A DYNAMIC BRADLEY-TERRY APPROACH TO LEARNING TIME-VARYING LEADERBOARDS FROM SPARSE COMPARISONS IN SPORTS AND LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

State-of-the-art rankings of large language models, such as those of `Arena.AI`, are summarized in leaderboards that are continually updated over time based on a steady stream of head-to-head comparisons between models. Such time-varying leaderboards must learn from sparse and uneven data while accommodating temporary model absences. This challenge also arises in international sports, where, unlike in league competitions, teams play sporadically and with highly imbalanced schedules in mostly bilateral series. We develop a dynamic Bradley–Terry (BT) estimator for jointly estimating smooth, item-specific ability trajectories and time-varying contextual effects using a temporal basis representation and penalized logistic likelihood. By pooling information across time, the method can recover latent abilities during temporary gaps in an item's comparison history when the observations on either side provide sufficient information to identify the chosen temporal representation. We establish recovery guarantees and estimation rates that depend on the ability-curves' smoothness and the conditioning of the joint time–comparison design, thereby accounting for uneven comparison information across items and time. Extensive simulations examine performance across diverse temporal patterns, imbalanced comparison schedules, and temporary item absences, and demonstrate the benefits of temporal information pooling. Applications to international T20 cricket and LLM evaluations from `Arena.AI` (previously Chatbot Arena) further illustrate the methodology, yielding lower held-out errors than static and kernel BT methods under constructed interpolation holdouts.

1 INTRODUCTION

We study the estimation of time-varying abilities from pairwise comparisons that are sparse, unevenly distributed, and suffer from temporary item absences. The Bradley–Terry (BT) model relates comparison probabilities to latent abilities (Bradley & Terry, 1952), but its standard form assumes constant abilities. Dynamic extensions use local likelihood, spectral smoothing, or explicit evolution models (Bong et al., 2020; Karlé & Tyagi, 2023; Tian et al., 2024; Cattelan et al., 2013). Sparse designs pose distinct challenges for these approaches. Local methods may have too few comparisons or a disconnected comparison graph within a smoothing window, while evolution models bridge gaps through assumptions on temporal dynamics. We propose a global basis representation that pools information across time to recover relative abilities with limited local comparison information.

Our Contributions The logistic transformation of pairwise ability differences, proposed by Bradley & Terry (1952) for modeling win probabilities, is widely used in tournament-ranking problems. Although powerful, the standard static BT model cannot accommodate settings in which item abilities evolve over time. Such temporal variation is central to contemporary applications such as LLM evaluation, where models develop rapidly, as well as to team sports such as cricket.

Our main contribution is a centered basis representation of time-varying, item-specific abilities in the BT model with joint estimation by penalized likelihood. In this new framework we study identification and existence building on static BT analyses of estimator existence (Ford, 1957) and graph-dependent accuracy (Simons & Yao, 1999; Yan et al., 2012; Han et al., 2020). We characterize identification through the joint time–comparison design. We derive asymptotic bounds on integrated squared ability error and excess log loss that account for approximation error for general basis expansions and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

quantify the effect of weakening global comparison information. These are extensions of related work on estimation and ranking accuracy in the static BT model (Shah et al., 2016; Hendrickx et al., 2020; Chen et al., 2019; Gao et al., 2023). We adopt the sum-to-zero constraint, that is the item abilities sum to zero at each time, a standard approach in log-linear paired-comparison models (Fienberg & Larntz, 1976) that avoids choosing a reference item, and is known to minimize total asymptotic variance among linear identifying constraints for the static BT model (Wu et al., 2022).

Our second contribution is an extension of the proposed dynamic BT model that incorporates covariates. This is motivated by the connection of the proposed dynamic BT model to varying-coefficient regression, with the ability trajectories serving as coefficients of item-contrast indicators (Hastie & Tibshirani, 1993; Fan & Zhang, 2008). Building on paired-comparison models with covariates (Dittrich et al., 1998; Fan et al., 2024; Li & Li, 2025), we allow item-specific covariate effects to vary over time. Following classical penalized-likelihood methods (O'Sullivan et al., 1986; Wahba, 1990; Wood, 2017), we develop an estimation framework based on increasing-dimensional basis expansions and roughness penalties and study their roles in the dynamic BT setting.

Third, we conduct extensive simulations under modern time-varying pairwise comparison scenarios that face challenges such as temporary item absences, sparse comparisons, and time-varying contextual effects, demonstrating the advantages of the proposed method over competing approaches when there is limited nearby comparison information. We apply the proposed method to Chatbot Arena and T20 cricket datasets, and illustrate these gains in prediction over held-out intervals with few comparisons available at neighboring times.

Related Work on Dynamic BT Models Bong et al. (2020) estimate time-varying abilities by kernel-smoothing pairwise win counts and fitting an unpenalized BT model at each target time. The bandwidth controls how strongly comparisons at other times contribute, balancing temporal resolution against the information available for local estimation. Smoothing pools comparisons across time, although sparse observations and long absences can limit recovery. The authors establish time-specific conditions for existence and uniqueness and derive oracle bounds for estimation error and excess risk that allow model misspecification. Dynamic Rank Centrality (Karlé & Tyagi, 2023) and Kernel Rank Centrality (Tian et al., 2024) similarly combine temporal smoothing with spectral ranking and provide estimation guarantees. Other approaches model temporal variation through outcome histories or global trajectories. Cattelan et al. (2013) use exponentially weighted past outcomes to model home and away abilities while allowing ties. Krese & Štrumbelj (2021) represent ability trajectories through barycentric rational interpolation and Gaussian-process priors, while Araki et al. (2019) use piecewise quadratic splines to model age effects. Kickscore (Maystre et al., 2019) models latent scores through continuous-time Gaussian processes, uses approximate Bayesian inference, and incorporates context through a feature-based formulation.

