=1}^{T}\frac{W_h(t_k,t)}{\sum_{k'}W_h(t_{k'},t)}f(t_k)-f(t)\right|\le C_sh\tag{62}$$

for a univeral constant C_s depending only on D_m, D_M, W, and L_f.

9.1.5 Proof of Theorem 5.3

In Section 9.1.3, we showed that

$$\begin{aligned}&\max_i\frac{|e^{\hat\beta_i(t)-\beta_i^*(t)}-1|}{|e^{\hat\beta_i(t)-\beta_i^*(t)}-1|+1}\\&\le8M(t)\left(2\delta_h(t)+\max_i\frac{|\Delta_i^{(var)}(t)|+|\Delta_i^{(bias)}(t)|}{N-1}\right)\end{aligned}\tag{63}$$

Since the bound for $\Delta_i^{(bias)}(t)$ depends only on D_m, D_M, W, and L_f, it is sufficient to find a bound for

$$\sup_{t\in[0,1]}\max_i|\Delta_i^{(var)}(t)|.\tag{64}$$

We use a covering approach. For $L \in \mathbb{N}$, let $\bar{t}_l = \frac{2l-1}{2L}$ for $l = 1, 2, \ldots, L$. Then for any $t \in [0,1]$ there exists l^* such that $|t - \bar{t}_{l^*}| \le \frac{1}{2L}$ and

$$\begin{aligned}
&\max_i \left\{ \Delta_i^{(var)}(t) \right\} \\
&\le \max_i \left\{ \begin{array}{l} \displaystyle\sum_{j:j\neq i} \frac{\sum_k W_h(t_k,t)\mathbf{1}(i \text{ defeats } j \text{ at } t_k)}{\sum_k W_h(t_k,t)} \\ \displaystyle - \sum_{j:j\neq i} \frac{\sum_k W_h(t_k,\bar{t}_{l^*(t)})\mathbf{1}(i \text{ defeats } j \text{ at } t_k)}{\sum_k W_h(t_k,\bar{t}_{l^*(t)})} \end{array} \right\} \\
&+ \max_i \Delta_i^{(var)}(\bar{t}_{l^*})
\end{aligned} \tag{65}$$

where t_k here stands $t_k^{(i,j)}$ for brevity.

In order to bound the second term in the curly brackets, we bound each of its summands as follows:

$$\begin{aligned}
&\left| \frac{W_h(t_k,t)}{\sum_k W_h(t_k,t)} - \frac{W_h(t_k,\bar{t}_{l^*(t)})}{\sum_k W_h(t_k,\bar{t}_{l^*(t)})} \right| \\
&\le \left| \frac{W_t S_{\bar{t}} - W_{\bar{t}} S_t}{S_t S_{\bar{t}}} \right| \\
&\le \frac{|W_t - W_{\bar{t}}|}{S_{\bar{t}}} + \frac{|S_{\bar{t}} - S_t| W_{\bar{t}}}{S_t S_{\bar{t}}}
\end{aligned} \tag{66}$$

where we denote $W_t = W_h(t_k,t)$, $W_{\bar{t}} = W_h(t_k,\bar{t}_{l^*(t)})$, $S_t = \sum_k W_h(t_k,t)$, and $S_{\bar{t}} = \sum_k W_h(t_k,\bar{t}_{l^*(t)})$ for brevity.

We have seen in Section 9.1.4 that, for any a sufficiently small h,

$$S_t,\ S_{\bar{t}} \ge \frac{D_m T}{6}. \tag{67}$$

Thus,

$$\begin{aligned}
&\frac{|W_t - W_{\bar{t}}|}{S_{\bar{t}}} + \frac{|S_{\bar{t}} - S_t| W_{\bar{t}}}{S_t S_{\bar{t}}} \\
&\le \frac{6}{D_m T} \frac{L_W}{h} |t - \bar{t}_{l^*(t)}| + \left(\frac{6}{D_m T} \right)^2 T \frac{L_W}{h} |t - \bar{t}_{l^*(t)}| \\
&\le \frac{36}{D_m^2 L h^2 T}
\end{aligned} \tag{68}$$

