Journal of Artificial Intelligence Research 70 (2021) 1481-1515 Submitted 12/2020; published 04/2021

Loss Functions, Axioms, and Peer Review

Ritesh Noothigattu RITESHN@CMU.EDU
Machine Learning Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Nihar B. Shah NIHARS@CS.CMU.EDU
Machine Learning Department and Computer Science Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Ariel D. Procaccia ARIELPRO@SEAS.HARVARD.EDU
School of Engineering and Applied Sciences
Harvard University
150 Western Avenue
Boston, MA 02134

Abstract

It is common to see a handful of reviewers reject a highly novel paper, because they view, say, extensive experiments as far more important than novelty, whereas the community as a whole would have embraced the paper. More generally, the disparate mapping of criteria scores to final recommendations by different reviewers is a major source of inconsistency in peer review. In this paper we present a framework inspired by empirical risk minimization (ERM) for learning the community's aggregate mapping. The key challenge that arises is the specification of a loss function for ERM. We consider the class of $L(p, q)$ loss functions, which is a matrix-extension of the standard class of L_p losses on vectors; here the choice of the loss function amounts to choosing the hyperparameters $p, q \in [1, \infty]$. To deal with the absence of ground truth in our problem, we instead draw on computational social choice to identify desirable values of the hyperparameters p and q. Specifically, we characterize $p = q = 1$ as the only choice of these hyperparameters that satisfies three natural axiomatic properties. Finally, we implement and apply our approach to reviews from IJCAI 2017.

1. Introduction

The essence of science is the search for objective truth, yet scientific work is typically evaluated through peer review — a notoriously subjective process (Church, 2005; Lamont, 2009; Bakanic et al., 1987; Hojat et al., 2003; Mahoney, 1977; Kerr et al., 1977). One prominent source of subjectivity is the disparity across reviewers in terms of their emphasis on the various criteria used for the overall evaluation of a paper. Lee (2015) refers to this disparity as *commensuration bias*, and describes it as follows:

> *"In peer review, reviewers, editors, and grant program officers must make interpretive decisions about how to weight the relative importance of qualitatively different peer review criteria — such as novelty, significance, and methodological*

> *soundness — in their assessments of a submission's final/overall value. Not all peer review criteria get equal weight; further, weightings can vary across reviewers and contexts even when reviewers are given identical instructions."*

Lee (2015) further argues that commensuration bias "illuminates how intellectual priorities in individual peer review judgments can collectively subvert the attainment of community-wide goals" and that it "permits and facilitates problematic patterns of publication and funding in science." There have been, however, very few attempts to address this problem.

A fascinating exception, which serves as a case in point, is the 27th AAAI Conference on Artificial Intelligence (AAAI 2013). Reviewers were asked to score papers, on a scale of 1–6, according to the following criteria: technical quality, experimental analysis, formal analysis, clarity/presentation, novelty of the question, novelty of the solution, breadth of interest, and potential impact. The admirable goal of the program chairs was to select "exciting but imperfect papers" over "safe but solid" papers, and, to this end, they provided detailed instructions on how to map the foregoing criteria to an overall recommendation. For example, the preimage of 'strong accept' is "a 5 or 6 in some category, no 1 in any category," that is, reviewers were instructed to strongly accept a paper that has a 5 or 6 in, say, clarity, but is below average according to each and every other criterion (i.e., a clearly boring paper). It turns out that the handcrafted mapping did not work well, and many of the reviewers chose to not follow these instructions. Indeed, handcrafting such a mapping requires specifying an 8-dimensional function, which is quite a non-trivial task.[1] Consequently, in this paper we do away with a manual handcrafting approach to this problem.

Instead, we propose a data-driven approach based on ideas from machine learning, designed to learn a mapping from criteria scores to recommendations capturing the opinion of the entire (reviewer) community. From a machine learning perspective, the examples are reviews, each consisting of criteria scores (the input point) and an overall recommendation (the label). We make the assumption that each reviewer has a *monotonic* mapping in mind, in the sense that a paper whose scores are at least as high as those of another paper on every criterion would receive an overall recommendation that is at least as high; the reviews submitted by a particular reviewer can be seen as observations of that mapping. Given this data, our goal is to learn a *single monotonic mapping* that minimizes a loss function (which we discuss momentarily). We can then apply this mapping to the criteria scores associated with each review to obtain new overall recommendations (which can either replace the original ones or can be provided alongside the original ones as additional information for the program chairs).

Our approach to learn this mapping is inspired by empirical risk minimization (ERM). In more detail, for some loss function, our approach is to find a mapping that, among all monotonic mappings from criteria scores to the overall scores, minimizes the loss between its outputs and the overall scores given by reviewers across all reviews. However, the choice of loss function may significantly affect the final outcome, so that choice is a key issue.

1. See also Mogul and Anderson (2008) for a similar case in the peer-review process of the OSDI conference. OSDI 2006 did not allow reviewers to report an overall score, but instead, the PC co-chairs synthesized this score from a weighted combination of criteria scores. Here, we instead take a community-based approach and learn a mapping common to the community of reviewers. Further, we do not assume linear aggregation of the criteria scores, but allow a more general monotonic mapping.

Specifically, we focus on the family of $L(p, q)$ loss functions, with hyperparameters $p, q \in [1, \infty]$, which is a matrix-extension of the more popular family of L_p losses on vectors. Our question, then, is:

> What values of the hyperparameters $p \in [1, \infty]$ and $q \in [1, \infty]$ in the specification of the $L(p, q)$ loss function should be used?

A challenge we must address is the absence of any ground truth in peer review. To this end, we take the perspective of *computational social choice* (Brandt et al., 2016), since our framework aggregates individual opinions over mappings into a consensus mapping. From this viewpoint, it is natural to select the loss function so that the resulting aggregation method satisfies socially desirable properties, such as *consensus* (if all reviewers agree then the aggregate mapping should coincide with their recommendations), *efficiency* (if one paper dominates another then its overall recommendation should be at least as high), and *strategyproofness* (reviewers cannot pull the aggregate mapping closer to their own recommendations by misreporting them).

With this background, the main contributions of this paper are as follows. We first provide a principled framework for addressing the issue of subjectivity regarding the various criteria in peer review.

Our main theoretical result is a characterization theorem that gives a decisive answer to the question of choosing the loss function for ERM: the three aforementioned properties are satisfied *if and only if* the hyperparameters are set as $p = q = 1$. This result singles out an instantiation of our approach that we view as particularly attractive and well grounded.

We also provide empirical results, which analyze properties of our approach when applied to a dataset of 9197 reviews from IJCAI 2017. One vignette is that the papers selected by $L(1, 1)$ aggregation have a 79.2% overlap with the actual list of accepted papers, suggesting that our approach makes a significant difference compared to the status quo (arguably for the better).

Finally, we note that the approach taken in this paper may find other applications. Indeed, the problem of selecting a loss function is ubiquitous in machine learning (Rosasco et al., 2004; Masnadi-Shirazi and Vasconcelos, 2008; Mei et al., 2018), and the axiomatic approach provides a novel way of addressing it. Going beyond loss functions, machine learning researchers frequently face the difficulty of picking an appropriate hypothesis class or values for certain hyperparameters.[2] Thus, in problem settings where such choices must be made — particularly in emerging applications of machine learning (such as peer review) — the use of natural axioms can help guide these choices.

2. Our Framework

Suppose there are n reviewers $\mathcal{R} = \{1, 2, \ldots, n\}$, and a set $\mathcal{P}$ of m papers, denoted using letters such as a, b, c. Each reviewer i reviews a subset of papers, denoted by $P(i) \subseteq \mathcal{P}$. Conversely, let $R(a)$ denote the set of all reviewers who review paper a. Each reviewer assigns scores to each of their papers on d different criteria, such as novelty, experimental analysis, and technical quality, and also gives an overall recommendation. We denote the

2. Popular techniques such as cross-validation for choosing hyperparameters also in turn depend on specification of a loss function.

criteria scores given by reviewer i to paper a by $\mathbf{x}_{ia}$, and the corresponding *overall recommendation* by y_{ia}. Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d$ denote the domains of the d criteria scores, and let $\mathbb{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$. Also, let $\mathbb{Y}$ denote the domain of the overall recommendations. For concreteness, we assume that each $\mathcal{X}_k$ as well as $\mathbb{Y}$ is the real line. However, our results hold more generally, even if these domains are non-singleton intervals in $\mathbb{R}$, for instance.

