Journal of Machine Learning Research 19 (2018) 1-12 Submitted 3/17; Revised 8/18; Published 11/18

Supplementary Material for: A New and Flexible Approach to the Analysis of Paired Comparison Data

Ivo F. D. Oliveira IVODAVID@GMAIL.COM
Department of Science, Engineering and Technology
UFVJM - Federal University of the Valleys of Jequitinhonha and Mucuri
Teofilo Otoni, Minas Gerais, Brazil

Nir Ailon NAILON@CS.TECHNION.AC.IL
Department of Computer Science
Technion - Israel Institute of Technionlogy
Haifa, Israel

Ori Davidov DAVIDOV@STAT.HAIFA.AC.IL
Department of Statistics
University of Haifa
Haifa, Israel

Editor: To be defined November 6, 2018

1. Simulations

In this file we reproduce experiments 1,2 and 3 and the illustrative example of Sections 5.1 and 5.2 utilizing the $\mathcal{L}_1$ and the $\mathcal{L}_\infty$ norms. For completeness we provide a full description of the experiments and follow up with a discussion comparing the main results for different norms. The simulations are performed 1000 times and we report the average performance under the specified conditions.

For convenience we provide a selection of results that serve as theoretical benchmarks for comparison between norms and which will help us to calibrate our expectations. The proofs are omitted for brevity.

Proposition 1 *Let* $\boldsymbol{P}^* = F_{\hat{\boldsymbol{\beta}}}(\Delta\hat{\boldsymbol{\mu}})$ *where* $\hat{\boldsymbol{\beta}}$ *and* $\hat{\boldsymbol{\mu}}$ *are estimated with* PolyRank *using either the* $\mathcal{L}_1$ *or the* $\mathcal{L}_\infty$ *norm. Assume that the true probability matrix is given by* $\boldsymbol{P} = F(\Delta\boldsymbol{\mu})$ *for some* $\boldsymbol{\mu}$ *and some unknown* L*-Lipschitz continuous function* F *with an analytic inverse function* F^{-1} *whose coefficients are upper-bounded by* U*. Then, if* $\boldsymbol{\beta}$ *and* $\boldsymbol{\mu}$ *were estimated using the* $\mathcal{L}_1$ *norm, then:*

$$||\boldsymbol{P}^* - \boldsymbol{P}||_1 \leq (1 + 4LU)||\hat{\boldsymbol{P}} - \boldsymbol{P}||_1 + (1/2^D)LUI^2 \tag{1}$$

and if $\boldsymbol{\beta}$ *and* $\boldsymbol{\mu}$ *were estimated using the* $\mathcal{L}_\infty$ *norm, then:*

$$||\boldsymbol{P}^* - \boldsymbol{P}||_\infty \leq (1 + 4LU)||\hat{\boldsymbol{P}} - \boldsymbol{P}||_\infty + (1/2^D)LU. \tag{2}$$

Corollary 2 *Under the conditions of the above Proposition, if* $\boldsymbol{\beta}$ *and* $\boldsymbol{\mu}$ *were estimated using the* $\mathcal{L}_1$ *norm, then:*

$$||\boldsymbol{P}^* - \boldsymbol{P}||_2 \leq (1 + 4LU)I||\hat{\boldsymbol{P}} - \boldsymbol{P}||_2 + (1/2^D)LUI^2 \tag{3}$$

if $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ were estimated using the $\mathcal{L}_\infty$ norm, then:

$$||\boldsymbol{P}^* - \boldsymbol{P}||_2 \leq (1 + 4LU)I||\hat{\boldsymbol{P}} - \boldsymbol{P}||_2 + (1/2^D)LUI. \tag{4}$$

Both Proposition 1 and Corollary 2 provide sensitivity bounds analogous to those obtained for the $\mathcal{L}_2$ norm in Theorem 6 of the main text. In Proposition 1 bounds are given with respect to the corresponding norms utilized by PolyRank, whereas Corollary 2 gives the same bounds with with respect to the $\mathcal{L}_2$ norm. The $\mathcal{L}_2$ bounds provide a reference point to which other simulation results are compared. Notice that equations (3) and (4) of Corollary 2 show that the constant associated with the estimation error, when measured in terms of $\mathcal{L}_2$, is $(1 + 4LU)I$ for both the $\mathcal{L}_1$ and the $\mathcal{L}_\infty$ norms, i.e., they are I times bigger than the constant associated with the $\mathcal{L}_2$ norm. Also, the approximation error associated with the $\mathcal{L}_1$ norm, given by $(1/2^D)LUI^2$, is I times bigger than that of the $\mathcal{L}_2$ and the $\mathcal{L}_\infty$ norms. Thus, we expect to see higher estimation errors than those found with the $\mathcal{L}_2$ norm and thus a slower convergence. Nevertheless, for sufficiently large D and provided sufficient data, both norms are guaranteed to converge even in this agnostic setting.

We proceed to the experiments.