as $D_m \le 1$ and W is L_W-Lipschitz by assumption. Hence,

$$\begin{aligned}
&\max_i \left\{ \begin{array}{l} \displaystyle\sum_{j:j\neq i} \frac{\sum_k W_h(t_k,t)\mathbf{1}(i \text{ defeats } j \text{ at } t_k)}{\sum_k W_h(t_k,t)} \\ \displaystyle - \sum_{j:j\neq i} \frac{\sum_k W_h(t_k,\bar{t}_{l^*(t)})\mathbf{1}(i \text{ defeats } j \text{ at } t_k)}{\sum_k W_h(t_k,\bar{t}_{l^*(t)})} \end{array} \right\} \\
&\le \frac{36(N-1)}{D_m^2 L h^2}
\end{aligned} \tag{69}$$

On the other hand,

$$\begin{aligned}
&\max_i \Delta_i^{(var)}(\bar{t}_{l^*}) \le \max_{l,i} \Delta_i^{(var)}(\bar{t}_l) \\
&= \max_{l,i} \left\{ \sum_{j:j\neq i} \left(\frac{\sum_k W_h(t_k,t)(\mathbf{1}_{ij}(t_k) - p_{ij}(t_k))}{\sum_k W_h(t_k,t)} \right) \right\}
\end{aligned} \tag{70}$$

where, again,$\mathbf{1}_{ij}(t_k)$ here stands $\mathbf{1}(i \text{ defeats } j \text{ at } t_k)$ for simplicity.

Using Eq. (42) and a union bound, we get that

$$\begin{aligned}
&\mathbb{P}\left(\max_{l,i} \Delta_i^{(var)}(\bar{t}_l) \ge \epsilon \right) \\
&\le 2NL \exp\left(-\frac{\epsilon^2 h D_m T}{18(1-p_{\min})(N-1)} \right),
\end{aligned} \tag{71}$$

for $\epsilon \le p_{\min}$.

Next we plug in $\epsilon = \sqrt{\frac{36(1-p_{\min})(N-1)\log(NL)}{hD_mT}}$ to obtain the bounds

$$\left| \max_{l,i} \Delta_i^{(var)}(\bar{t}_{l^*}) \right| \le \sqrt{\frac{36(1-p_{\min})(N-1)\log(NL)}{hD_mT}} \tag{72}$$

and, in turn,

$$\begin{aligned}
&\sup_{t,i} \frac{|e^{\hat{\beta}_i(t) - \beta_i^*(t)} - 1|}{|e^{\hat{\beta}_i(t) - \beta_i^*(t)} - 1| + 1} \\
&\le \sup_t 8M(t) \left(\begin{array}{l} 2\delta_h(t) + \dfrac{36}{D_m^2 L h^2} \\ + \sqrt{\dfrac{36(1-p_{\min})\log(NL)}{hD_m(N-1)T}} + C_s h \end{array} \right)
\end{aligned} \tag{73}$$

with probability at least $1 - \frac{2}{NL}$ and as long as $\epsilon \le p_{\min}$.

Plugging in $h = \max\left\{ \left(\frac{1}{T}\right)^{1+\eta}, \left(\frac{36(1-p_{\min})\log(NT^{3+3\eta})}{C_s^2 D_m (N-1) T} \right)^{\frac{1}{3}} \right\}$ and $L = \lceil h^{-3} \rceil$, we conclude that

$$\begin{aligned}
&\sup_{t,i} \frac{|e^{\hat{\beta}_i(t) - \beta_i^*(t)} - 1|}{|e^{\hat{\beta}_i(t) - \beta_i^*(t)} - 1| + 1} \\
&\le \sup_t 8M(t) \left(\begin{array}{l} 2\delta_h(t) + \dfrac{36}{D_m^2 L h^2} + C_s h \\ + \sqrt{\dfrac{36(1-p_{\min})\log(NL)}{hD_m(N-1)T}} \end{array} \right) \\
&\le \sup_t 8M(t) \left(\begin{array}{l} 2\delta_h(t) + \left(C_s + \dfrac{72}{D_m^2} \right) h \\ + \sqrt{\dfrac{36(1-p_{\min})\log(Nh^{-3})}{hD_m(N-1)T}} \end{array} \right) \\
&\le \sup_t 16M(t) \left(\delta_h(t) + \left(C_s + \frac{36}{D_m^2} \right) h \right)
\end{aligned} \tag{74}$$

with probability at least $1 - \frac{2h^3}{N}$ when $\epsilon \le p_{\min}$. Since $\epsilon \le \sqrt{3(1+\eta)}C_s h$ given the choice of h, this bound holds for all sufficiently small h.