We further assume that each reviewer has a monotonic function in mind that they use to compute the overall recommendation for a paper from its criteria scores. By a monotonic function, we mean that given any two score vectors $\mathbf{x}$ and $\mathbf{x}'$, if $\mathbf{x}$ is greater than or equal to $\mathbf{x}'$ on all coordinates, then the function's value on $\mathbf{x}$ must be at least as high as its value on $\mathbf{x}'$. Formally, for each reviewer i, there exists $g_i^\star \in \mathcal{F}$ such that $y_{ia} = g_i^\star(\mathbf{x}_{ia})$ for all $a \in P(i)$, where

$$\mathcal{F} = \{f : \mathbb{X} \to \mathbb{Y} \mid \forall \mathbf{x}, \mathbf{x}' \in \mathbb{X}, \mathbf{x} \geq \mathbf{x}' \Rightarrow f(\mathbf{x}) \geq f(\mathbf{x}')\}$$

is the set of all monotonic functions.

2.1 Loss Functions

Recall that our goal is to use all criteria scores, and their corresponding overall recommendations, to learn an aggregate function $\widehat{f}$ that captures the opinions of all reviewers on how criteria scores should be mapped to recommendations. Inspired by empirical risk minimization, we do this by computing the function in $\mathcal{F}$ that minimizes the $L(p, q)$ loss on the data. In more detail, given hyperparameters $p \in [1, \infty], q \in [1, \infty]$, we compute

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i \in \mathcal{R}} \left[\sum_{a \in P(i)} |y_{ia} - f(\mathbf{x}_{ia})|^p \right]^{\frac{q}{p}} \right\}^{\frac{1}{q}}. \tag{1}$$

This class of $L(p, q)$ losses ($p \in [1, \infty], q \in [1, \infty]$) is a matrix-extension of the more common L_p losses on vectors, and represents a general and popular class as we discuss below. In words, the loss is computed by taking the L_q norm over the loss associated with individual reviewers, where the loss associated to a reviewer is defined as the L_p norm computed on the error of f with respect to the reviewer's overall recommendations. We refer to aggregation by minimizing $L(p, q)$ loss as defined in Equation (1) as "$L(p, q)$ aggregation."

For a function f, the $L(p, q)$ loss is simply the $L(p, q)$ matrix norm of the difference between the matrix $[y_{ia}]_{i \in \mathcal{R}, a \in \mathcal{P}}$ and the matrix $[f(\mathbf{x}_{ia})]_{i \in \mathcal{R}, a \in \mathcal{P}}$ (the entry of the matrices is set to zero if the reviewer does not review the paper). The class of $L(p, q)$ norms represents the standard "entrywise" class of matrix norms. It includes various popular matrix norms as special cases such as the Frobenius norm ($p = q = 2$), the max norm ($p = q = \infty$), and the 1-norm ($p = 1, q = \infty$). This class has had numerous applications in machine learning, statistics, and signal processing (Kowalski, 2009; Ding et al., 2006; Kong et al., 2011; Nie et al., 2010; Zhaoshui and Cichocki, 2008; Rahimpour et al., 2017; Kashlak and Kong, 2021; Cai et al., 2011). Moreover, unlike some other matrix norm classes (like Schatten or induced norm classes) the entrywise $L(p, q)$ class is quite interpretable; for instance, the $L(1, 1)$ loss simply sums up the absolute differences between the overall scores given by reviewers and those given by the function f.

Equation (1) does not specify how to break ties between multiple minimizers. For concreteness, we use the minimum L_2 norm for tie-breaking (although of our results hold

under any reasonable tie-breaking method, such as the minimum L_ℓ norm for any $\ell \in (1, \infty)$). Formally, letting

$$\widehat{F} = \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i \in \mathcal{R}} \left[\sum_{a \in P(i)} |y_{ia} - f(\mathbf{x}_{ia})|^p \right]^{\frac{q}{p}} \right\}^{\frac{1}{q}}$$

be the set of all $L(p, q)$ loss minimizers, we break ties by choosing

$$\widehat{f} \in \operatorname*{argmin}_{f \in \widehat{F}} \sqrt{\sum_{i \in \mathcal{R}} \sum_{a \in P(i)} f(\mathbf{x}_{ia})^2}. \tag{2}$$

Observe that since the $L(p, q)$ loss and constraint set are convex, $\widehat{F}$ is also a convex set. Hence, $\widehat{f}$ as defined by Equation (2) is unique.

Once the function $\widehat{f}$ has been computed, it can be applied to every review (for all reviewers i and papers a) to obtain a new overall recommendation $\widehat{f}(\mathbf{x}_{ia})$. There is a separate — almost orthogonal — question of how to aggregate the overall recommendations of several reviewers on a paper into a single recommendation. In our theoretical results we are agnostic to how this additional aggregation step is performed, but we return to it in our experiments in Section 4.

We remark that an alternative approach would be to learn a monotonic function $\widehat{g}_i : \mathbb{X} \to \mathbb{Y}$ for each reviewer (which best captures their recommendations), and then aggregate these functions into a single function $\widehat{f}$. We chose not to pursue this approach, because in practice there are very few examples per reviewer, so it is implausible that we would be able to accurately learn the reviewers' individual functions.

2.2 Axiomatic Properties

In social choice theory, the most common approach — primarily attributed to Arrow (1951) — for comparing different aggregation methods is to determine which desirable axioms they satisfy. We take the same approach in order to determine the values of the hyperparameters p and q for the $L(p, q)$ aggregation in Equation (1).

We stress that axioms are defined for aggregation methods and not aggregate functions. Informally, an aggregation method is a function that takes as input all the reviews $\{(\mathbf{x}_{ia}, y_{ia})\}_{i \in \mathcal{R}, a \in P(i)}$, and outputs an aggregate function $\widehat{f} : \mathbb{X} \to \mathbb{Y}$. We do not define an aggregation method formally to avoid introducing cumbersome notation that will largely be useless later. It is clear that for any choice of hyperparameters $p, q \in [1, \infty]$, $L(p, q)$ aggregation (with tie-breaking as defined by Equation 2) is an aggregation method.

Social choice theory essentially relies on counterfactual reasoning to identify scenarios where it is clear how an aggregation method should behave. To give one example, the *Pareto efficiency* property of voting rules states that if all voters prefer alternative a to alternative b, then b should *not* be elected; this situation is extremely unlikely to occur, yet Pareto efficiency is obviously a property that any reasonable voting must satisfy. With this principle in mind, we identify a setting in our problem where the requirements are very clear, and then define our axioms in that setting.

For all of our axioms, we restrict attention to scenarios where every reviewer reviews every paper, that is, $P(i) = \mathcal{P}$ for every i. Moreover, we assume that the papers have 'objective' criteria scores, that is, the criteria scores given to a paper are the same across all reviewers, so the only source of disagreement is how the criteria scores should be mapped to an overall recommendation. We can then denote the criteria scores of a paper a simply as $\mathbf{x}_a$, as opposed to $\mathbf{x}_{ia}$, since they do not depend on i. We stress that our framework does *not* require these assumptions to hold — they are only used in our axiomatic characterization, namely Theorem 1 in the next section.

An axiom is satisfied by an aggregation method if its statement holds for every possible number of reviewers n and number of papers m, and for all possible criteria scores and overall recommendations. We start with the simplest axiom, consensus, which informally states that if there is a paper such that all reviewers give it the same overall recommendation, then $\widehat{f}$ must agree with the reviewers; this axiom is closely related to the *unanimity* axiom in social choice.

Axiom 1 (Consensus)**.** *For any paper $a \in \mathcal{P}$, if all reviewers report identical overall recommendations $y_{1a} = y_{2a} = \cdots = y_{ma} = r$ for some $r \in \mathbb{Y}$, then $\widehat{f}(\mathbf{x}_a) = r$.*

Before presenting the next axiom, we require another definition: we say that paper $a \in \mathcal{P}$ *dominates* paper $b \in \mathcal{P}$ if there exists a bijection $\sigma : \mathcal{R} \to \mathcal{R}$ such that for all $i \in \mathcal{R}$, $y_{ia} \geq y_{\sigma(i)b}$. Equivalently (and less formally), paper a dominates paper b if the *sorted* overall recommendations given to a pointwise-dominate the *sorted* overall recommendations given to b. Intuitively, in this situation, a should receive a (weakly) higher overall recommendation than b, which is exactly what the axiom requires; it is similar to the classic Pareto efficiency axiom mentioned above.

Axiom 2 (Efficiency)**.** *For any pair of papers $a, b \in \mathcal{P}$, if a dominates b, then $\widehat{f}(\mathbf{x}_a) \geq \widehat{f}(\mathbf{x}_b)$.*

Our positive result, which will be presented shortly, satisfies this notion of efficiency. On the other hand, we also use this axiom to prove a negative result; an important note is that *the negative result requires a condition that is significantly weaker than the aforementioned definition of efficiency.* We revisit this point about requiring a much weaker condition for the negative result at the end of Section 3.2.1.