2 METHODOLOGY

We first describe the dynamic BT model without covariates. Let J denote the number of items. For rows $r = 1, \dots, R$, let $i_r < j_r$ denote the compared items, $t_r \in [0, 1]$ the normalized comparison time, and W_{rA}, W_{rB} their respective decisive-win counts, with $m_r = W_{rA} + W_{rB} > 0$. Conditional on the observed time–matchup design $\mathcal{D} = \{(t_r, i_r, j_r, m_r)\}_{r=1}^{R}$, the rows are assumed independent, with $W_{rA} \mid \mathcal{D} \sim \text{Binomial}(m_r, p_r)$ and $\text{logit}(p_r) = \beta_{i_r}(t_r) - \beta_{j_r}(t_r)$. The total sample size is $n = \sum_r m_r$. Up to a parameter-independent factor, this likelihood is equivalent to the likelihood obtained by treating decisive battles as conditionally independent, with all battles within a row sharing the same comparison time and item pair. Comparison times and item pairs are treated as given, and win probabilities are conditional on a decisive outcome; comparisons labeled "tie" or "both bad" are excluded. For identifiability, we impose the centering constraint $\mathbf{1}_J^\top \boldsymbol{\beta}(t) = 0$, where $\boldsymbol{\beta}(t) = \{\beta_1(t), \dots, \beta_J(t)\}^\top$. We target the invariant pairwise log odds $\Delta_{ij}(t) = \beta_i(t) - \beta_j(t)$ and win probabilities $p_{ij}(t) = \sigma\{\Delta_{ij}(t)\}$, where $\sigma(a) = \{1 + \exp(-a)\}^{-1}$.

Note that the unknown true population ability curves lie in $L^2([0, 1], \mu)$, where μ is the time law. We next consider modeling the ability curves using a M-dimensional basis in $L^2([0, 1], \mu)$. Let $\mathcal{S}_M = \text{span}\{\phi_1, \dots, \phi_M\}$ be an M-dimensional subspace of $L^2([0, 1], \mu)$ with linearly independent, measurable basis functions and specified pointwise versions for evaluation. Write $\boldsymbol{\phi}(t) = (\phi_m(t))_{m=1}^M$ and choose an orthonormal item-contrast matrix $C \in \mathbb{R}^{J\times(J-1)}$ satisfying $C^\top C = I_{J-1}$ and $C^\top \mathbf{1}_J = 0$. Every centered ability vector in $\mathcal{S}_M^J$ has a unique representa-

tion $\boldsymbol{\beta}(t) = CA\boldsymbol{\phi}(t)$, where $A \in \mathbb{R}^{(J-1)\times M}$. With $\theta = \mathrm{vec}(A^\top)$ in column-major order and $d_{ij} = e_i - e_j$, the pairwise log odds become $\Delta_{ij}(t) = z_{ij}(t)^\top\theta$, where $z_{ij}(t) = C^\top d_{ij} \otimes \boldsymbol{\phi}(t)$. This parameterization enforces centering without selecting a reference item and accommodates natural splines, Fourier systems, wavelets, and other finite bases.

We now present penalized likelihood estimation of the parameters in the BT model. For a symmetric positive-semidefinite penalty matrix Ω_M, define $P_M = I_{J-1} \otimes \Omega_M$ and $J_M(\boldsymbol{\beta}_\theta) = \theta^\top P_M \theta$. For example, $(\Omega_M)_{mm'} = \langle \phi''_m, \phi''_{m'} \rangle$ yields an integrated squared-curvature penalty when the basis functions have square-integrable second derivatives. Frequency or scale weights provide alternative quadratic penalties. Writing $z_r = z_{i_r j_r}(t_r)$, the normalized negative log likelihood, up to a parameter-independent constant, and the penalized estimator for any $\rho \geq 0$ are given by:

$$\widehat{\theta}_\rho \in \underset{\theta\in\mathbb{R}^{(J-1)M}}{\operatorname{argmin}} \{\mathcal{L}_n(\theta) + \rho J_M(\boldsymbol{\beta}_\theta)\} \text{ and } \mathcal{L}_n(\theta) = \frac{1}{n}\sum_{r=1}^{R} \left[m_r \log\{1 + \exp(z_r^\top\theta)\} - W_{rA} z_r^\top\theta\right]. \tag{1}$$

Implementation details are provided in Appendix A. Large-sample conditions for the uniqueness of the proposed estimator $\widehat{\theta}_\rho$ are provided in Section 3, and a deterministic finite-sample existence and uniqueness criterion is given in Proposition 2 of the Appendix.

Dynamic BT with covariates Let $x_{ri} = x_i(u_r, c_r) \in \mathbb{R}^p$ encode user or population information u_r and comparison context c_r relevant to item i. Following the varying-coefficient framework (Hastie & Tibshirani, 1993; Fan & Zhang, 2008), define $\alpha_i(x,t) = \beta_i(t) + x^\top\gamma_i(t)$, $\gamma_i(t) = g(t) + \widetilde{\gamma}_i(t)$ and $\sum_i \widetilde{\gamma}_i(t) = 0$, where $g(t)$ is the common covariate effect and $\widetilde{\gamma}_i(t)$ is the item-specific deviation. Conditional on the feature-augmented design, rows are assumed independent, with each item's feature vector held fixed across battles within a row. The decisive-win probability is $\sigma(\eta_r)$, where

$$\eta_r = \beta_{i_r}(t_r) - \beta_{j_r}(t_r) + (x_{ri_r} - x_{rj_r})^\top g(t_r) + x_{ri_r}^\top \widetilde{\gamma}_{i_r}(t_r) - x_{rj_r}^\top \widetilde{\gamma}_{j_r}(t_r).$$