9.1.6 Proof of Theorem 5.4

For convenience, we omit the time index t for $\hat{\boldsymbol{\beta}}(t)$, $\boldsymbol{\beta}^*(t)$, and $p_{\min}(t)$, unless it is required for clarification.

We seek to replace $M(t)$ in Eq. (16) by a term of $p_{\min}$. This requires $\exp(\beta_i^* - \beta_j^*)$ to be bounded above by a function of $p_{\min}$. The following lemma provides the desired bound. The proof is in Section 9.1.7.

Lemma 9.2.

$$\max_{i,j:i\neq j} |\beta_i^* - \beta_j^*| - \frac{1}{p_{min}} \tag{75}$$

is upper-bounded by a universal constant, and hence

$$\max_{i,j:i\neq j} \exp(|\beta_i^* - \beta_j^*|) \le C_p \exp\left(\frac{1}{p_{min}}\right) \tag{76}$$

for some universal constant $1 < C_p < 1.5$.

Plugging in the new bound on $\exp(\beta_i^* - \beta_j^*)$, we get

$$\|\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}^*(t)\|_\infty \le 72K\left(\delta_h(t) + C_s h\right) \tag{77}$$

instead of $48M(t)\left(\delta_h(t) + C_s h\right)$ in Eq. (16)

This result easily extends to the uniform case Eq. (21).

9.1.7 Proof of Lemma 9.2

Let d_0 be the difference in scores which implies a bias of probability $\frac{p_{\min}}{2}$:

$$\frac{1}{1+\exp(d_0)} = \frac{p_{\min}}{2} \tag{78}$$

Suppose that

$$i_{\max} = \arg\max_i \beta_i^* \text{ and } i_{\min} = \arg\min_i \beta_i^* \tag{79}$$

and that

$$\beta_{\max}^* = \max_i \beta_i^* \text{ and } \beta_{\min}^* = \min_i \beta_i^* \tag{80}$$

Then, the maximal difference between $\boldsymbol{\beta}_i^*$'s $d_{\max}$ is

$$d_{\max} = \max_{i,j:i\neq j} \beta_i^* - \beta_j^* = \beta_{\max}^* - \beta_{\min}^* \tag{81}$$

Let $I_1 = \{i : \beta_i < \beta_{\min} + d_0\}$. Plugged in $i = i_{\min}$, Eq. (34) implies

$$\begin{aligned}(N-1)p_{\min} &\le \sum_{j:j\neq i_{\min}} p_{i_{\min}j}(t) \\ &= \sum_{j:j\neq i_{\min}} \frac{1}{1+\exp(\beta_j^* - \beta_{\min}^*)} \\ &\le \frac{|I_1|-1}{2} + (N-1)\frac{p_{\min}}{2}\end{aligned} \tag{82}$$

Hence, $|I_1| \ge (N-1)p_{\min} + 1$.

Now, let $I_2 = \{i : \beta_i < \beta_{\min} + 2d_0\}$. Summing Eq. (34) plugged in $i \in I$, we get

$$\begin{aligned}(N - |I_1|)|I_1|p_{\min} &\le \sum_{j\in I_1^C}\sum_{i\in I_1} p_{ij}(t) \\ &= \sum_{j\in I_1^C}\sum_{i\in I_1} \frac{1}{1+\exp(\beta_j^* - \beta_i^*)} \\ &\le \frac{(|I_2| - |I_1|)|I_1|}{2} + (N-|I_1|)|I_1|\frac{p_{\min}}{2}\end{aligned} \tag{83}$$

and hence

$$\begin{aligned}|I_2| &\ge |I_1| + (N-|I_1|)p_{\min} \\ &\ge Np_{\min} + |I_1|(1-p_{\min}) \\ &\ge (N-1)(1-(1-p_{\min})^2) + 1\end{aligned} \tag{84}$$