Experiment 1: In this experiment we investigate the empirical performance of the LS refinement (given the correct comparison function) with PolyRank using a low degree polynomial. For this we generate $I = 20$ items with merits μ_i sampled uniformly from $[0, 10]$. A total of 50 pairs, selected randomly, were compared assuming a Bradley-Terry-Luce (BTL) model. We refine the estimator $\hat{p}_{ij} = (Y_{ij}+1)/(m_{ij}+2)$ with PolyRank using $D = 5$. We also compute the LS estimator with known F. Figures 1, 2 show the average distance $||\boldsymbol{P}^* - \boldsymbol{P}||_2$ of the estimator refined with the $\mathcal{L}_1$ and $\mathcal{L}_\infty$ norms respectively.

As expected, in both cases the LS method with the correct comparison function performs best, PolyRank performs almost as well and both substantially outperform the initial estimates. In this experiment, no significant difference is noticed when PolyRank is used with either the $\mathcal{L}_1, L_2$ or the $\mathcal{L}_\infty$ norms.

Experiment 2 In this experiment we investigate the empirical performance of PolyRank in the round-robin setting with an increasing number of items. For this we generate a sequence of round-robin tournaments with $I = 10, 20, 30, 40$ and 50 items. The data is generated with a polynomial quantile function of degree $D = 5$. The matrix $\boldsymbol{P}$ is estimated using the isotonic regression estimator of (Chatterjee, S. and Mukherjee, S. 2016) and refined using PolyRank with $D \in \{3, 5, 7\}$ with the $\mathcal{L}_1$ and $\mathcal{L}_\infty$ norms. Figures 3 and 4 display, respectively, the average of $||\boldsymbol{P}^* - \boldsymbol{P}||^2/I^2$ when the $\mathcal{L}_1$ and $\mathcal{L}_\infty$ norms are used in the refinement step. Figures 5 and 6 display the average of $||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \hat{\boldsymbol{\mu}} - \boldsymbol{\mu})||^2/(I + D)$ for the respective norms.

Figures 3 and 4 show four curves, all of which decrease with I. The solid blue curve is the risk for the unrefined isotonic-regression based estimator. The estimators refined by PolyRank correspond to the remaining curves. When the $\mathcal{L}_1$ norm is used we find that the refined estimators always do better than the unrefined estimators. When the $\mathcal{L}_\infty$ norm is used this is not the case. Even though the $\mathcal{L}_\infty$ norm does not perform as well on all cases, precision is lost at most by a constant factor in the worst case, and, in the best case we find an improvement in the overall estimator. This is consistent with our theoretical results.

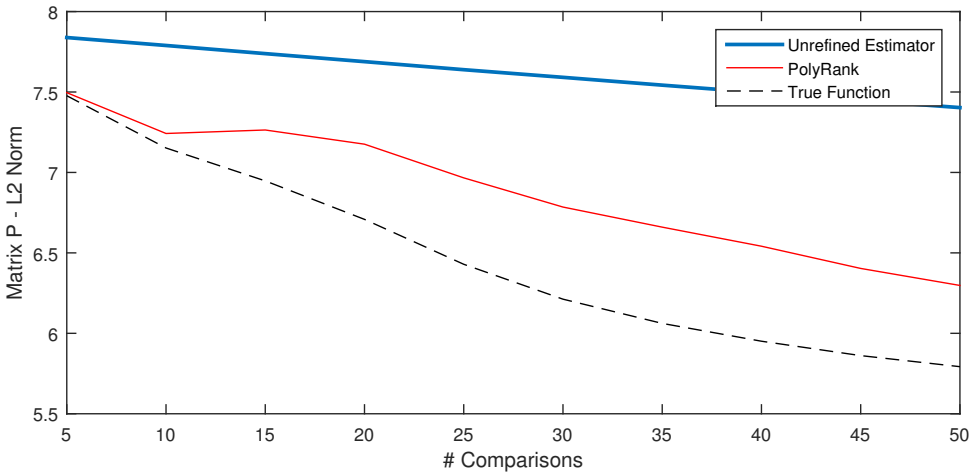

Figure 1: Comparison of estimator refined using PolyRank with the $\mathcal{L}_1$ norm with low sampling.