Our final axiom is *strategyproofness*, a game-theoretic property that plays a major role in social choice theory (Moulin, 1983). For the application of peer review, we consider strategyproofness motivated by the many instances of strategic behavior uncovered and studied recently in peer review (Balietti et al., 2016; Xu et al., 2019; Vijaykumar, 2020a,b; Jecmen et al., 2020; Stelmakh et al., 2021a). Intuitively, in our problem setting, strategyproofness means that reviewers have no incentive to misreport their overall recommendations: they cannot bring the aggregate recommendations — the community's consensus about the relative importance of various criteria — closer to their own through strategic manipulation.[3]

Axiom 3 (Strategyproofness)**.** *For each reviewer $i \in \mathcal{R}$, and all possible manipulated recommendations $\mathbf{y}'_i \in \mathbb{Y}^m$, if $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})$ is replaced with $\mathbf{y}'_i$, then*

$$\|(\widehat{f}(\mathbf{x}_1), \ldots, \widehat{f}(\mathbf{x}_m)) - \mathbf{y}_i\|_2 \leq \|(\widehat{g}(\mathbf{x}_1), \ldots, \widehat{g}(\mathbf{x}_m)) - \mathbf{y}_i\|_2, \tag{3}$$

3. This is vaguely reminiscent of work on impartial peer evaluation (Alon et al., 2011; Kurokawa et al., 2015; Kahng et al., 2018; Aziz et al., 2019; Xu et al., 2019), but in that setting the utility of a reviewer depends on whether their own paper was selected — a strategic issue that is orthogonal to our work.

where $\widehat{f}$ and $\widehat{g}$ are the aggregate functions obtained from the original and manipulated reviews, respectively.

The use of the L_2 norm in the definition (3) of the strategyproofness axiom is made only for concreteness, and all our results hold for any norm L_ℓ, $\ell \in [1, \infty]$.

3. Main Result

In Section 2, we introduced $L(p, q)$ aggregation as a family of rules for aggregating individual opinions towards a consensus mapping from criteria scores to recommendations. But that definition, in and of itself, leaves open the question of how to choose the values of p and q in a way that leads to the most socially desirable outcomes. The axioms of Section 2.2 allow us to give a satisfying answer to this question. Specifically, our main theoretical result is a characterization of $L(p, q)$ aggregation in terms of the three axioms.

Theorem 1. *$L(p, q)$ aggregation, where $p, q \in [1, \infty]$, satisfies consensus, efficiency, and strategyproofness if and only if $p = q = 1$.*

We remark that for $p = q$, Equation (1) does not distinguish between different reviewers, that is, the aggregation method pools all reviews together. We find this interesting, because the $L(p, q)$ aggregation framework does have enough power to make that distinction, but the axioms guide us towards a specific solution, $L(1, 1)$, which does not.

Turning to the proof of the theorem, we start from the easier 'if' direction.

3.1 $p = q = 1$ Satisfies All Three Axioms

Lemma 1. *$L(p, q)$ aggregation with $p = q = 1$ satisfies consensus, efficiency and strategyproofness.*

Proof. The key idea of the proof lies in the form taken by the minimizer of $L(1, 1)$ loss. When each reviewer reviews every paper and the papers have objective criteria scores, $L(1, 1)$ aggregation reduces to computing

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i \in \mathcal{R}} \sum_{a \in \mathcal{P}} |y_{ia} - f(\mathbf{x}_a)| \right\}, \tag{4}$$

where ties are broken by picking the minimizer with minimum L_2 norm. We claim that the aggregate function is given by

$$\widehat{f}(\mathbf{x}_a) = \text{left-med}(\{y_{ia}\}_{i \in \mathcal{R}}) \quad \forall a \in \mathcal{P},$$

where left-med$(\cdot)$ of a set of points is their left median. We prove this claim by showing four parts:

(i) $\widehat{f}$ is a valid function,

(ii) $\widehat{f}$ is an unconstrained minimizer of the objective in (4),

(iii) $\widehat{f}$ satisfies the constraints of (4), i.e., $\widehat{f} \in \mathcal{F}$, and

(iv) $\widehat{f}$ has the minimum L_2 norm among all minimizers of (4).

We start by proving part (i). This part can only be violated if there are two papers a and b such that $\mathbf{x}_a = \mathbf{x}_b$, but left-med$(\{y_{ia}\}_{i\in\mathcal{R}}) \neq$ left-med$(\{y_{ib}\}_{i\in\mathcal{R}})$, leading to $\widehat{f}$ having two function values for the same $\mathbf{x}$-value. However, we assumed that each reviewer i has a function $g_i^\star$ used to score the papers. So, for the two papers a and b, we would have $y_{ia} = g_i^\star(\mathbf{x}_a) = g_i^\star(\mathbf{x}_b) = y_{ib}$ for every i, giving us left-med$(\{y_{ia}\}_{i\in\mathcal{R}}) =$ left-med$(\{y_{ib}\}_{i\in\mathcal{R}})$. Therefore, $\widehat{f}$ is a valid function.

For part (ii), consider the optimization problem (4) without any constraints. Denote the objective function as $G(f)$. Rearranging terms, we obtain

$$G(f) = \sum_{a\in\mathcal{P}} \sum_{i\in\mathcal{R}} |y_{ia} - f(\mathbf{x}_a)| \,. \tag{5}$$

Consider the inner summation $\sum_{i\in\mathcal{R}} |y_{ia} - f(\mathbf{x}_a)|$; it is well known that this quantity is minimized when $f(\mathbf{x}_a)$ is any median of the $\{y_{ia}\}_{i\in\mathcal{R}}$ values. Hence, we have

$$\begin{aligned} G(f) &= \sum_{a\in\mathcal{P}} \sum_{i\in\mathcal{R}} |y_{ia} - f(\mathbf{x}_a)| \\ &\geq \sum_{a\in\mathcal{P}} \sum_{i\in\mathcal{R}} |y_{ia} - \text{left-med}(\{y_{ia}\}_{i\in\mathcal{R}})| \\ &= G(\widehat{f}), \end{aligned} \tag{6}$$

where f is an arbitrary function. Therefore, $\widehat{f}$ minimizes the objective function even in the absence of any constraints, proving part (ii).

Turning to part (iii), we show that $\widehat{f}$ satisfies the monotonicity constraints, i.e., $\widehat{f} \in \mathcal{F}$. Suppose $a, b \in \mathcal{P}$ are such that $\mathbf{x}_a \geq \mathbf{x}_b$. Using the fact that each reviewer i scores papers based on the function $g_i^\star$, we have $y_{ia} = g_i^\star(\mathbf{x}_a)$ and $y_{ib} = g_i^\star(\mathbf{x}_b)$. And since $g_i^\star \in \mathcal{F}$ obeys monotonicity constraints, we obtain $y_{ia} \geq y_{ib}$ for every i. This trivially implies that left-med$(\{y_{ia}\}_{i\in\mathcal{R}}) \geq$ left-med$(\{y_{ib}\}_{i\in\mathcal{R}})$, i.e., $\widehat{f}(\mathbf{x}_a) \geq \widehat{f}(\mathbf{x}_b)$, completing part (iii).

Finally, we prove part (iv). Observe that Equation (6) is a strict inequality if there is a paper a for which $f(\mathbf{x}_a)$ is not a median of the $\{y_{ia}\}_{i\in\mathcal{R}}$ values. In other words, the only functions f that have the same objective function value as $\widehat{f}$ are of the form

$$f(\mathbf{x}_a) \in \text{med}(\{y_{ia}\}_{i\in\mathcal{R}}) \quad \forall a \in \mathcal{P}, \tag{7}$$

where med$(\cdot)$ of a collection of points is the set of all points between (and including) the left and right medians. Hence, all other minimizers of (4) must satisfy Equation (7). Observe that $\widehat{f}$ is pointwise smaller than any of these functions, since it computes the left median at each of the $\mathbf{x}$-values. Therefore, $\widehat{f}$ has the minimum L_2 norm among all possible minimizers of (4), completing the proof of part (iv).

Combining all four parts proves that $\widehat{f}$ is indeed the aggregate function chosen by $L(1,1)$ aggregation. We use this to prove that $L(1,1)$ aggregation satisfies consensus, efficiency and strategyproofness.