When the context is shared, so that $x_{ri_r} = x_{rj_r}$, the common effect cancels. Alternative-specific features can differ between the two items in the same comparison: a home-status indicator, for example, is 1 for the home team and 0 for its opponent. The common effect then need not cancel, so setting $g = 0$ imposes an additional modeling restriction. We impose $g = 0$ in the estimator below and all reported contextual fits. Expanding the baseline and item-specific contextual coefficient functions in the same temporal basis and writing $\widetilde{x}_{ri} = (1, x_{ri}^\top)^\top$ gives the augmented design $z_r^c = \{C^\top e_{i_r} \otimes \widetilde{x}_{ri_r} - C^\top e_{j_r} \otimes \widetilde{x}_{rj_r}\} \otimes \boldsymbol{\phi}(t_r)$. Let $\theta_c \in \mathbb{R}^{(J-1)(p+1)M}$ collect the basis coefficients and $\eta_r = (z_r^c)^\top\theta_c$. Let Q_c be a positive-semidefinite matrix specifying temporal roughness penalties for the baseline and contextual coefficients and additional ridge shrinkage for the contextual coefficients. Its detailed construction is given in Appendix A. We estimate the coefficients by:

$$\widehat{\theta}_c[Q_c] \in \underset{\theta_c}{\operatorname{argmin}} \left\{\mathcal{L}_n^c(\theta_c) + \theta_c^\top Q_c \theta_c\right\} \text{ where, } \mathcal{L}_n^c(\theta_c) = \frac{1}{n}\sum_{r=1}^{R} \left[m_r \log\{1 + \exp(\eta_r)\} - W_{rA}\eta_r\right]. \tag{2}$$

Within the restricted finite-basis model, the coefficient vector is identifiable from the observed design precisely when the matrix with rows $(z_r^c)^\top$ has full column rank. Appendix A.2 and Proposition 3 provide details on the optimization in (2).

Tuning For each study, we fix the basis family and construct bases at prespecified dimensions using training times. We select dimensions and penalties by time-block cross-validation on the training data, using the same folds for all tuned methods. The quadratic penalty does not perform sparse coefficient selection. Cross-validation comparisons require both items to appear in the corresponding training sample. The experimental sections specify the interpolation holdouts, and Appendix A gives the tuning protocol, including the treatment of fitting failures. For an assessment set $\mathcal{S}$, the per-battle log loss of method a, denoted by $\mathrm{LL}_{\mathcal{S}}^{(a)}$, is evaluated as $-\sum_{r\in\mathcal{S}}[W_{rA}\log\widehat{p}_r^{(a)} + W_{rB}\log\{1-\widehat{p}_r^{(a)}\}]/\{\sum_{r\in\mathcal{S}} m_r\}$, where $\widehat{p}_r^{(a)}$ is the probability predicted by method a that item i_r defeats item j_r in assessment row r. Methods are compared on a common assessment set, with failures reported separately. This loss weights comparisons according to the observed schedule and differs from uniform-pair simulation risk.

3 STATISTICAL PROPERTIES

We theoretically characterize how temporal approximation and comparison information affect the performance of our proposed estimator. For ease of presentation, we state our main results for the dynamic BT model without covariates. These results depend on the structure of the comparison graph. We provide analogous results for the model with covariates in Appendix B.9. For the model with covariates, we study the joint time–pair–covariate design because the comparison graph alone does not capture all relevant information.

We first characterize the information available in the fitted parametric space and study the uniqueness and stability of the penalized log-likelihood estimator when the comparison graph may be disconnected at individual times. We then establish asymptotic rates for the integrated squared ability error and excess log loss. Here, J is the fixed number of items, n is the total number of decisive comparisons, and $M = M_n$ is the temporal basis dimension. We study the regime $n \to \infty$ with $M_n \to \infty$ so that the temporal basis can increasingly well approximate the relative ability curves. However, the rate at which M_n can grow while retaining estimability of the coefficients depends on the sparsity of the comparison graph, as described below. We illustrate the implications of these estimation results for well-conditioned and polynomially weakening designs. We also show that the exact fixed-basis model (M fixed) follows as a special case of the general result. All proofs are provided in Appendix B. For the rate analysis, we impose a random-design assumption: at each sample size, the individual comparisons $(T_\ell, I_\ell, J_\ell, Y_\ell)$, $\ell = 1, \dots, n$, are independent and identically distributed, with $I_\ell < J_\ell$ and $\Pr(Y_\ell = 1 \mid T_\ell = t, I_\ell = i, J_\ell = j) = p^0_{ij}(t) = \sigma\{\beta^0_i(t) - \beta^0_j(t)\}$. The true trajectories are centered and square-integrable on $[0, 1]$ with fixed measurable versions. We allow the distribution of (T, I, J) to depend on n. We denote expectation under this distribution by $\mathbb{E}_n$ and its time marginal by μ_n. Let $q_{ij,n}(t) = \Pr(I = i, J = j \mid T = t)$ and $L_{q,n}(t) = \sum_{i<j} q_{ij,n}(t) d_{ij} d_{ij}^\top$, where $d_{ij} = e_i - e_j$. We distinguish integrated ability error from error over the comparisons sampled by the design. For a vector of functions h, define the integrated ability loss $\mathcal{A}$ and the design-weighted pairwise loss $\mathcal{P}_n$ by $\mathcal{A}(h) = \sum_{i=1}^J \int_0^1 h_i(t)^2 \, dt$ and

$$\mathcal{P}_n(h) = \int_0^1 \sum_{i<j} q_{ij,n}(t)\{h_i(t) - h_j(t)\}^2 \, d\mu_n(t) = \int_0^1 h(t)^\top L_{q,n}(t) h(t) \, d\mu_n(t).$$

Here, $\mathcal{A}$ measures squared ability error with respect to Lebesgue measure, whereas $\mathcal{P}_n$ weights pairwise log-odds error according to the comparison design and its time marginal μ_n. Under the time-density bounds imposed for the smoothness corollaries, $\mathcal{A}$ is uniformly equivalent to its μ_n-weighted analogue. The functional $\mathcal{P}_n$ is a squared seminorm on unrestricted trajectories, because changes confined to unobserved comparisons need not affect it. Let $\mathcal{B}_n = \{\boldsymbol{\beta} \in \mathcal{S}^J_{M_n} : \mathbf{1}_J^\top \boldsymbol{\beta}(t) = 0\}$ and $d_n = (J-1)M_n$. We next define κ_n which measures the information provided by the joint time–comparison design on the fitted trajectory space: $\kappa_n = \inf_{h \in \mathcal{B}_n : \mathcal{A}(h) > 0} \mathcal{P}_n(h)/\mathcal{A}(h)$. Thus, κ_n measures the weakest amount of information that the joint time–comparison design provides about any direction in the chosen space of ability trajectories. Note that the condition $\kappa_n > 0$ means that every nonzero direction $h \in \mathcal{B}_n$ has $\mathcal{P}_n(h) > 0$, so every such change affects the available comparisons. Thus, $\kappa_n > 0$ ensures identifiability of the parameters in the dynamic BT model within the fitted space. This condition does not require $L_{q,n}(t)$ to have a positive spectral gap at every time. For example, observations on either side of an item's temporary absence may jointly identify the coefficients of a sufficiently restrictive temporal representation.