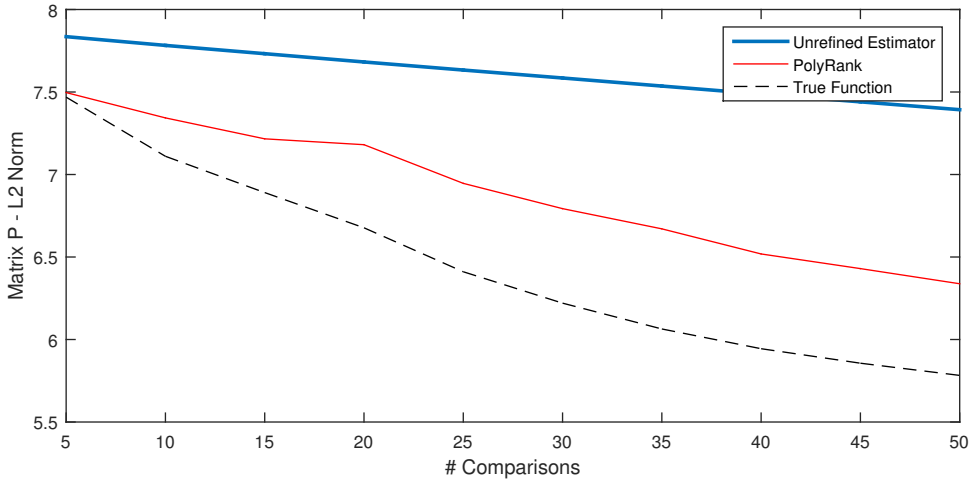

Figure 2: Comparison of estimator refined using PolyRank with the $\mathcal{L}_\infty$ norm with low sampling.

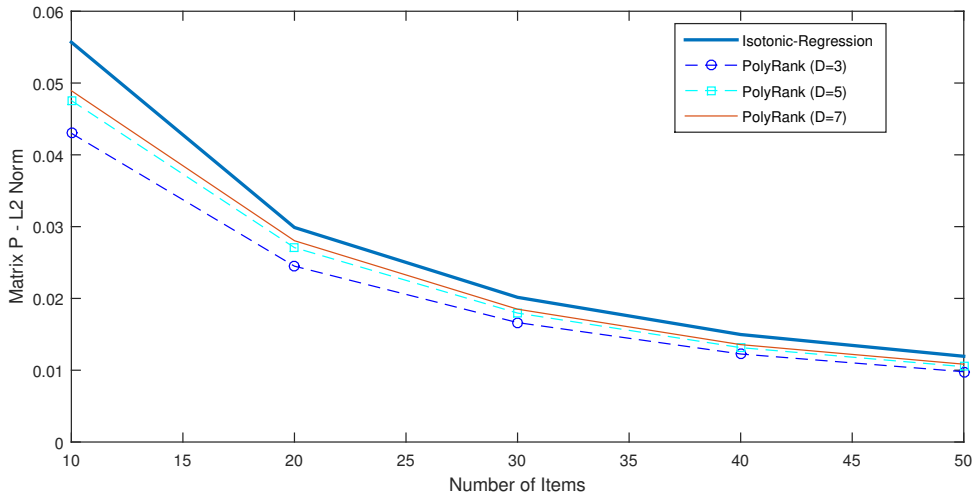

Figure 3: Refined estimators for round-robin tournaments using the $\mathcal{L}_1$ norm.

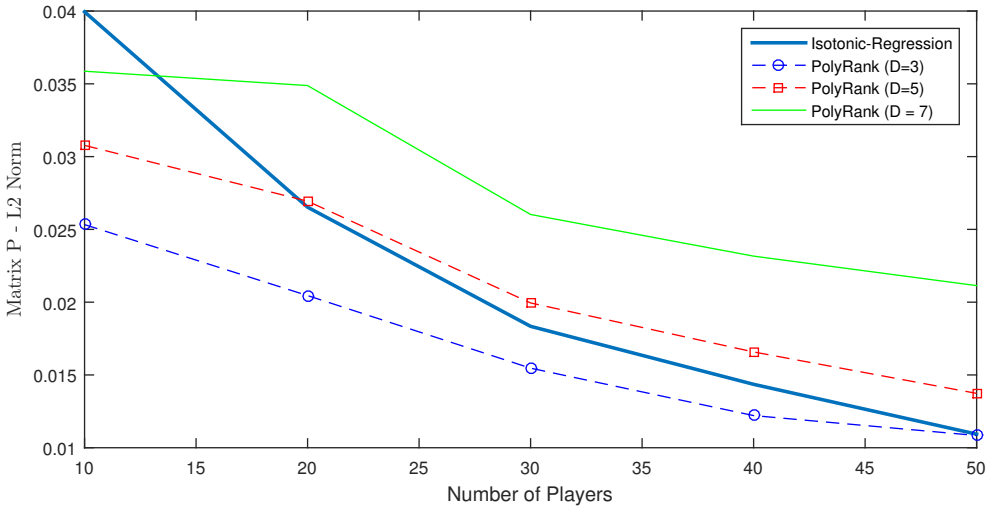

Figure 4: Refined estimators for round-robin tournaments using the $\mathcal{L}_\infty$ norm.