Similarly for $I_k = \{i : \beta_i < \beta_{\min} + kd_0\}$ and $J_k = \{j : \beta_j > \beta_{\max} - kd_0\}$,

$$\begin{aligned}|I_k| &\ge (N-1)(1-(1-p_{\min})^k) + 1, \\ |J_k| &\ge (N-1)(1-(1-p_{\min})^k) + 1.\end{aligned} \tag{85}$$

Now, without loss of generality we assume that $d_{\max} > 2kd_0$. Then, by the optimality of $\boldsymbol{\beta}^*$ for $\mathcal{R}(\boldsymbol{\beta})$,

$$\begin{aligned}\log 2 =& \mathcal{R}(\mathbf{0}) \ge \mathcal{R}(\beta^*) \\ =& \frac{1}{\binom{N}{2}} \sum_{i,j:i\neq j} p_{ij}(t)\log(1+\exp(\beta_j^* - \beta_i^*)) \\ \ge& \frac{1}{\binom{N}{2}} \sum_{i\in I_k}\sum_{j\in J_k} p_{\min}\log(1+\exp(d_{\max} - 2kd_0)) \\ \ge& 2p_{\min}(1-(1-p_{\min})^k)^2(d_{\max} - 2kd_0).\end{aligned} \tag{86}$$

Thus, $d_{\max} \le \frac{\log 2}{2p_{\min}(1-(1-p_{\min})^k)^2} + 2kd_0$ for any k. Plugging in $k = \lceil \log(\frac{1}{p_{\min}}) \rceil$, we get that

$$\begin{aligned}d_{\max} \le& \frac{\log 2}{2(1-1/e)^2 p_{\min}} \\ &+ 2\log\left(\frac{2}{p_{\min}} - 1\right)\left(\log\left(\frac{1}{p_{\min}}\right) + 1\right) \\ \le& \frac{1}{p_{\min}} + C,\end{aligned} \tag{87}$$

for some universal constant C since the derivative of $2\log(2x-1)(\log x + 1)$ is positive and converges to 0 as $x \to \infty$. Then $2\log(2x-1)(\log x+1)$ has a upper-bounding tangent line with slope $1 - \frac{\log 2}{2(1-1/e)^2}$, and C is its y-intercept. This also yields that

$$\max_{i,j:i\neq j} \exp(\beta_i^* - \beta_j^*) \le C_p \exp\left(\frac{1}{p_{\min}}\right), \tag{88}$$

for some universal constant C_p. In particular, $1 < C_p < 1.5$.

9.2 Tuning kernel bandwidth in practical settings

As noted in Section 2.2, the bandwidth $h \in \mathbb{R}_{>0}$ of the kernel function serves as an effective global smoothing parameter between subsequent time periods and allows to borrow information across contiguous time points. Increasing h, all else held constant, leads to parameter estimates (and hence the derived global rankings) becoming "smoothed" together across time.

Naturally the question remains on how to *tune* h in practical applications in a principled data-driven manner. This is a fundamentally challenging question not just in our problem but, more generally, in nonparametric regression. Here we present a way to tuning h with some degree of objectivity based on leave-one-out cross-validation (LOOCV).

In general settings where we have independent and identically distributed (i.i.d.) samples, LOOCV assesses the performance of a predictive model on a single held-out i.i.d. sample. In our case, each pairwise comparison can be considered an i.i.d. sample if we take the compared teams and the time point on which they are compared as covariates. Remember that (i_m, j_m, t_m) denotes m-th pairwise comparison where team i_m won against team j_m at time point t_m for $m = 1, \ldots, M$. Then, for a given smoothing penalty parameter h, LOOCV is adapted to our estimation approach as follows:

1. For $m = 1, \ldots, M$, given $h > 0$:
   (a) fit our model with kernel bandwidth h on the dataset with the m-th comparison held-out;
   (b) calculate the negative log-likelihood (nll) of the previous solution to (i_m, j_m, t_m).
2. Take the average of the negative log-likelihoods to obtain nll_h as a loss in the predictive performance of time-varying Bradley-Terry estimator for given h on our dataset.
3. Choose the bandwith h^* with the smallest nll_h value.

We apply this data-driven methodology to the experiments and real-life application in Section 6 and Section 7.