For a generic pair–time point (t, i, j), let $z_{ij}(t)$ denote the design vector and define $G_{M_n} = \int_0^1 \sum_{i<j} q_{ij,n}(t) z_{ij}(t) z_{ij}(t)^\top \, d\mu_n(t)$, $H_{M_n} = \int_0^1 \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^\top \, dt$, and $V_n = I_{J-1} \otimes H_{M_n}$. For these matrices and the Lebesgue-integrated loss, we assume $\phi_m \in L^2([0,1], dt)$ and $0 < \lambda_{\min}(H_{M_n}) \le \lambda_{\max}(H_{M_n}) < \infty$ for all n. The following result relates κ_n to the smallest eigenvalue of the joint time–comparison design. Thus, κ_n connects sparsity in this design to identifiability and the stability of our estimator.

Proposition 1 (Identifiability)**.** *In the dynamic BT model, $\kappa_n = \lambda_{\min}\{V_n^{-1/2} G_{M_n} V_n^{-1/2}\}$. The centered basis coefficients are identifiable from the population comparison probabilities supported by the design if and only if $G_{M_n} \succ 0$, or equivalently, $\kappa_n > 0$. For an L^2-orthonormal temporal basis, $V_n = I_{d_n}$ and $\kappa_n = \lambda_{\min}(G_{M_n})$.*

Only positivity is needed for this population identification statement. Stable estimation requires more. If $\kappa_n \downarrow 0$ as $n \to \infty$, as may occur with very sparse designs, distinct trajectories can produce increasingly similar probabilities for the available comparisons. The rates below track this loss of information explicitly. Positivity of κ_n alone does not ensure a unique finite empirical fit. Proposition 2 in the appendix provides a deterministic fitting criterion, while the asymptotic conditions below ensure empirical existence and uniqueness. To quantify the estimation error, we separate approximation error from sampling error. Choose a deterministic centered approximant $\boldsymbol{\beta}_n^* \in \mathcal{B}_n$ and write $a_n^2 = \mathcal{A}(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}^0)$ and $\tilde{a}_n^2 = \mathcal{P}_n(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}^0)$. The first measures approximation error over the full time interval and all items, whereas the second measures its effect on sampled comparisons. When the truth lies in the fitted space, both vanish by taking $\boldsymbol{\beta}_n^* = \boldsymbol{\beta}^0$. Let $\widehat{\boldsymbol{\beta}}_n$ be the trajectory corresponding to $\widehat{\theta}_{\rho_n}$ in Section 2, with deterministic $\rho_n \geq 0$ and the positive-semidefinite quadratic penalty J_{M_n}. Define

$$w_n = \tilde{a}_n^2 + d_n n^{-1} + \rho_n J_{M_n}(\boldsymbol{\beta}_n^*), \qquad \ell_n^2 = \operatorname*{ess\,sup}_{(t,i,j)} z_{ij}(t)^\top G_{M_n}^{-1} z_{ij}(t).$$

The three terms in w_n account for approximation, coefficient estimation, and regularization, respectively. The last term in w_n depends on the penalty of the approximant and does not represent the exact smoothing bias. The essential supremum defining ℓ_n is taken over pair–time points (t, i, j) in the support of the sampling distribution; it need not include item–time combinations that are never observed. Once the basis choice and penalty structure are fixed, we impose the following conditions.

(A1) Assume, for every sufficiently large n, that $\int_0^1 \sum_{i<j} q_{ij,n}(t)\|z_{ij}(t)\|_2^2 \, d\mu_n(t) < \infty$, $G_{M_n} \succ 0$, and $\ell_n < \infty$. The dimension M_n may be fixed or increasing.

(A2) For a constant B independent of n, the centered approximant satisfies $a_n, \tilde{a}_n < \infty$ and $\operatorname{ess\,sup}_{(t,i,j)} |\beta_{n,i}^*(t) - \beta_{n,j}^*(t)| \leq B$.

(A3) The approximation, complexity, and penalty satisfy $\ell_n^2 w_n \to 0$ as $n \to \infty$.

Condition A3 limits the size of fitted log-odds perturbations on the observed support, ensuring sufficient curvature of the logistic likelihood. It is a sufficient condition, not a necessary condition for estimation. No time-by-time connectivity assumption or exact finite-basis specification is imposed. The next theorem first controls the error in observed contrasts and then extends this control to the full ability trajectories. This extension incurs a factor of $1/\kappa_n$ because a weak global design may conceal substantial differences between trajectories. The approximation term remains outside this factor because the true trajectories need not belong to the fitted space.

Theorem 1 (Estimation error)**.** *Under Conditions A1–A3, the penalized negative log-likelihood criterion has a unique minimizer with probability tending to one and as $n \to \infty$,*

$$\mathcal{P}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) = O_p(w_n), \quad \mathcal{A}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) = O_p\left(a_n^2 + \frac{w_n}{\kappa_n}\right). \tag{3}$$