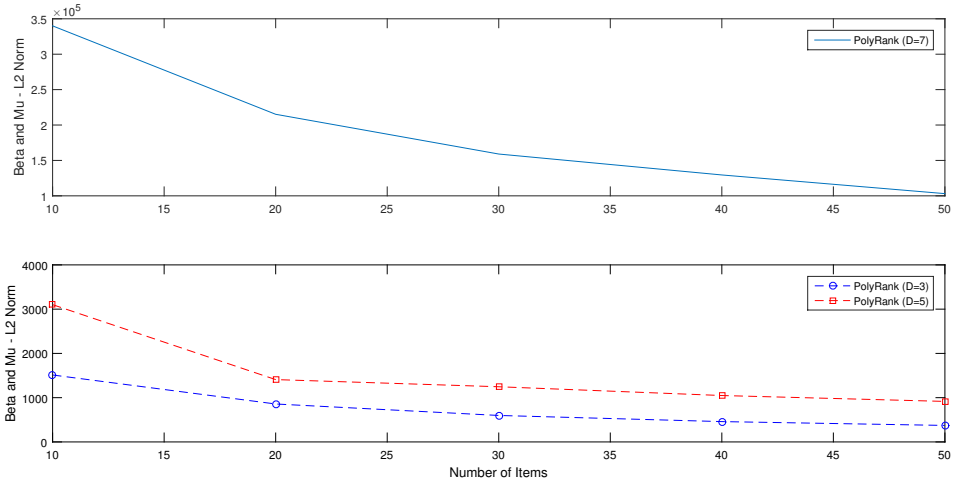

Figure 5: Estimated parameters for round-robin tournaments using the $\mathcal{L}_1$ norm.

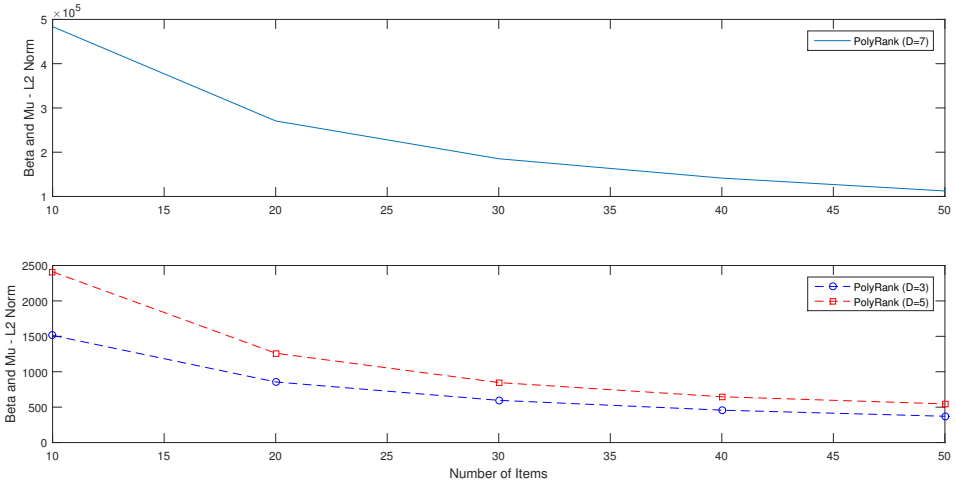

Figure 6: Estimated parameters for round-robin tournaments using the $\mathcal{L}_\infty$ norm.

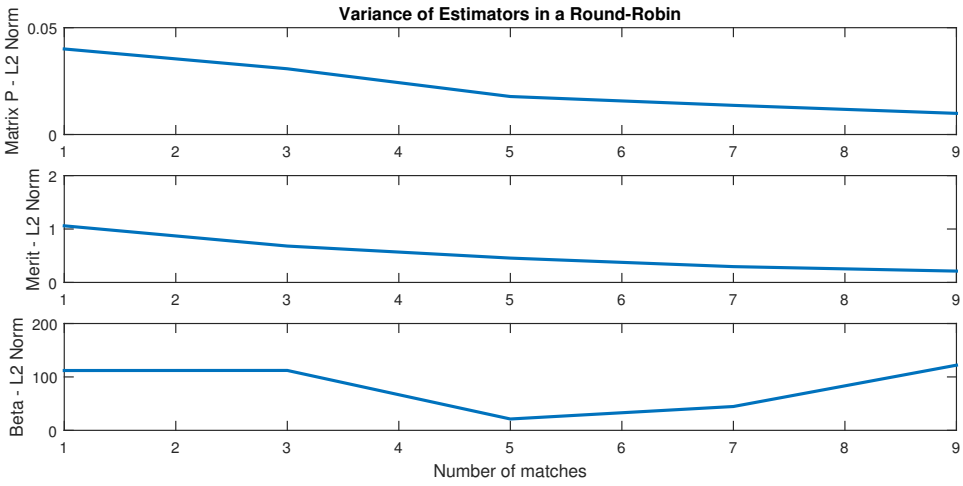

Figure 7: Variance of estimators in a round-robin with increasing number of matches between each pair using the $\mathcal{L}_1$ norm.

Notice that in both experiments over-fitting, i.e., $D = 7$, usually results in higher estimation errors when compared to $D = 3, 5$ and the lowest estimation error is found when $D = 3$, i.e., when under-fitting. This behavior is similar to what was found with the $\mathcal{L}_2$ norm. In Figures 5 and 6 we see that the average error of the estimated parameters decreases as a function of I and once again under-fitting seems to outperform the true model, and the true model outperforms overfitting.