Consensus. Let $a \in \mathcal{P}$ be a paper such that $y_{1a} = y_{2a} = \cdots = y_{ma} = r$ for some r. Then, left-med$(\{y_{ia}\}_{i\in\mathcal{R}}) = r$. Hence, $\widehat{f}(\mathbf{x}_a) = r$, satisfying consensus.

Efficiency. Let $a, b \in \mathcal{P}$ be such that a dominates b. In other words, the sorted overall recommendations given to a pointwise-dominate the sorted overall recommendations given to b. So, by definition, left-med$(\{y_{ia}\}_{i\in\mathcal{R}})$ is at least as large as left-med$(\{y_{ib}\}_{i\in\mathcal{R}})$. That is, $\widehat{f}(\mathbf{x}_a) \geq \widehat{f}(\mathbf{x}_b)$, satisfying efficiency.

Strategyproofness. Let i be an arbitrary reviewer. Observe that in this setting, the aggregate score $\widehat{f}(\mathbf{x}_a)$ of a paper a depends only on the score y_{ia} and not on other scores $\{y_{ib}\}_{b\neq a}$ given by reviewer i. In other words, the only way to manipulate $\widehat{f}(\mathbf{x}_a) = \text{left-med}(\{y_{i'a}\}_{i'\in\mathcal{R}})$ is by changing y_{ia}. Consider three cases. Suppose $y_{ia} < \widehat{f}(\mathbf{x}_a)$. In this case, if reviewer i reports $y'_{ia} \leq \widehat{f}(\mathbf{x}_a)$, then there is no change in the aggregate score of a. On the other hand, if $y'_{ia} > \widehat{f}(\mathbf{x}_a)$, then either the aggregate score of a remains the same or increases, making things only worse for reviewer i. The other case of $y_{ia} > \widehat{f}(\mathbf{x}_a)$ is symmetric to $y_{ia} < \widehat{f}(\mathbf{x}_a)$. Consider the third case, $y_{ia} = \widehat{f}(\mathbf{x}_a)$. In this case, manipulation can only make things worse since we already have $|y_{ia} - \widehat{f}(\mathbf{x}_a)| = 0$. In summary, reporting y'_{ia} instead of y_{ia} cannot help decrease $|y_{ia} - \widehat{f}(\mathbf{x}_a)|$. Also, recall that y_{ia} does not affect the aggregate scores of other papers, and hence manipulation of y_{ia} does not help them either. Therefore, by manipulating any of the y_{ia} scores, reviewer i cannot bring the aggregate recommendations closer to her own, proving strategyproofness. □

3.2 Violation of the Axioms When $(p, q) \neq (1, 1)$

We now tackle the harder 'only if' direction of Theorem 1. We do so in three steps: efficiency is violated by $p \in (1, \infty)$ and $q = 1$ (Lemma 2), strategyproofness is violated by $L(p, q)$ aggregation for all $q > 1$ (Lemma 3), and consensus is violated by $p = \infty$ and $q = 1$ (Lemma 4). Together, the three lemmas leave $p = q = 1$ as the only option. Below we state the lemmas and give some proof ideas; the theorem's full proof is relegated to Appendix A.

It is worth noting that, although we have presented the lemmas as components in the proof of Theorem 1, they also have standalone value (some more than others). For example, if one decided that only strategyproofness is important, then Lemma 3 below would give significant guidance on choosing an appropriate method.

3.2.1 Violation of Efficiency

In our view, the following lemma presents the most interesting and counter-intuitive result in the paper.

Lemma 2. *$L(p, q)$ aggregation with $p \in (1, \infty)$ and $q = 1$ violates efficiency.*

It is quite surprising that such reasonable loss functions violate the simple requirement of efficiency. In what follows explain this phenomenon via a connection between our problem and the notion of the 'Fermat point' of a triangle (Spain, 1996). The explanation provided here demonstrates the negative result for $L(2, 1)$ aggregation. The complete proof of the lemma for general values of $p \in (1, \infty)$ is quite involved, as can be seen in Appendix A.

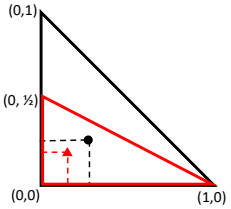
The construction of the negative result is illustrated in Figure 1 and described in more detail here. Consider a setting with 3 reviewers and 2 papers, where each reviewer reviews both papers. We let $\mathbf{x}_1$ and $\mathbf{x}_2$ denote the respective objective criteria scores of the two papers. Assume that no score in $\{\mathbf{x}_1, \mathbf{x}_2\}$ is pointwise greater than or equal to the other score in that set; an example is shown in Figure 1(a). Let the overall recommendations

Criteria scores: $\mathbf{x}_1 = [1/4, 3/4], \mathbf{x}_2 = [3/4, 1/4]$
Overall scores (y_{ij}'s):

	Paper 1	Paper 2
Reviewer 1	0	0
Reviewer 2	1	0
Reviewer 3	0	z

(a) Example data with 2 papers, 3 reviewers, and 2 criteria. When $z = 1$ the data for the two papers is symmetric. Setting $z < 1$ makes Paper 1 dominate Paper 2 and the efficiency axiom mandates $\widehat{f}(\mathbf{x}_1) \geq \widehat{f}(\mathbf{x}_2)$.

$(\widehat{f}(\mathbf{x}_1), \widehat{f}(\mathbf{x}_2))$ is the Fermat point of triangle with vertices $(y_{11}, y_{12}), (y_{21}, y_{22}), (y_{31}, y_{32})$:



(b) The Fermat point is (.21, .21) when $z = 1$ (black circle), but (.12, .15) when $z = 1/2$ (red triangle). Hence $L(2,1)$ aggregation with $z = 1/2$ results in $\widehat{f}(\mathbf{x}_1) < \widehat{f}(\mathbf{x}_2)$.

Figure 1: An example of aggregation under $L(2,1)$ loss, and violation of the efficiency axiom.

given by the reviewers be $y_{11} = 0$, $y_{21} = 1$, $y_{31} = 0$ to the first paper and $y_{12} = 0$, $y_{22} = 0$ and $y_{23} = z$ to the second paper. Under these scores, let $\widehat{f}$ denote the aggregate function that minimizes the $L(2,1)$ loss. We see that in this data, when $z < 1$, the first paper dominates the second, and hence the efficiency axiom mandates $\widehat{f}(\mathbf{x}_1) \geq \widehat{f}(\mathbf{x}_2)$.

The outcome of the $L(2,1)$ aggregation is related to the notion of the Fermat point of a triangle. The Fermat point of a triangle is a point such that the sum of its (Euclidean) distances from all three vertices is minimum. Consider a triangle in $\mathbb{R}^2$ with vertices $(y_{11}, y_{12}), (y_{21}, y_{22}), (y_{31}, y_{32})$; see Figure 1(b). Then by definition, the Fermat point of this triangle is exactly $(\widehat{f}(\mathbf{x}_1), \widehat{f}(\mathbf{x}_2))$. Intuitively, the $p = 2$ in the $L(p = 2, q = 1)$ loss relates to the Euclidean distances used in the Fermat point, and the $q = 1$ relates to summing the distances to all vertices.

In our specific example, we have $(y_{11}, y_{12}) = (0, 0), (y_{21}, y_{22}) = (1, 0)$, and $(y_{31}, y_{32}) = (0, z)$. When $z = 1$ the Fermat point equals $(0.21, 0.21)$ and is symmetric across both coordinates. Now when the vertex $(z, 0)$ is moved down (by decreasing z), the Fermat point paradoxically moves to the left and down. For instance, setting $z = \frac{1}{2}$, the Fermat point of this triangle as $(0.12, 0.15)$. Thus the aggregate score of paper 1 under $L(2,1)$ aggregation is strictly smaller than of paper 2, thereby violating efficiency for the $L(2,1)$ loss.

As a final but important remark, the proof of Lemma 2 only requires a *significantly weaker notion of efficiency*. In this weaker notion, we first consider two papers 1 and 2 such that their reviews are *symmetric*: formally, switching the labels "1" and "2" and switching the labels of some reviewers and criteria leaves the data unchanged.[4] The weaker version of efficiency says that *reducing* the review scores of paper 2 mandates $\widehat{f}(\mathbf{x}_1) \geq \widehat{f}(\mathbf{x}_2)$. In the example of Figure 1(a), when $z = 1$, switching the labels of the two papers, the labels of reviewers 2 and 3, and the labels of the two criteria yields data identical to the original. In the example above, reducing z to $z < 1$ breaks the symmetry and makes paper 2 inferior to paper 1 in this data. The axiom requires $\widehat{f}(\mathbf{x}_1) \geq \widehat{f}(\mathbf{x}_2)$ in this case.