The above result shows that full-trajectory L^2 consistency holds whenever $a_n^2 + w_n/\kappa_n \to 0$. We describe this asymptotic condition for different designs below. Next, we assess prediction for a new comparison drawn independently from the same sampling distribution as the training data. Let (t, i, j, y) denote such a test comparison, where $i < j$ are the two randomly selected items and $y = 1$ indicates that item i defeats item j. Define the log loss $\mathcal{R}_n(\boldsymbol{\beta}) = \mathbb{E}_n[-y \log p_{\boldsymbol{\beta},ij}(t) - (1-y)\log\{1 - p_{\boldsymbol{\beta},ij}(t)\}]$ where, $p_{\boldsymbol{\beta},ij}(t) = \sigma\{\beta_i(t) - \beta_j(t)\}$. The corresponding excess log-loss risk is

$$\mathcal{E}_n := \mathcal{R}_n(\widehat{\boldsymbol{\beta}}_n) - \mathcal{R}_n(\boldsymbol{\beta}^0) = \int_0^1 \sum_{i<j} q_{ij,n}(t)\, \mathrm{KL}\big\{\mathrm{Ber}(p_{ij}^0(t)), \mathrm{Ber}(\widehat{p}_{ij,n}(t))\big\}\, d\mu_n(t), \tag{4}$$

where $\widehat{p}_{ij,n} = p_{\widehat{\boldsymbol{\beta}}_n,ij}$ and KL denotes the Kullback–Leibler divergence. We also define the design-weighted squared probability error as $\mathcal{Q}_n = \int_0^1 \sum_{i<j} q_{ij,n}(t)\{\widehat{p}_{ij,n}(t) - p_{ij}^0(t)\}^2 \, d\mu_n(t)$. These losses weight each pair–time combination according to its sampling frequency. Consequently, we can control these losses using the observed-contrast error without reconstructing the full ability trajectories. The following result makes this distinction explicit. Consistency is defined in terms of excess log loss because the risk at the true probabilities includes irreducible Bernoulli uncertainty and is generally nonzero.

Theorem 2 (Prediction log-loss error)**.** *Under the assumptions of Theorem 1,*

$$\mathcal{Q}_n \le \mathcal{P}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)/16, \qquad \mathcal{E}_n \le \mathcal{P}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)/8. \tag{5}$$

Consequently, $\mathcal{Q}_n = O_p(w_n)$ *and* $\mathcal{E}_n = O_p(w_n)$ *as* $n \to \infty$.

The error bounds in Theorems 1 and 2 do not require a particular temporal basis. To express them in terms of the smoothness of the ability curves, we now consider basis families with explicit approximation and inverse bounds. For these specializations, assume that $\mu_n(dt) = f_n(t)\,dt$ with $0 < \underline{f} \le f_n(t) \le \overline{f} < \infty$ almost everywhere, uniformly in n. Suppose also that, for a fixed integer $s \ge 2$, the centered true curves satisfy $\sum_i \|\beta_i^0\|^2_{W^{s,2}[0,1]} \le R_0^2$ and $\max_i \|\beta_i^0\|_\infty \le B_0$, where the constants are independent of n. Here $W^{s,2}$ consists of functions whose weak derivatives through order s are square-integrable. Assume that the chosen spaces admit centered approximants satisfying

$$\mathcal{A}(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}^0) = O(M_n^{-2s}), \quad \max_i \|\beta_{n,i}^* - \beta_i^0\|_\infty = o(1), \quad J_{M_n}(\boldsymbol{\beta}_n^*) = O(1), \tag{6}$$

and, for a constant C_0 independent of n, $\max_i \|h_i\|_\infty \le C_0\{M_n \cdot \mathcal{A}(h)\}^{1/2}$ for all $h \in \mathcal{B}_n$. For a concrete example, consider a quasi-uniform polynomial spline family of fixed degree $d_0 \ge s$, with an integrated squared-derivative penalty of order $1 \le r_0 \le s$ and the corresponding approximation bounds (Huang & Su, 2021).

Under these conditions, a sufficient condition for consistent estimation and prediction is

$$M_n v_n \kappa_n^{-1} \to 0 \text{ where, } v_n = M_n^{-2s} + (J-1)M_n n^{-1} + \rho_n. \tag{7}$$

In such regimes, squared L^2 error is $O_p(v_n/\kappa_n)$ whereas the excess log loss is $O_p(v_n)$. The condition $\liminf_{n\to\infty} \kappa_n > 0$ means that the joint time-comparison design remains uniformly informative on the fitted trajectory spaces. A sufficient condition, independent of the basis dimension, is that the second eigen-value of $L_{q,n}(t) \ge \lambda_0 > 0$ for every t, uniformly in n. The time-density lower bound then gives $\kappa_n \ge \underline{f}\,\lambda_0$. In particular, if $q_{ij,n}(t) \ge \underline{q} > 0$ uniformly over pairs, for all t and n, then $\kappa_n \ge J\underline{f}\,\underline{q}$. Deterioration can occur when comparison information becomes increasingly weak in directions represented by the fitted spaces, including through vanishing sampling probabilities or item-absence intervals. Instantaneous connectivity is not necessary, since temporal pooling can maintain global conditioning, as illustrated in Appendix B.7. Under uniform conditioning, the following corollary gives the standard nonparametric estimation rate.

Corollary 1. *Suppose* $\liminf_{n\to\infty} \kappa_n > 0$*. Then, under the preceding time-density, smoothness, approximation, and inverse conditions, the proposed methodology with* $M_n \asymp n^{1/(2s+1)}$ *and* $0 \le \rho_n = O\{n^{-2s/(2s+1)}\}$ *satisfies*

$$\mathcal{A}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) = O_p\{n^{-2s/(2s+1)}\}, \text{ and, } \mathcal{Q}_n,\ \mathcal{E}_n = O_p\{n^{-2s/(2s+1)}\}. \tag{8}$$

Note that the above results show that, when $s = 2$, the squared-error and excess-risk rates are both $n^{-4/5}$. We next allow the restricted comparison information to weaken with the basis dimension, subject to a polynomial lower bound. The following result shows that, under the same tuning parameters, prediction for the sampled comparisons retains the same rate, whereas recovery of the full trajectories incurs an additional conditioning cost.