In this experiment, no significant difference is noticed when PolyRank is used with either the $\mathcal{L}_1, \mathcal{L}_2$ or the $\mathcal{L}_\infty$ norms other than the overall speed of convergence of the estimators. When speed of convergence is critical, then, the $\mathcal{L}_2$ norms should be prefered.

Experiment 3: In this experiment we investigate the empirical performance of PolyRank in the round-robin setting with a fixed number of items. For this we generate a sequence of round-robin tournaments with $I = 10$ and an increasing number of matches. The data is generated using a polynomial quantile function of degree $D = 3$. The matrix $\boldsymbol{P}$ is estimated using the standard frequency estimator for $\hat{p}_{ij}$ and is refined using PolyRank (with $D = 3$). Figures 7 and 8 display the average of $||\boldsymbol{P}^* - \boldsymbol{P}||^2/I^2$, of $||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2/I$ and of $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2$ for $m_{ij} = 1, 3, 5, 7$ and 9 for all pairs (i, j) when the $\mathcal{L}_1$ and the $\mathcal{L}_\infty$ norms are used in the refinement step. Figures 9 and 10 show the sequence of estimated functions when the $\mathcal{L}_1$ and the $\mathcal{L}_\infty$ norms are used in the refinement step.

For both norms we observe a similar behavior. The top two curves of Figures 7 and 8 show that the variance of the estimators $\boldsymbol{P}^*$ and $\hat{\boldsymbol{\mu}}$ decreases with the amount of paired comparisons in both experiments. The bottom curves of Figures 7 and 8 show the variance of the estimator $\hat{\boldsymbol{\beta}}$, and, in both cases the estimator did not converge. Figures 9 and 10 show that the graph of the estimated comparison functions do not match the true function due to the slow convergence.

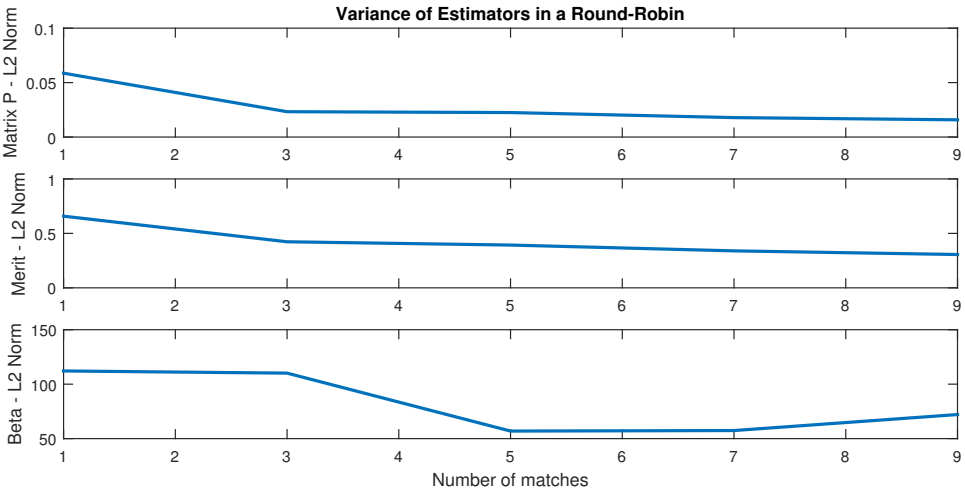

Figure 8: Variance of estimators in a round-robin with increasing number of matches between each pair using the $\mathcal{L}_\infty$ norm.

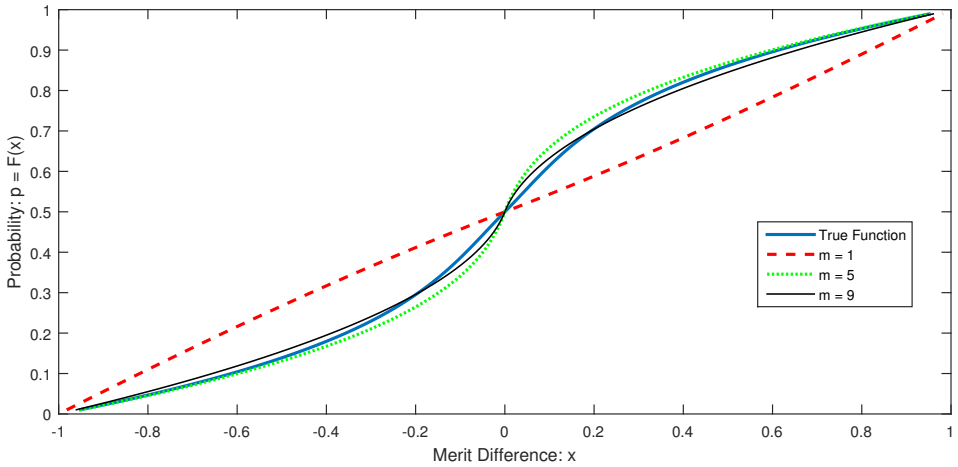

Figure 9: Recovered function using the $\mathcal{L}_1$ norm.