4. Note that we assume no prior importance on various criteria and any such importance should be learnt from the data.

3.2.2 Violation of Strategyproofness

Lemma 3. *$L(p, q)$ aggregation with $q \in (1, \infty]$ violates strategyproofness.*

We prove the lemma via a simple construction with just one paper and two reviewers, who give the paper overall recommendations of 1 and 0, respectively. For $q \in (1, \infty)$, the aggregate score is

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathbb{R}} \Big\{ |1 - f|^q + |f|^q \Big\},$$

and for $q = \infty$, it is

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathbb{R}} \ \max\big(|1 - f|, |f|\big).$$

Either way, the unique minimum is obtained at an aggregate score of 0.5. If reviewer 1 reported an overall recommendation of 2, however, the aggregate score would be 1, which matches her 'true' recommendation, thereby violating strategyproofness. See Appendix A.2 for the complete proof.

3.2.3 Violation of Consensus

Lemma 4. *$L(p, q)$ aggregation with $p = \infty$ and $q = 1$ violates consensus.*

Lemma 4 is established via another simple construction: two papers, two reviewers, and overall recommendations

$$\mathbf{y} = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix},$$

where y_{ia} denotes the overall recommendation given by reviewer i to paper a. Crucially, the two reviewers agree on an overall recommendation of 1 for paper 2, hence the aggregate score of this paper must also be 1. But we show that $L(\infty, 1)$ aggregation would *not* return an aggregate score of 1 for paper 2. The formal proof appears in Appendix A.3.

4. Implementation and Experimental Results

In this section, we provide an empirical analysis of a few aspects of peer review through the approach of this paper. We employ a dataset of reviews from the 26$^{\text{th}}$ International Joint Conference on Artificial Intelligence (IJCAI 2017), which was made available to us by the program chair. To our knowledge, we are the first to use this dataset.

At submission time, authors were asked if review data for their paper could be included in an anonymized dataset, and, similarly, reviewers were asked whether their reviews could be included; the dataset provided to us consists of all reviews for which permission was given. Each review is tagged with a reviewer ID and paper ID, which are anonymized for privacy reasons. The criteria used in the conference are 'originality', 'relevance', 'significance', 'quality of writing' (which we call 'writing'), and 'technical quality' (which we call 'technical'), and each is rated on a scale from 1 to 10. Overall recommendations are also on a scale from 1 to 10. In addition, information about which papers were accepted and which were rejected is included in the dataset.

The number of papers in the dataset is 2380, of which 649 were accepted, which amounts to 27.27%. This is a large subset of the 2540 submissions to the conference, of which 660

# of reviews by a reviewer	1	2	3	4	5	6	7	8	≥ 9
Frequency	238	96	92	120	146	211	628	187	7

Table 1: Distribution of number of papers reviewed by a reviewer.

were accepted, for an actual acceptance rate of 25.98%. The number of reviewers in the dataset is 1725, and the number of reviews is 9197. All but nine papers in the dataset have three reviews (485 papers), four reviews (1734 papers), or five reviews (152 papers). Table 1 shows the distribution of the number of papers reviewed by reviewers.

We apply $L(1,1)$ aggregation (i.e., $p = q = 1$), as given in Equation (1), to this dataset to learn the aggregate function. Let us denote that function by $\widetilde{f}$.[5] The optimization problem in Equation (1) is convex, and standard optimization packages can efficiently compute the minimizer. Hence, importantly, computational complexity is a nonissue in terms of implementing our approach.

Once we compute the aggregate function $\widetilde{f}$, we calculate the aggregate overall recommendation of each paper a by taking the median of the aggregate reviewer scores for that paper obtained by applying $\widetilde{f}$ to the objective scores:

$$y_{\widetilde{f}}(a) = \text{median}(\{\widetilde{f}(\mathbf{x}_{ia})\}_{i \in R(a)}) \quad \forall a \in \mathcal{P}. \tag{8}$$

In case of multiple medians in (8), we took the mean of all medians. Recalling that 27.27% of the papers in the dataset were actually accepted to the conference, in our experiments we define the set of papers accepted by the aggregate function $\widetilde{f}$ as the the top 27.27% of papers according to their respective $y_{\widetilde{f}}$ values. We now present the specific experiments we ran, and their results.

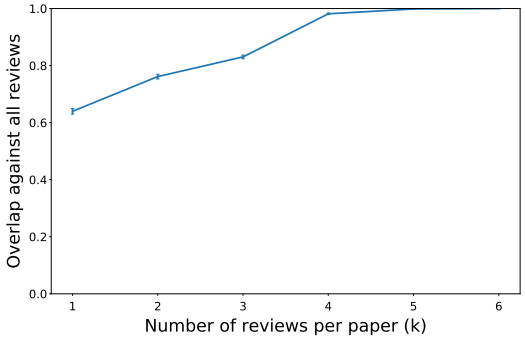


Figure 2: Fraction overlap as number of reviews per paper is restricted. Error bars depict 95% confidence intervals, but may be too small to be visible for $k = 4, 5$.

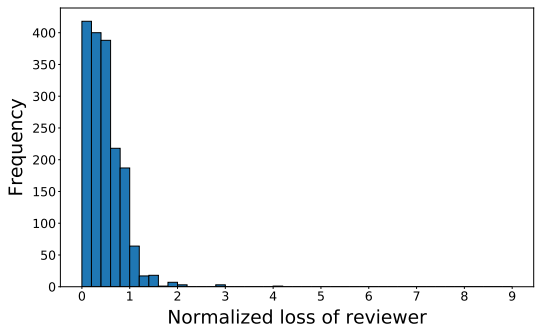


Figure 3: Frequency of losses of the reviewers for $L(1,1)$ aggregation, normalized by the number of papers reviewed by the respective reviewers.

5. Code available at `https://github.com/ritesh-noothigattu/choosing-how-to-choose-papers`.

4.1 Varying Number of Reviewers

In our first experiment, for each value of a parameter $k \in \{1, \ldots, 5\}$, we subsampled k distinct reviews for each paper uniformly at random from the set of all reviews for that paper (if the paper had fewer than k to begin with then we retained all the reviews). We then computed an aggregate function, $\widehat{f}_k$, via $L(1,1)$ aggregation applied only to these subsampled reviews. Next, we found the set of top 27.27% papers as given by $\widehat{f}_k$ applied to the subsampled reviews. Finally, we compared the overlap of this set of top papers for every value of k with the set of top 27.27% papers as dictated by the overall aggregate function $\widetilde{f}$.

The results from this experiment are plotted in Figure 2, and lead to several observations. First, the incremental overlap from $k = 4$ to 5 is very small because there are very few papers that had 5 or more reviews. Second, we see that the amount of overlap monotonically increases with the number of reviewers per paper k, thereby serving as a sanity check on the data as well as our methods. Third, we observe the overlap to be quite high ($\approx 60\%$) even with a single reviewer per paper.

4.2 Loss Per Reviewer

Next, we look at the loss of different reviewers, under $\widetilde{f}$ (obtained by $L(1,1)$ aggregation). In order for the losses to be on the same scale, we normalize each reviewer's loss by the number of papers reviewed by them. Formally, the normalized loss of reviewer i (for $p = 1$) is

$$\frac{1}{|P(i)|} \sum_{a \in P(i)} |y_{ia} - \widetilde{f}(\mathbf{x}_{ia})|.$$

The normalized loss averaged across reviewers is found to be 0.470, and the standard deviation is 0.382. Figure 3 shows the distribution of the normalized loss of all the reviewers. Note that the normalized loss of a reviewer can fall in the range $[0, 9]$. These results thus indicate that the function $\widetilde{f}$ is indeed at least a reasonable representation of the mapping of the broader community.

4.3 Overlap of Accepted Papers

We also compute the overlap between the set of top 27.27% papers selected by $L(1,1)$ aggregation $\widetilde{f}$ with the *actual* 27.27% accepted papers. It is important to emphasize that we believe the set of papers selected by our method is *better* than any hand-crafted or rule-based decision using the scores, since this aggregate represents the opinion of the community. Hence, to be clear, we do *not* have a goal of maximizing the overlap. Nevertheless, a very small overlap would mean that our approach is drastically different from standard practice, which would potentially be disturbing. We find that the overlap is 79.2%, which we think is quite fascinating — our approach does make a significant difference, but the difference is not so drastic as to be disconcerting.

Out of intellectual curiosity, we also computed the pairwise overlaps of the papers accepted by $L(p, q)$ aggregation, for $p, q \in \{1, 2, 3\}$. We find that the choice of the reviewer-norm hyperparameter q has more influence than the paper-norm hyperparameter p; we refer the reader to Appendix B.1 for details.