Corollary 2. *Suppose* $\kappa_n \ge c_0 M_n^{-\alpha}$ *for* $c_0 > 0$ *and* $0 \le \alpha < 2s - 1$*. Then, under the preceding time-density, smoothness, approximation, and inverse conditions, with* $M_n \asymp n^{1/(2s+1)}$ *and* $0 \le \rho_n = O\{n^{-2s/(2s+1)}\}$*, the proposed estimator exists uniquely with probability tending to one, and*

$$\mathcal{A}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) = O_p\{n^{-(2s-\alpha)/(2s+1)}\}, \text{ and, } \mathcal{Q}_n,\ \mathcal{E}_n = O_p\{n^{-2s/(2s+1)}\}. \tag{9}$$

For example, if $s = 2$ and $\kappa_n \gtrsim M_n^{-1}$, the full-trajectory squared-error bound is $O_p(n^{-3/5})$, while the design-weighted excess log loss is $O_p(n^{-4/5})$. Appendix B.7 gives examples illustrating how local disconnection can coexist with global identification. It also illustrates the role of observation gaps relative to the basis resolution. Unlike Corollary 1, Corollary 2 gives a faster upper rate for excess log loss than for integrated squared ability error. This difference arises because the two evaluation criteria are not equivalent when the design is not well-conditioned. Full-trajectory L^2 error counts errors for every item throughout time, including periods during which the item has no comparisons. Design-weighted log loss and squared probability error assign zero weight to comparisons that the design never samples.

Remark 1. For a fixed design distribution and a fixed bounded finite-dimensional basis with $G_M \succ 0$, suppose the true centered ability curves lie in the fitted space and the fixed penalty is finite at the truth. Then Theorems 1–2 imply that our proposed method with $\rho_n = O(n^{-1})$, including $\rho_n = 0$, satisfies $\mathcal{A}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0), \mathcal{Q}_n, \mathcal{E}_n = O_p(n^{-1})$. Appendix B.6 provides further discussion.

4 Simulation Studies

We conduct several empirical experiments to examine the performance of the proposed approach. In particular, we focus on two scenarios: how well can temporal pooling recover pairwise preferences during temporary item absences when local-smoothing approaches may not be effective, and when does adjusting for covariates improve estimation in dynamic preferences; additional experiments on other scenarios are relegated to Appendix C. We compare the proposed dBT estimator with static BT and kernel BT (Bong et al., 2020) as baselines. In the contextual setting, we compare the proposed contextual dBT estimator with the static contextual BT estimator, the context-free dBT estimator, and kernel BT. For each configuration, we use 500 simulated datasets. For tuning parameters, we use time-block cross-validation on the training data, followed by refitting on that sample. For details on the simulation designs, we refer to Appendix C.1.

Throughout the simulation figures in the main text and supplement, MSE and computation time use logarithmic axes; ordering error uses a linear scale. Diamonds in all boxplots denote arithmetic means computed on the original scale.

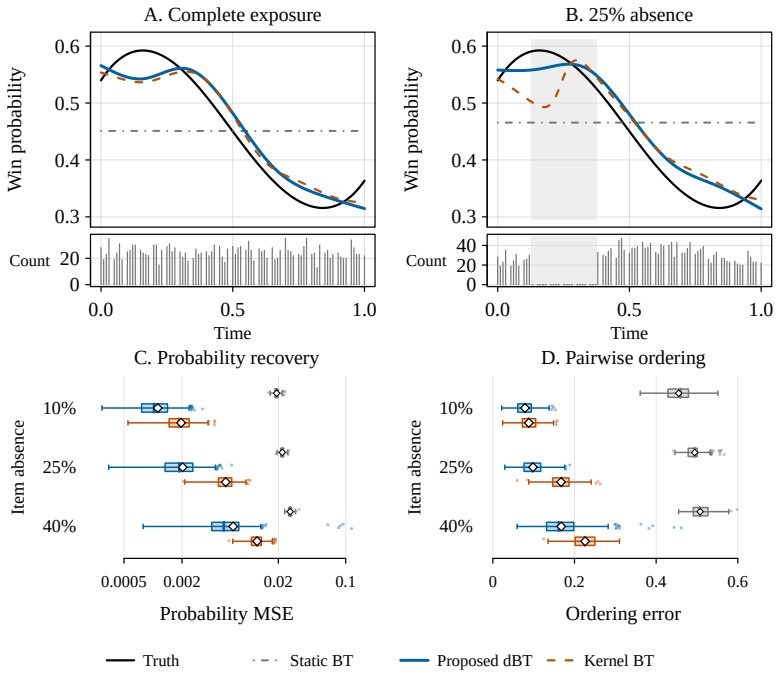

Figure 1: Recovery during item absences. A–B: True and fitted probabilities that item 3 defeats item 1 in simulated dataset 1, under complete exposure and a 25% absence of item 3 (gray shading). Count strips show item 3's training appearances. C–D: Probability MSE and pairwise ordering error across 500 datasets, averaged over all absence-affected pairs and times.

Estimation under temporary item absences. We first study how well relative preferences can be recovered during an item's absence using observations before and after that interval, a scenario which is challenging for local-smoothing based approaches. Our target is the estimation of $p_{ij}^0(t) = \sigma\{\beta_i^0(t) - \beta_j^0(t)\}$, the probability that item i defeats item j at time t. We consider eight teams, with centered ability trajectories constructed using polynomial trends and smooth crossings, together with one localized and one nearly constant component (see Appendix C.1 for details). Under fixed total comparison budget, we compare complete exposure with staggered absences of increasing duration, where two items remain available throughout and the other six are unobserved on the given intervals.

The evaluation set comprises all grid times and unordered pairs for which at least one item is within its designated absence interval, including training-grid times; in the complete-exposure setting, designated intervals are used for evaluation without removing observations. We report the mean squared error (MSE) of the estimated win probabilities and pairwise ordering error on this evaluation set, where ordering error measures recovery of the sign of $\beta_i^0(t) - \beta_j^0(t)$, with true ties excluded and estimated ties assigned half an error.

Figure 1 shows that the proposed method has lower mean probability MSE and ordering error at each absence duration, and errors increase with duration for all temporal estimators. The five largest dBT errors at 40% absence decrease substantially with stronger penalties at the same spline dimension, but we retain the original cross-validation choices. Additional experiments in Appendix C.3 illustrate that recovery also depends on the number and allocation of comparisons outside the absence intervals. We report the centered-ability and rank trajectories, log-odds MSE, and sensitivity analyses using a common evaluation region across durations in the appendix. In Appendix C.7 we show that the opponent-averaged team win-ratio MSE follows the same order among the methods.