9.3 Details of Experiments

Here we explain some details of the setting of the numerical experiments in Section 6.

9.3.1 Bradley-Terry Model as the True Model

We set the number of teams $N = 50$ and the number of time points $M = 50$. We set $n_{ij}(t) = 1$ for all $i, j \in [N]$ and $t \in [M]$.

For the Gaussian process to generate a path for $\boldsymbol{\beta}_i^*$ at $t = 1, \ldots, M$, we use the same covariance matrix $\Sigma_i = \Sigma \in \mathbb{R}^{M \times M}$ for all $i \in [N]$, and Σ is set to be a Teoplitz matrix defined by

$$\Sigma_{ij} = 1 - M^{-\alpha}|i - j|^r,$$

and in our experiment we set $(\alpha, r) = (1, 1)$. The mean vector is set to be a constant over time, i.e., $\mu_i(t) = u_i$ for $t = 1, \ldots, M$, and $u_1, \ldots, u_N$ are *i.i.d.* generated from uniform distribution on $[0, 1]$.

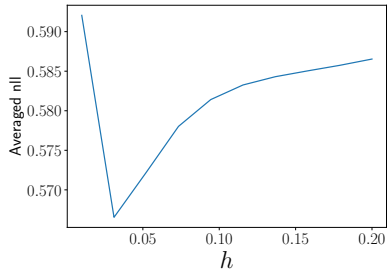

Figure 3: LOOCV curve of our Dynamic Bradley-Terry model fitted with Gaussian kernel. y-axis: averaged negative log-likelihood. The optimal h^* is 0.03.

Fig. 3 shows the curve of LOOCV of our dynamic Bradley-Terry model fitted with a Gaussian kernel in one repetition of our experiment. The curve is for the setting here and for the agnostic model setting the CV curve has similar shape. The curve shows a typical shape of CV curve for tuning parameter. The kernel bandwidth, h, with smallest nll_h is $h^* = 0.03$. The LOOCV procedure is described in Section 9.2.

9.3.2 Agnostic Model Setting

Again we set the number of teams $N = 50$, the number of time points $M = 50$, and $n_{ij}(t) = 1$ for all $i, j \in [N]$ and $t \in [M]$. The covariance matrix is also the same as section 9.3.1. The only difference lies in the mean vector. Now the mean vector is still constant over time, or $\mu_i(t) = u_i$ for $t = 1, \ldots, M$, but u_i's are generated in a following group-wise way:

1. Set the number of groups G and the index set of each group $I_1, \ldots, I_G$ so that $\sum_i |I_i| = N$. Set the base support to be $[0, b]$ and the group gap to be a.

2. For each $i \in \{1, \ldots, G\}$, generate u_j from Uniform$(a(i-1), a(i-1)+b)$ for all $j \in I_i$.

In our experiment we set $G = 5$ with each group containing two randomly picked indices, $b = 0.5$ and $a = 1.5$. Such group-wise generation intends to ensure that different teams have distinguishable perofrmance in pairwise comparison so that the ranking is more reasonable.

9.3.3 Running Time

Fig. 4 compares the time it takes to fit our model and the original Bradley-Terry model under 3 different settings, where N is the number of teams and M is the number of time points:

- Fix N, vary M.
- Fix M, vary N.
- Vary N and M together while keeping $N = M$.

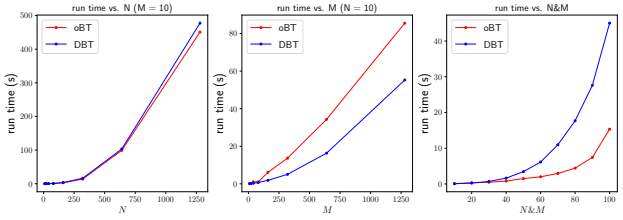

Figure 4: Comparison of running time of original Bradley Terry model (oBT) and our Dynamic Bradley Terry model (DBT). The values are averaged over 20 repetitions.

For our dynamic Bradley-Terry model, the running time here is measured for fitting the model with a given kernel parameter h, hence it contains the time cost of kernel smoothing step and the optimization step. In real applications, if one wants to select the best h from a range of values with cross-validation, then the total computation time would be approximately the running time here multiplied by the number of cross-validations.