4.4 A Visualization of the Learnt Mapping

In Appendix B.2 we present visualizations and interpretations of $L(1,1)$ aggregate mapping learnt from the IJCAI 2017 data, which provide insights into the preferences of the community. We present here the key takeaways based on visual inspection of the visualizations, and refer the reader to the appendix for more detail. First, we observe that writing and relevance do not have a significant influence on the overall recommendations: Really bad writing or relevance is a significant downside, excellent writing or relevance is appreciated, but everything else in between in irrelevant. Second, technical quality and significance exert a high and approximately linear influence. Third, if modeling this mapping, linear models are partially applicable — for some criteria one may indeed assume a linear model, but not for all.

5. Limitations, Discussion, and Open Problems

We address the problem of subjectivity in peer review by combining approaches from machine learning and social choice theory. A key challenge in the setting of peer review (e.g., when choosing a loss function) is the absence of ground truth, and we overcome this challenge via a principled, axiomatic approach.

Our work also contributes to recent endeavors in *understanding* the peer review process (Lawrence and Cortes, 2014; Shah et al., 2018; Tomkins et al., 2017; Stelmakh et al., 2021b, 2020, 2021c). Specifically, the mapping learnt via L(1,1) aggregation can be used to understand the community's aggregate preferences over various criteria. We illustrate this via an empirical analysis in Section 4 and in Appendix B.

A critical aspect of peer review is confidentiality or privacy (Ding et al., 2020) with respect to who reviewed which paper. There are other values of p and q where the aggregate mapping could potentially reveal some information about individual reviewers (e.g., that two specific reviews were written by the same person). On the other hand, L(1,1) aggregation performs the optimization (1) by simply pooling all reviews together, and does not use any association of who reviewed which paper. It thus guards against this issue, and can even be executed on publicly available data for conferences following open review (i.e., where all reviews are public but reviewer identities are private).

One can think of the theoretical results of Section 3 as supporting $L(1,1)$ aggregation using the tools of social choice theory, whereas the empirical results of Section 4 focus on studying its behavior on real data. Understanding this helps clear up a possible source of confusion: are we not overfitting by training on a set of reviews, and then applying the aggregate function to the same reviews? The answer is negative, because the process of learning the function $\widehat{f}$ amounts to an aggregation of opinions about how criteria scores should be mapped to overall recommendations. Applying it to the data yields recommendations in $\mathbb{Y}$, whereas this function from $\mathbb{X}$ to $\mathbb{Y}$ lives in a different space.

We now conclude with a discussion of the limitations of our work and relevant open problems.

- Our framework assumes that the set of criteria listed by the program chairs encapsulates the criteria used by any reviewer for evaluating a paper. To address situations where this is violated, the program chairs may solicit information on the insufficiency

of the criteria from reviewers directly, and this information can also help improve the choice of criteria used in subsequent conferences. On a technical front, this leads to an open problem of designing statistical methods to detect the insufficiency of given criteria in conference peer review (see also Shah et al., 2018, Section 3.9).

- It is of interest to understand the statistical aspects of estimating the community's consensus mapping function, assuming the existence of a ground truth. In more detail, suppose each reviewer's true function $g_i^\star$ is a noisy version of some underlying function $f^{\star\star}$ that represents the community's beliefs. Then how can one recover $f^{\star\star}$ in a statistically efficient manner, perhaps via $L(1,1)$ aggregation or otherwise? Conceptually this non-parametric estimation problem is related to isotonic regression (Shah et al., 2017; Gao and Wellner, 2007; Chatterjee et al., 2018). The key difference is that the observations in our setting consist of evaluations of multiple functions, where each such function is a noisy version of the original monotonic function. In contrast, isotonic regression is primarily concerned with noisy evaluations of a common function. Nevertheless, insights from isotonic regression suggest that the monotonicity assumption of our setting can yield attractive — and sometimes near-parametric (Shah et al., 2017, 2019, 2020) — rates of estimation.

- It is of interest to further incorporate additional information from reviews such as self-reported confidence (MacKay et al., 2017) or self-reported expertise of reviewers (by, e.g., reweighting the terms in the $L(1,1)$ aggregation accordingly) or even the review text (Hua et al., 2019; Manzoor and Shah, 2021).

- There are various other problems in peer review such as miscalibration (Ge et al., 2013; Roos et al., 2011; Wang and Shah, 2019), noise (Stelmakh et al., 2019a), fraud (Vijaykumar, 2020a,b; Jecmen et al., 2020), biases (Tomkins et al., 2017; Stelmakh et al., 2019b; Nielsen et al., 2021). These problems have been treated independently of each other in the literature, and addressing them jointly along with the problem of subjectivity is a challenging and important open problem.

- Our framework assumes that reviewers use criteria to come up with an overall score. In practice, some reviewers may first arrive at a overall judgment and tailor criteria scores to fit their overall judgment. The instructions for reviewing could be designed to mitigate this issue.

- Our work focuses on learning one representative aggregate mapping for the entire community of reviewers. Instead, the program chairs of a conference may wish to allow for multiple mappings that represent the aggregate opinions of different subcommunities (e.g., theoretical or applied researchers). In this case, one may modify our framework to also learn this (unknown) partition of reviewers and/or papers into multiple sub-communities with different mapping functions, and frame the problem in terms of learning a mixture model. The design of computationally efficient algorithms for $L(p,q)$ aggregation under such a mixture model is a challenging open problem.

As a final remark, we see our work as an unusual synthesis between computational social choice and machine learning. We hope that our approach will inspire exploration of additional connections between these two fields of research, especially in terms of viewing choices

made in machine learning — often in an *ad hoc* fashion — through the lens of computational social choice.

Acknowledgments

We thank the JAIR reviewers for their valuable feedback, and to the associate editor Haris Aziz for his efficient handling of the paper. We are grateful to Francisco Cruz for compiling the IJCAI 2017 review dataset, and to Carles Sierra for making it available to us.

Shah was supported in part by NSF grants CRII-CCF-1755656 and CCF-1763734. Noothigattu and Procaccia were supported in part by NSF grants IIS-1350598, IIS-1714140, CCF-1525932, and CCF-1733556; by ONR grants N00014-16-1-3075 and N00014-17-1-2428; and by a Sloan Research Fellowship and a Guggenheim Fellowship.

Appendix A. Proof Of Theorem 1

Recall that the proof of our main result, Theorem 1, includes four lemmas. Here we prove the three lemmas whose proofs were omitted from the main text.

A.1 Proof of Lemma 2

Consider $L(p, 1)$ aggregation with an arbitrary $p \in (1, \infty)$. We show that efficiency is violated using the following construction. There are 2 papers, 3 reviewers and each reviewer reviews both papers. Assume that the papers have objective criteria scores $\mathbf{x}_1$ and $\mathbf{x}_2$, and that neither of these scores is pointwise greater than or equal to the other. Let the overall recommendations by the reviewers for the papers be defined by the matrix

$$\mathbf{y} = \begin{bmatrix} z & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

where z is a constant strictly bigger than 1 and y_{ia} denotes the overall recommendation by reviewer i to paper a. Observe that paper 1 dominates paper 2. But, we will show that there exists a value $z > 1$ such that the aggregate score of paper 1 is strictly smaller than the aggregate score of paper 2.

Let f_i denote the value of function f on paper i, i.e. $f_i := f(\mathbf{x}_i)$. And let $\widehat{f}_i(z)$ denote the aggregate score of paper i; observe that we write it as a function of z because the aggregate score of each paper would depend on the chosen score z. Since we are minimizing $L(p, 1)$ loss, the aggregate function satisfies:

$$(\widehat{f}_1(z), \widehat{f}_2(z)) \in \operatorname*{argmin}_{(f_1,f_2)\in\mathbb{R}^2} \left\{ \big\|(z, 0) - (f_1, f_2)\big\|_p + \big\|(0, 1) - (f_1, f_2)\big\|_p + \big\|(f_1, f_2)\big\|_p \right\}. \quad (9)$$

We do not have any monotonicity constraints in (9) because the two papers have incomparable criteria scores. For simplicity, let $\mathbf{f} := (f_1, f_2)$, $\widehat{\mathbf{f}}(z) := (\widehat{f}_1(z), \widehat{f}_2(z))$, and denote the objective function in Equation (9) by $G_z(\mathbf{f})$. That is,

$$G_z(f_1, f_2) = \Big[|z - f_1|^p + |f_2|^p\Big]^{\frac{1}{p}} + \Big[|f_1|^p + |1 - f_2|^p\Big]^{\frac{1}{p}} + \Big[|f_1|^p + |f_2|^p\Big]^{\frac{1}{p}}. \quad (10)$$

For the overall proof to be easier to follow, proofs of all claims are given at the end of this proof. Also, just to re-emphasize, the whole proof assumes $z > 1$.