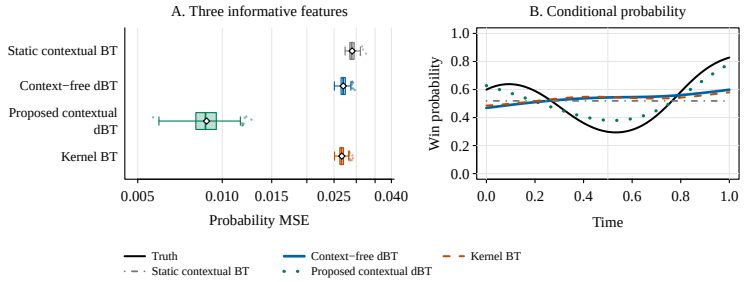

Figure 2: Recovery with covariates. A: Probability MSE across 500 datasets, averaged over held-out times, unordered pairs, and the distribution of the contexts. B: True and fitted probabilities that item 1 defeats item 5 in dataset 1, at $x = (1, 1, 1)$.

Estimation with covariates. Next, we study whether joint adjustment for time and context improves recovery over either adjustment alone. As before, we target $p_{ij}^0(x, t)$, the win probability now conditional on the covariate domain $x \in \{-1, 1\}^3$ and time t. We compare the eight items at 51 equally spaced times, with 60 comparisons per time and uniform pair sampling. We use the mixed smooth curves for the baseline abilities and allow all three context coefficients to vary over time. We generate features with balanced marginals, correlation 0.5 between the first two, and independence of the third from the other two. We average probability MSE uniformly over held-out times and unordered pairs, and take its expectation over the context distribution used to generate the data (Appendix C.6). Figure 2 illustrates that contextual dBT has lower probability MSE than both the context-free estimators and static contextual BT. Hence, in the presence of relevant covariates, joint time–context improves recovery. In Appendix C.5 we provide additional experiments where we study scenarios with irrelevant features or missing context categories. In Appendix C.7, we show that opponent-averaged team win-ratio MSE leads to the same conclusions.

5 Real Data Applications

5.1 Chatbot Arena data

Models with temporal variation. We use 14232 decisive battles from April–July 2025 among 16 endpoint-supported Arena 140K models (Zheng et al., 2023; Chiang et al., 2024). For evaluation of the proposed method, we select eight models with the largest noise-adjusted temporal variation in weekly win fractions within a fixed *discovery sample* of 10295 battles that are not in the test set but are used for fitting (Appendix D.2). For training, we retain all the 16 models as eligible opponents.

Low-information interpolation. We evaluate 15 seven-day windows for the selected models, and exclude discovery-sample battles in the assessment. Each test participant appears in comparisons before and after its holdout. In each replication, we sample one eligible window uniformly for each selected model and fit each method separately for that holdout, using interlaced date cross-validation for tuning. We compute average win-ratio MSE by equally weighting the selected models and averaging predicted win probabilities over their respective test opponents. Across 100 draws, we obtain 52 distinct sets of holdout windows for sensitivity analysis to window selection within the

same dataset. Figure 3 illustrates that the proposed dBT approach has lower win-ratio MSE than both comparators in all 100 holdout collections. Log loss and Brier scores also favor the proposed method on average. We relegate the full implementation details and additional illustrations to Appendix D.2.

5.2 T20 Cricket data

Prediction with limited nearby matches. International cricket provides a natural example of uneven temporal exposure. We use men's international T20 matches from Cricsheet (Rushe, 2026), excluding ties and no-results. From an initial cohort of 20 teams with at least 50 decisive matches and eight years of participation, we retain eight teams with at least 200 appearances within that cohort: Australia, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, and West Indies. Their 763 head-to-head matches span February 2005–July 2026. We include the implementation and holdout construction details and additional results in Appendix D.3.

Low-information interpolation. We consider short participation episodes with little nearby training information, identifying 75 eligible three-match episodes with only 0–3 other focal-team matches within 90 days of their midpoint. Each test participant remains observed before and after its withheld matches. In each of 100 replications, we randomly select one eligible episode per team. For each selected episode, we withhold its three matches and fit each method to the 760 remaining matches. We then compare the average predicted win ratio of each team with their observed fraction of wins, and average across the teams to obtain one MSE per method and replication, as illustrated in the box plots in Figure 4 which shows that the proposed dBT approach has lower mean win-ratio MSE than both comparators, with lower paired error in 75 of 100 collections against static BT and 79 against kernel BT. Log loss and Brier scores show a similar pattern across methods. In Appendix D.3.1 we show fitted abilities, win proportions, and ranks to further support our analysis.

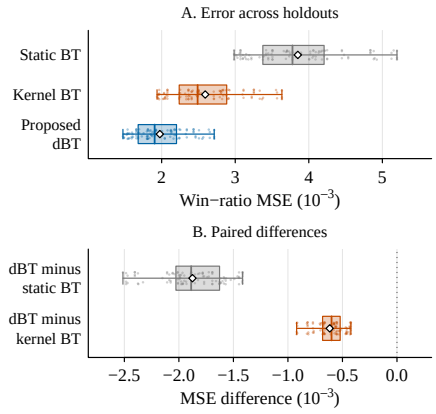

Figure 3: Arena 140K data analysis under low local exposure. A: Opponent-matched average win-ratio MSE. B: Paired differences, dBT minus comparator.

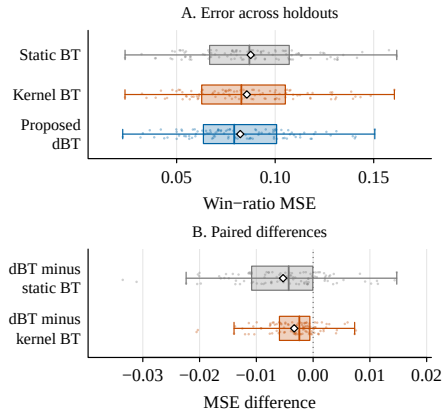

Figure 4: T20 cricket data analysis under low local exposure. A: Opponent-matched win-ratio MSE. B: Paired differences, dBT minus comparator.