The results in Fig. 4 shows that with all advantages our model can bring with, it does not cost much more in terms of computation time. Furthermore, when the number of time points M is large while N is relatively small, our model can cost even less time than the original Bradley-Terry model.

If one wants to do LOOCV to select h when N and M are huge, then it could take a long time to finish the whole procedure. However, in this case we observed in some extended experiments that with a pre-determined h in a reasonable range, our model can give fairly good estimate close to the one given by the best h^* selected by LOOCV. The supporting files of our experiments can be found in our GitHub repository[†].

9.3.4 MLE of the Bradley-Terry Model

Table 4 shows the frequency with which Condition 4.1 holds at a single time point for the original pairwise comparison data for different M and N, where $n_{ij}(t) = 1$ for all i, j, t. To be clear, here we just regard the matrix $\tilde{X}(t)$ in Condition 4.1 as the original data rather than the smoothed data, as it originally was in Ford (1957). Given $\{X(t), t \in [M]\}$, the frequency here refers to $\#\{t : \text{The condition holds for } X(t)\}/M$.

The data are generated as described in Section 6.1, and the frequency is averaged over 50 repetitions. When $N = M = 10$ and $n_{ij}(t) = 4$ for all i, j, t, the frequency arises to 0.988, illustrating how $n_{ij}(t)$ controls the sparsity of the game matrix and consequently whether Condition 4.1 holds or not.

(N, M)	(5,5)	(10,10)	(20,10)	(30,10)	(40,10)	(50,10)
Freq.	0.248	0.622	0.902	0.950	0.984	0.984

Table 4: Frequency that Condition 4.1 holds at a single time point for the original pairwise comparison data. $n_{ij}(t) = 1$.

As a comparison, under the same setting, for the kernel-smoothed pairwise comparison data, Condition 4.1 always holds in the experiment. This fact demonstrates the advantage of using kernel-smooth, and partly explains why in our experiments where the data is sparse our model performs the best.

The frequencies in Table 4 seem high for $N > 20$, but from a global perspective, the induced frequency that Condition 4.1 holds *for all* M time points could be much lower. Table 5 shows such frequency in some settings. Again the values are averaged over 50 repetitions. Remember that in these settings the condition always holds for kernel-smoothed data.

(N, M)	(10,10)	(20,10)	(30,10)	(40,10)	(50,10)	(60,10)
Freq.	0.02	0.44	0.62	0.86	0.84	0.88

Table 5: Frequency that Condition 4.1 holds for all M time points for the original pairwise comparison data. $n_{ij}(t) = 1$.

To make it clearer how $n_{ij}(t)$ affects the global connectivity, we make Table 6. In the table we fix $(N, M) = (10, 10)$.

[†]Code available at `https://github.com/shamindras/bttv-aistats2020`

$n_{ij}(t)$	1	2	4	6	8	10
Freq.	0.02	0.48	0.92	0.94	0.96	1.0

Table 6: Frequency that Condition 4.1 holds for all M time points for the original pairwise comparison data. $(N, M) = (10, 10)$.

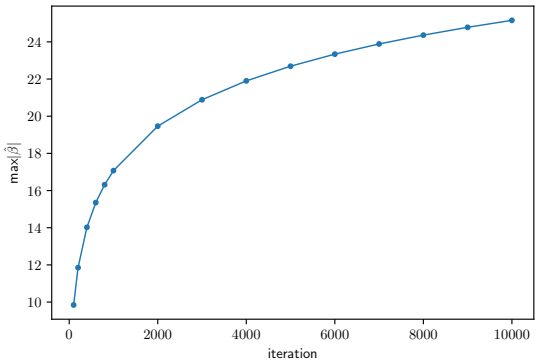

Figure 5: Divergence of $\max|\hat{\beta}|$ when MLE does not exist at some time points for the original Bradley-Terry model.

By direct inspection of the likelihood of the original Bradley-Terry model, it can be seen that, when the MLE does not exist, the norm of $\hat{\beta}$ will go to infinity if one uses gradient descent without any regularization. Fig. 5 shows an example where $N = M = 10$ and $n_{ij}(t) = 1$ for all i, j, t.