Claim 1. *G_z is a strictly convex objective function.*

Claim 1 states that G_z is strictly convex, implying that it has a unique minimizer $\widehat{\mathbf{f}}(z)$. Hence, there is no need to consider tie-breaking.

Claim 2. *$\widehat{f}_1(z)$ and $\widehat{f}_2(z)$ are bounded. In particular, $\widehat{f}_1(z) \in [0,1]$ and $\widehat{f}_2(z) \in [0,1]$.*

Claim 2 states that the aggregate score of both papers lies in the interval $[0,1]$ irrespective of the value of z. This allow us to restrict ourselves to the region $[0,1]^2$ when computing the minimizer of (10). Hence, for the rest of the proof, we only consider the space $[0,1]^2$. In this region, the optimization problem (9) can be rewritten as

$$(\widehat{f}_1(z), \widehat{f}_2(z)) = \underset{f_1\in[0,1], f_2\in[0,1]}{\operatorname{argmin}} \left\{ \left[\left(z - f_1\right)^p + f_2^p \right]^{\frac{1}{p}} + \left[f_1^p + \left(1 - f_2\right)^p \right]^{\frac{1}{p}} + \left[f_1^p + f_2^p \right]^{\frac{1}{p}} \right\}.$$

To start off, we analyze the objective function as we take the limit of z going to infinity. Later, we show that the observed property holds even for a sufficiently large finite z.

For the limit to exist, redefine the objective function as $H_z(f_1, f_2) = G_z(f_1, f_2) - G_z(0,0)$, i.e.,

$$H_z(f_1, f_2) = \left[\left(z - f_1\right)^p + f_2^p \right]^{\frac{1}{p}} - z + \left[f_1^p + \left(1 - f_2\right)^p \right]^{\frac{1}{p}} + \left[f_1^p + f_2^p \right]^{\frac{1}{p}} - 1. \tag{11}$$

For any value of z, the function H_z has the same minimizer as G_z, that is,

$$(\widehat{f}_1(z), \widehat{f}_2(z)) = \underset{f_1\in[0,1], f_2\in[0,1]}{\operatorname{argmin}} H_z(f_1, f_2).$$

Claim 3. *For any (fixed) $f_1 \in [0,1], f_2 \in [0,1]$,*

$$\lim_{z\to\infty} H_z(f_1, f_2) = H^\star(f_1, f_2),$$

where

$$H^\star(f_1, f_2) = -f_1 + \left[f_1^p + \left(1 - f_2\right)^p \right]^{\frac{1}{p}} + \left[f_1^p + f_2^p \right]^{\frac{1}{p}} - 1. \tag{12}$$

The proof proceeds by analyzing some important properties of the limiting function $H^\star$.

Claim 4. *The function $H^\star(\mathbf{f})$ is convex in $\mathbf{f} \in [0,1]^2$. Moreover, the function $H^\star(\mathbf{f})$ is strictly convex for $f_1 \in (0,1]$ and $f_2 \in [0,1]$.*

Claim 5. *$H^\star$ is minimized at $\widehat{\mathbf{v}} = (\widehat{v}_1, \widehat{v}_2)$, where*

$$\widehat{v}_1 = \frac{1}{2}\left[\frac{1}{(2^{\frac{p}{p-1}} - 1)} \right]^{\frac{1}{p}}, \qquad \widehat{v}_2 = \frac{1}{2}. \tag{13}$$

Claim 6. $\widehat{v}_1 < \widehat{v}_2$.

Observe that Claim 6 is the desired result, but for the limiting objective function $H^\star$. The remainder of the proof proceeds to show that this result holds even for the objective function H_z, when the score z is large enough. Define $\Delta = \widehat{v}_2 - \widehat{v}_1 > 0$. We first show that (i) there exists $z > 1$ such that $\|\widehat{\mathbf{f}}(z) - \widehat{\mathbf{v}}\|_2 < \frac{\Delta}{4}$, and then (ii) show that in this case, we have $\widehat{f}_1(z) < \widehat{f}_2(z)$.

To prove part (i), we first analyze how functions H_z and $H^\star$ relate to each other. Using Claim 3, for any fixed f_1, f_2, by definition of the limit, for any $\epsilon > 0$, there exists z_ϵ (which could be a function of f_1, f_2) such that, for all $z > z_\epsilon$, we have

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| < \epsilon. \tag{14}$$

For a given f_1, f_2, denote the corresponding value of z_ϵ by $z_\epsilon(f_1, f_2)$. And, let $\mathcal{Z}_\epsilon(f_1, f_2)$ denote the set of all values of $z > 1$ for which Equation (14) holds for (f_1, f_2).

Claim 7. *$\mathcal{Z}_\epsilon(1,1) \subset \mathcal{Z}_\epsilon(f_1, f_2)$ for every $(f_1, f_2) \in [0,1]^2$.*

Claim 7 says that if Equation (14) holds for a particular value of z for $f_1 = f_2 = 1$, then for the same value of z it holds for every other value of $(f_1, f_2) \in [0,1]^2$ as well. So, define

$$\widetilde{z}_\epsilon := z_\epsilon(1,1) + 1. \tag{15}$$

By definition, $\widetilde{z}_\epsilon \in \mathcal{Z}_\epsilon(1,1)$. And by Claim 7, $\widetilde{z}_\epsilon \in \mathcal{Z}_\epsilon(f_1, f_2)$ for every $(f_1, f_2) \in [0,1]^2$. So, set $z = \widetilde{z}_\epsilon$. Then, Equation (14) holds for all $(f_1, f_2) \in [0,1]^2$ simultaneously. In other words, for all $(f_1, f_2) \in [0,1]^2$, we simultaneously have

$$H^\star(f_1, f_2) - \epsilon < H_z(f_1, f_2) < H^\star(f_1, f_2) + \epsilon, \tag{16}$$

i.e. H_z is in an ϵ-band around $H^\star$ throughout this region. And observe that this band gets smaller as ϵ is decreased (which is achieved at a larger value of z).

To bound the distance between $\widehat{\mathbf{v}}$, the minimizer of $H^\star$, and $\widehat{\mathbf{f}}(z)$, the minimizer of H_z, we bound the distance between the objective function values at these points.

Claim 8. *$H^\star(\widehat{\mathbf{f}}(z)) < H^\star(\widehat{\mathbf{v}}) + 2\epsilon$.*

Although $\widehat{\mathbf{f}}(z)$ does not minimize $H^\star$, Claim 8 says that the objective value at $\widehat{\mathbf{f}}(z)$ cannot be more than 2ϵ larger than its minimum, $H^\star(\widehat{\mathbf{v}})$. We use this to bound the distance between $\widehat{\mathbf{f}}(z)$ and the minimizer $\widehat{\mathbf{v}}$. Observe that $\widehat{\mathbf{f}}(z)$ falls in the $[H^\star(\widehat{\mathbf{v}}) + 2\epsilon]$-level set of $H^\star$. So, we next look at a specific level set of $H^\star$.

Define

$$\tau := \min_{\mathbf{f} \in [0,1]^2 : \|\mathbf{f} - \widehat{\mathbf{v}}\|_2 = \frac{\Delta}{4}} H^\star(\mathbf{f}). \tag{17}$$

Observe that a minimum exists (infimum is not required) for the minimization in (17) because we are minimizing over the closed set $\{\mathbf{f} \in [0,1]^2 : \|\mathbf{f} - \widehat{\mathbf{v}}\|_2 = \frac{\Delta}{4}\}$ and $H^\star$ is continuous.

For any fixed $p \in (1, \infty)$, Equation (13) shows that $\widehat{v}_1$ is bounded away from 0. Hence, Claim 4 shows that $H^\star$ is strictly convex at and in the region around $\widehat{\mathbf{v}}$. Further, $H^\star$ is convex everywhere else. Coupling this with the fact that (17) minimizes along points not arbitrarily close to the minimizer $\widehat{\mathbf{v}}$, we have $\tau > H^\star(\widehat{\mathbf{v}})$.

Define the level set of $H^\star$ with respect to τ:

$$\mathcal{C}_\tau = \{\mathbf{f} \in [0,1]^2 : H^\star(\mathbf{f}) \le \tau\}.$$

Claim 9. *For every* $\mathbf{f} \in \mathcal{C}_\tau$*, we have* $\|\mathbf{f} - \widehat{\mathbf{v}}\|_2 \le \frac{\Delta}{4}$.