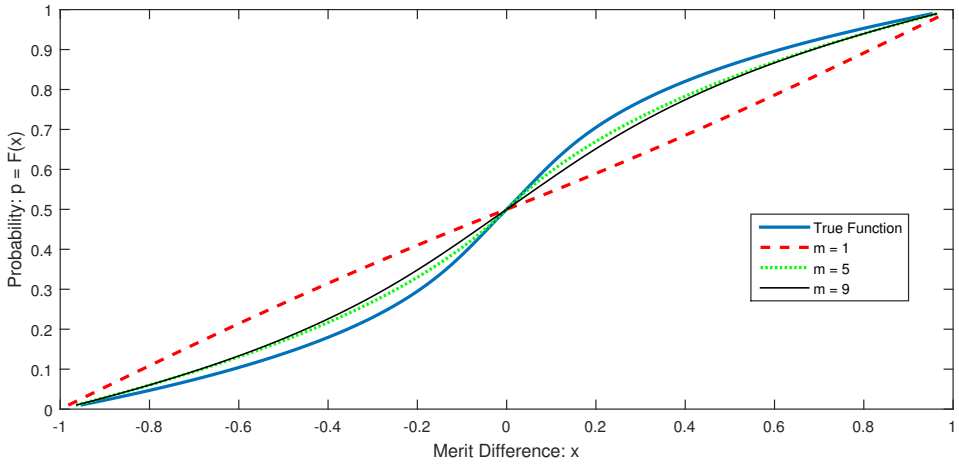

Figure 10: Recovered function using the $\mathcal{L}_\infty$ norm.

1.1 Illustrative Example

In this subsection we illustrate the use of PolyRank on a computer-chess dataset[1]. The dataset comprises of matches between 186 free single-CPU chess-engines. Each chess-engine played (roughly) 32 matches against 40 opponents. We use PolyRank to estimate model parameters from observed matches that resulted in a victory or defeat; ties are ignored. Figures 11 and 12 show the estimated comparison function for various values of D for the first 100 chess-engines using PolyRank with the $\mathcal{L}_1$ and the $\mathcal{L}_\infty$ norms.

In this experiment the comparison function found by PolyRank utilizing both the $\mathcal{L}_1$ and the $\mathcal{L}_2$ norms are very similar, whereas the $\mathcal{L}_\infty$ norm differs considerably. For both norms, the function considered seems to stabilize for D greater than or equal to 7. Figures 13 and 14 show the estimated function when the dimension $D = 7$ is fixed and the number of chess-engines I is gradually increased for the $\mathcal{L}_1$ and the $\mathcal{L}_\infty$ refined estimators. For I greater than or equal to 60 the estimated functions also seem to stabilize. Finally, Figures 15 and 16 compare the best fit functions recovered by PolyRank to the family of BTL models described by $F_{\text{BTL}}(x) = 1/(1 + \exp(-\kappa x))$ for various values of $\kappa > 0$. Again, the $\mathcal{L}_1$ norm seems to recover the same function as the $\mathcal{L}_2$ norm, which, as observed in the main paper considerably differs from the BTL models. The $\mathcal{L}_\infty$ norm, however, seem to not have converged, perhaps because in these experiments, the expected value of $||\hat{\boldsymbol{P}} - \boldsymbol{P}||_\infty$ grows with the increasing number of items when the number of matches remains fixed.

1.2 Concluding Remarks

We observe that the $\mathcal{L}_\infty$ norm seems to converge slower than the $\mathcal{L}_1$ and $\mathcal{L}_2$ norms. At times the $\mathcal{L}_\infty$ even performs worse than the unrefined estimators. The $\mathcal{L}_2$ norm seems to slightly outperform the $\mathcal{L}_1$ norm in most experiments, though from a practitioner's perspective, the both the $\mathcal{L}_2$ and the the $\mathcal{L}_1$ norms can be chosen. As a rule of thumb, if performance