6 Discussion

We propose a dynamic Bradley–Terry model that represents time-varying ability functions with basis functions and estimates their coefficients jointly using penalized logistic likelihood. This flexible formulation accommodates different temporal structures through the choice of basis and penalty, and extends to a varying-coefficient model with time-varying, item-specific covariate effects. Theoretically, we establish identification and deterministic fitting criteria, together with random-design asymptotic bounds for the penalized estimator which include approximation error and distinguish integrated ability recovery from prediction under the comparison design. Simulations and applications to the Chatbot Arena and T20 cricket datasets illustrate the advantages of this approach under temporary item absences and sparse comparisons where methods relying on local smoothing may struggle. Future work includes establishing guarantees for data-driven selection of the basis and penalty, extending the rate analysis to dependent comparison schedules, and inference for dynamic pairwise contrasts with limited comparison information.

Reproducibility Statement

We specify the evaluation targets, temporal splits, tuning grids, and treatment of fitting failures in the experimental descriptions. Our figures and tables are generated using reproducible results and versioned configurations, and all random seeds and fitting diagnostics for the simulation studies are retained. The accompanying files document the data versions and preprocessing steps for each application. Complete proofs of our theoretical results appear in the Appendix.

AI Use Statement

The research ideas, problem formulation, and original contributions are our own. We used generative AI tools to assist with code organization and implementation, simulation and data-analysis workflows, theoretical derivations, and manuscript editing. We reviewed and revised all AI-assisted material and take full responsibility for the originality and accuracy of the work, including all mathematical claims, code, analyses, and text.

References

Adelin Albert and John A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984. doi: 10.1093/biomet/71.1.1.

Kenji Araki, Yoshihiro Hirose, and Fumiyasu Komaki. Paired comparison models with age effects modeled as piecewise quadratic splines. *International Journal of Forecasting*, 35(2):733–740, 2019. doi: 10.1016/j.ijforecast.2018.02.006.

Heejong Bong, Wanshan Li, Shamindra Shrotriya, and Alessandro Rinaldo. Nonparametric estimation in the dynamic bradley-terry model. In *International Conference on Artificial Intelligence and Statistics*, pp. 3317–3326. PMLR, 2020.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Manuela Cattelan, Cristiano Varin, and David Firth. Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 62(1):135–150, 2013.

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204–2235, 2019.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024.

Regina Dittrich, Reinhold Hatzinger, and Walter Katzenbeisser. Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4):511–525, 1998.

Jianqing Fan and Wenyang Zhang. Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1(1):179–195, 2008. doi: 10.4310/SII.2008.v1.n1.a15.

Jianqing Fan, Jikai Hou, and Mengxin Yu. Uncertainty quantification of mle for entity ranking with covariates. *Journal of Machine Learning Research*, 25(358):1–83, 2024.

Stephen E. Fienberg and Kinley Larntz. Log linear representation for paired and multiple comparisons models. *Biometrika*, 63(2):245–254, 1976. doi: 10.1093/biomet/63.2.245.

Lester R. Ford, Jr. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8P2):28–33, 1957.

Chao Gao, Yandi Shen, and Anderson Y Zhang. Uncertainty quantification in the bradley–terry–luce model. *Information and Inference: A Journal of the IMA*, 12(2):1073–1140, 2023.

Ruijian Han, Rougang Ye, Chunxi Tan, and Kani Chen. Asymptotic theory of sparse bradley–terry model. *The Annals of Applied Probability*, 30(5):pp. 2491–2515, 2020. ISSN 10505164, 21688737. URL https://www.jstor.org/stable/26966014.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–796, 1993.

Julien Hendrickx, Alex Olshevsky, and Venkatesh Saligrama. Minimax rate for learning from pairwise comparisons in the btl model. In *International Conference on Machine Learning*, pp. 4193–4202. PMLR, 2020.

J. Z. Huang and Y. Su. Asymptotic properties of penalized spline estimators in concave extended linear models: Rates of convergence. *The Annals of Statistics*, 49(6):3383–3407, 2021. doi: 10.1214/21-AOS2088.

Eglantine Karlé and Hemant Tyagi. Dynamic ranking with the btl model: a nearest neighbor based rank centrality method. *Journal of Machine Learning Research*, 24(269):1–57, 2023.

Blaž Krese and Erik Štrumbelj. A bayesian approach to time-varying latent strengths in pairwise comparisons. *PLOS ONE*, 16(5):e0251945, 2021. doi: 10.1371/journal.pone.0251945.

Xiudi Li and Sijia Li. Efficient inference for covariate-adjusted bradley-terry model with covariate shift. *arXiv preprint arXiv:2503.18256*, 2025.

Lucas Maystre, Victor Kristof, and Matthias Grossglauser. Pairwise comparisons with flexible time-dynamics. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1236–1246. Association for Computing Machinery, 2019. doi: 10.1145/3292500.3330831.

Finbarr O'Sullivan, Brian S. Yandell, and William J. Raynor, Jr. Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, 81(393): 96–103, 1986.

Stephen Rushe. Cricsheet: Available match data downloads and json format. https://cricsheet.org/downloads/, 2026. Men's international T20 JSON archive, accessed September 4, 2026.

Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

Mervyn J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(3):310–313, 1981. doi: 10.1111/j.2517-6161.1981.tb01676.x.

Gordon Simons and Yi-Ching Yao. Asymptotics when the number of parameters tends to infinity in the bradley-terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.

Xinyu Tian, Jian Shi, Xiaotong Shen, and Kai Song. A spectral approach for the dynamic bradley–terry model. *Stat*, 13(3):e722, 2024.

Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton, 2 edition, 2017.

Weichen Wu, Brian W Junker, and Nynke Niezink. Asymptotic comparison of identifying constraints for bradley-terry models. *arXiv preprint arXiv:2205.04341*, 2022.

Ting Yan, Yaning Yang, and Jinfeng Xu. Sparse paired comparisons in the bradley-terry model. *Statistica Sinica*, pp. 1305–1318, 2012.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.