1. Publicly available at `http://kirill-kryukov.com/chess/kcec/games.html`.

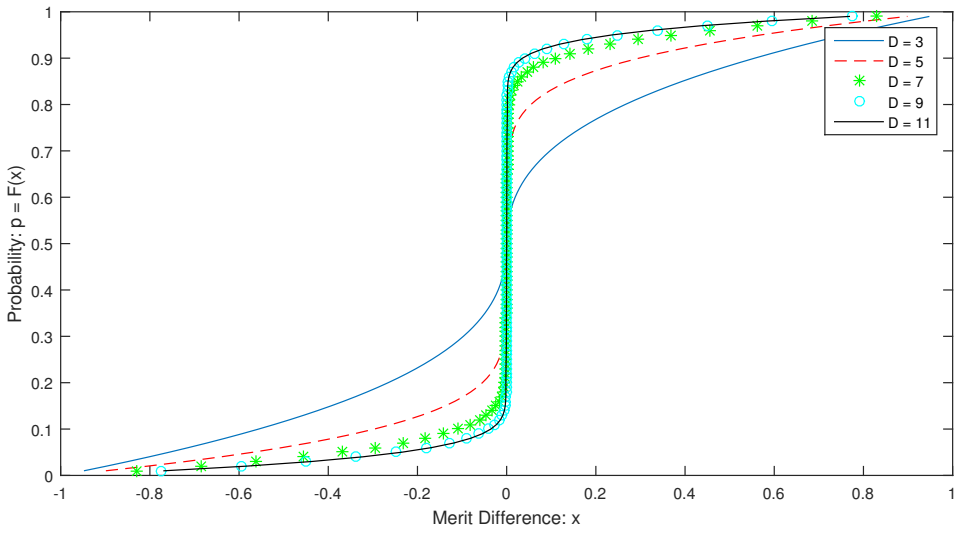

Figure 11: The effect of increasing the dimension D in estimating function F obtained with the $\mathcal{L}_1$ norm.

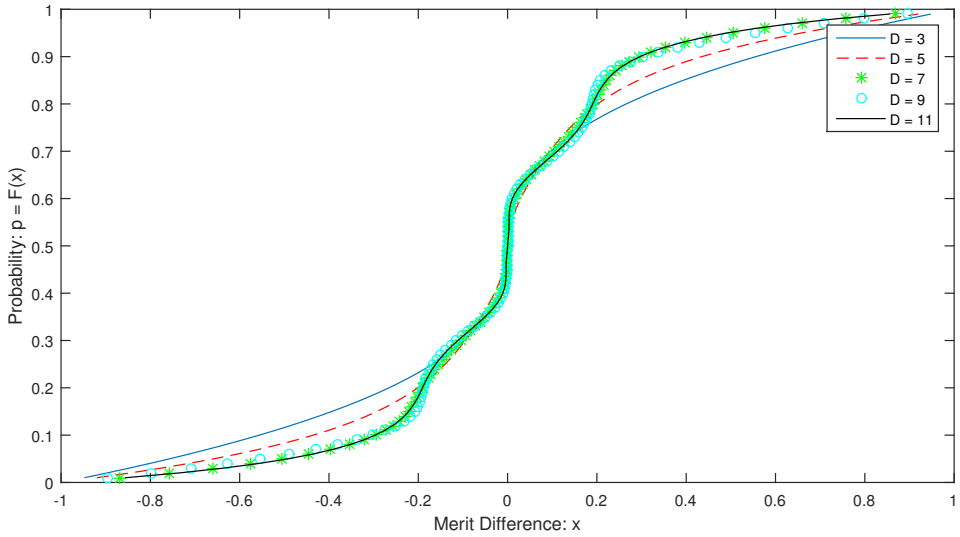

Figure 12: The effect of increasing the dimension D in estimating function F obtained with the $\mathcal{L}_\infty$ norm.

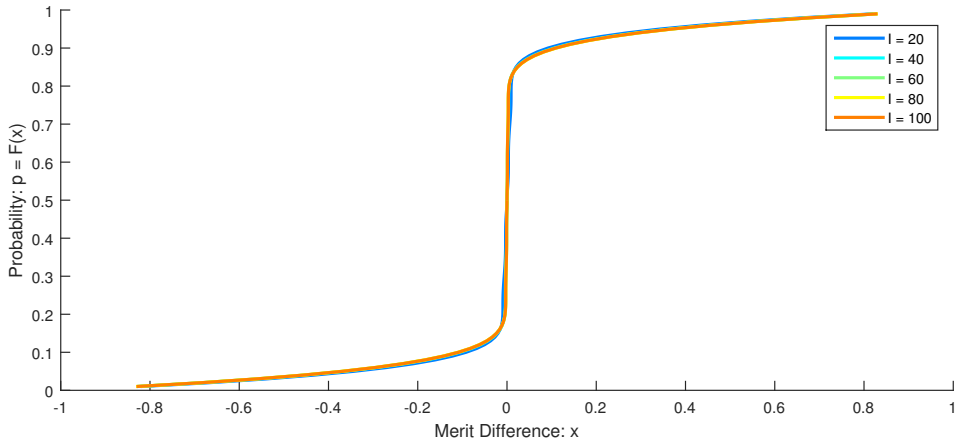

Figure 13: The effect of increasing the amount of data in estimating function F obtained with the $\mathcal{L}_1$ norm.

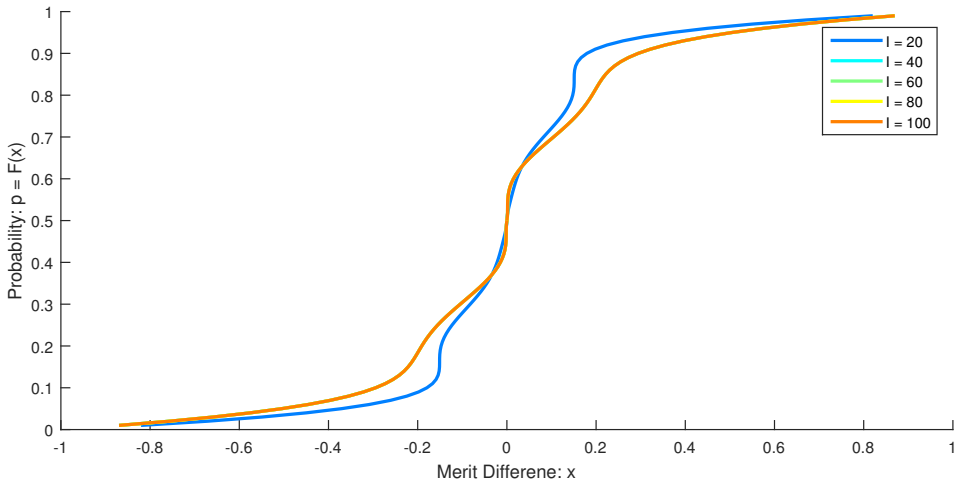

Figure 14: The effect of increasing the amount of data in estimating function F obtained with the $\mathcal{L}_\infty$ norm.

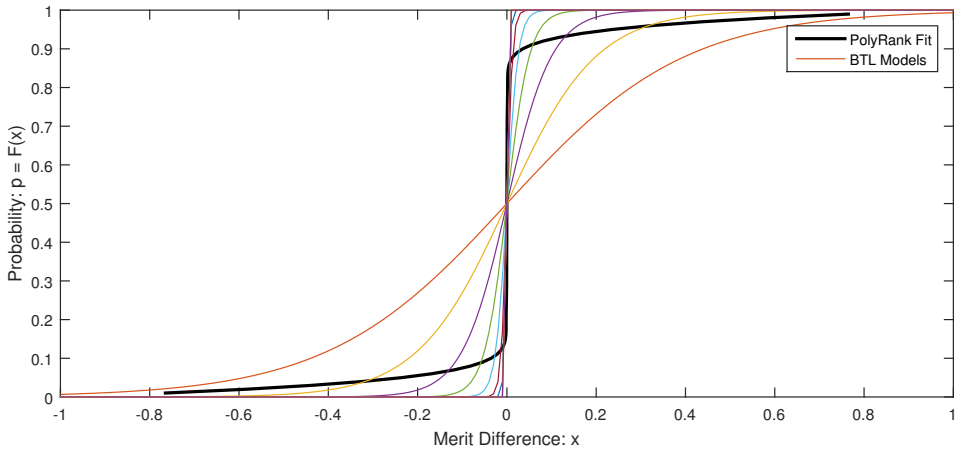

Figure 15: Bradley-Terry-Luce models compared to the best fit comparison function obtained with the $\mathcal{L}_1$ norm.

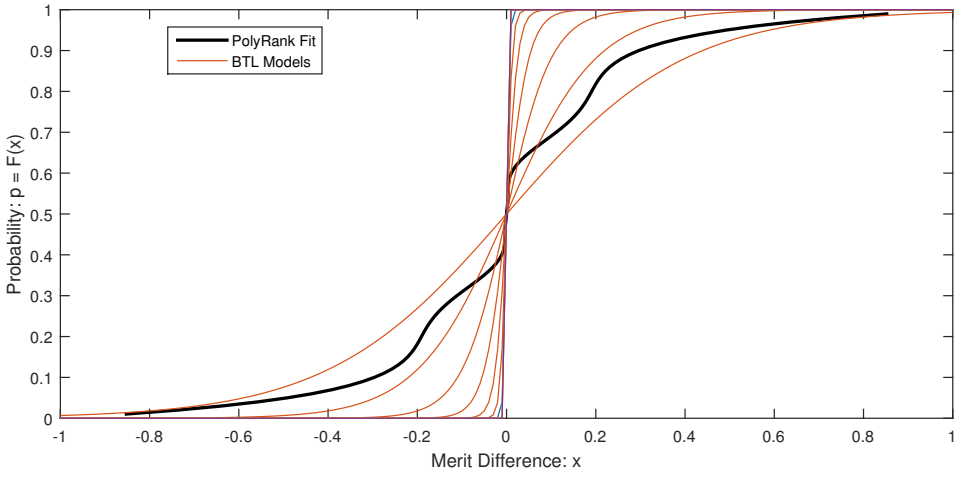

Figure 16: Bradley-Terry-Luce models compared to the best fit comparison function obtained with the $\mathcal{L}_\infty$ norm.

is measured in terms of variance of the estimators, then, we recommend the $\mathcal{L}_2$ norm minimization.