Define $\epsilon_o := \frac{\tau - H^\star(\widehat{\mathbf{v}})}{2}$, and set $\epsilon = \epsilon_o$. Then, set $z = \widetilde{z}_{\epsilon_o}$ as before. Applying Claim 8, we obtain

$$H^\star(\widehat{\mathbf{f}}(\widetilde{z}_{\epsilon_o})) < H^\star(\widehat{\mathbf{v}}) + 2\epsilon_o = \tau.$$

In other words, $\widehat{\mathbf{f}}(\widetilde{z}_{\epsilon_o}) \in \mathcal{C}_\tau$. And applying Claim 9, we obtain $\|\widehat{\mathbf{f}}(\widetilde{z}_{\epsilon_o}) - \widehat{\mathbf{v}}\|_2 \le \frac{\Delta}{4}$, completing part (i).

This implies that $\|\widehat{\mathbf{f}}(\widetilde{z}_{\epsilon_o}) - \widehat{\mathbf{v}}\|_\infty \le \frac{\Delta}{4}$, which means

$$\left|\widehat{f}_1(\widetilde{z}_{\epsilon_o}) - \widehat{v}_1\right| \le \frac{\Delta}{4} \quad \text{and} \quad \left|\widehat{f}_2(\widetilde{z}_{\epsilon_o}) - \widehat{v}_2\right| \le \frac{\Delta}{4}. \tag{18}$$

Using these properties, we have

$$\begin{aligned}
\widehat{f}_1(\widetilde{z}_{\epsilon_o}) &\le \widehat{v}_1 + \frac{\Delta}{4} \\
&= \widehat{v}_2 - \Delta + \frac{\Delta}{4} \\
&\le \widehat{f}_2(\widetilde{z}_{\epsilon_o}) + \frac{\Delta}{4} - \Delta + \frac{\Delta}{4} = \widehat{f}_2(\widetilde{z}_{\epsilon_o}) - \frac{\Delta}{2},
\end{aligned}$$

where the first inequality holds because of the first part of (18), the equality holds because $\Delta = \widehat{v}_2 - \widehat{v}_1$ and the second inequality holds because of the second part of (18). Therefore, for $z = \widetilde{z}_{\epsilon_o} > 1$, the aggregate scores of the two papers are such that

$$\widehat{f}_1(\widetilde{z}_{\epsilon_o}) < \widehat{f}_2(\widetilde{z}_{\epsilon_o}),$$

violating efficiency.

Proof of Claim 1 Take arbitrary $\mathbf{f}, \mathbf{g} \in \mathbb{R}^2$ with $\mathbf{f} \neq \mathbf{g}$, and let $\theta \in (0,1)$. We show that $G_z(\theta\mathbf{f} + (1-\theta)\mathbf{g}) < \theta G_z(\mathbf{f}) + (1-\theta)G_z(\mathbf{g})$. For this, we will first show that either (i) the vector $[(z,0) - \mathbf{f}]$ is not parallel to the vector $[(z,0) - \mathbf{g}]$, (ii) the vector $[(0,1) - \mathbf{f}]$ is not parallel to the vector $[(0,1) - \mathbf{g}]$ or (iii) the vector $\mathbf{f}$ is not parallel to the vector $\mathbf{g}$. For the sake of contradiction, assume that this is not true. That is, assume $[(z,0) - \mathbf{f}]$ is parallel to $[(z,0) - \mathbf{g}]$, $[(0,1) - \mathbf{f}]$ is parallel to $[(0,1) - \mathbf{g}]$, and $\mathbf{f}$ is parallel to $\mathbf{g}$. This implies that

$$\begin{bmatrix} z - f_1 \\ -f_2 \end{bmatrix} = r \begin{bmatrix} z - g_1 \\ -g_2 \end{bmatrix}, \qquad \begin{bmatrix} -f_1 \\ 1 - f_2 \end{bmatrix} = s \begin{bmatrix} -g_1 \\ 1 - g_2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = t \begin{bmatrix} g_1 \\ g_2 \end{bmatrix},$$

for some $r, s, t \in \mathbb{R}$.[6] Note that, none of r, s, t can be 1 because $\mathbf{f} \neq \mathbf{g}$. The second equation tells us that $f_1 = sg_1$ and the third one tells us that $f_1 = tg_1$. So, either $f_1 = g_1 = 0$ or $s = t$. But from the first equation, $z - f_1 = rz - rg_1$. So if $f_1 = g_1 = 0$, it says that $r = 1$ which is not possible. Therefore, $s = t$. The third equation now tells us that $f_2 = tg_2 = sg_2$. But, the second equation gives us $1 - f_2 = s - sg_2$, which implies that $s = 1$. But again this is not possible, leading to a contradiction. Therefore, at least one of (i), (ii) and (iii) is true.

6. A boundary case not captured here is when $\mathbf{g}$ is exactly one of the points $(z,0), (0,1)$ or $(0,0)$, leading to $1/r, 1/s$ or $1/t$ being zero respectively. But for this case, it is easy to prove that the other two pairs of vectors cannot be parallel unless $\mathbf{f} = \mathbf{g}$.

L_p norm with $p \in (1, \infty)$ is a convex norm, i.e. for any $x, y \in \mathbb{R}^2$,

$$\|\theta x + (1-\theta)y\|_p \leq \theta\|x\|_p + (1-\theta)\|y\|_p. \tag{19}$$

Further, since $p \in (1, \infty)$, the inequality in (19) is strict if x is not parallel to y. For our objective (in Equation (9)),

$$\begin{aligned} G_z(\theta\mathbf{f} + (1-\theta)\mathbf{g}) = &\left\|\theta[(z,0) - \mathbf{f}] + (1-\theta)[(z,0) - \mathbf{g}]\right\|_p \\ &+ \left\|\theta[(0,1) - \mathbf{f}] + (1-\theta)[(0,1) - \mathbf{g}]\right\|_p \\ &+ \left\|\theta\mathbf{f} + (1-\theta)\mathbf{g}\right\|_p. \end{aligned} \tag{20}$$

Because of convexity of the L_p norm, each of the three terms on the RHS of Equation (20) satisfies inequality (19). Further, because at least one of the pair of vectors in the three terms is not parallel (since either (i), (ii) or (iii) is true), at least one of them gives us a strict inequality. Therefore we obtain

$$G_z(\theta\mathbf{f} + (1-\theta)\mathbf{g}) < \theta G_z(\mathbf{f}) + (1-\theta)G_z(\mathbf{g}).$$

■

Proof of Claim 2 The claim has four parts: (i) $\widehat{f}_1(z) \geq 0$, (ii) $\widehat{f}_1(z) \leq 1$, (iii) $\widehat{f}_2(z) \geq 0$ and (iv) $\widehat{f}_2(z) \leq 1$. Observe that parts (i), (iii) and (iv) are more intuitive, since they show that the aggregate score of a paper is no higher than the maximum score given to it, and no lower than the minimum score given to it. Part (ii) on the other hand is stronger; even though paper 1 has a score of $z > 1$ given to it, this part shows that $\widehat{f}_1(z) \leq 1$ (which is much tighter than an upper bound of z, especially when z is large). We prove the simpler parts (i), (iii) and (iv) first.

For the sake of contradiction, suppose $\widehat{f}_1(z) < 0$. Then

$$\begin{aligned} G_z(\widehat{f}_1(z), \widehat{f}_2(z)) &= \left[|z - \widehat{f}_1(z)|^p + |\widehat{f}_2(z)|^p\right]^{\frac{1}{p}} + \left[|\widehat{f}_1(z)|^p + |1 - \widehat{f}_2(z)|^p\right]^{\frac{1}{p}} + \left[|\widehat{f}_1(z)|^p + |\widehat{f}_2(z)|^p\right]^{\frac{1}{p}} \\ &> \left[|z|^p + |\widehat{f}_2(z)|^p\right]^{\frac{1}{p}} + \left[0 + |1 - \widehat{f}_2(z)|^p\right]^{\frac{1}{p}} + \left[0 + |\widehat{f}_2(z)|^p\right]^{\frac{1}{p}} = G_z(0, \widehat{f}_2(z)), \end{aligned}$$

contradicting the fact that $(\widehat{f}_1(z), \widehat{f}_2(z))$ is optimal. Therefore, $\widehat{f}_1(z) \geq 0$, completing proof of (i). Similarly, if $\widehat{f}_2(z) < 0$, we can show that $G_z(\widehat{f}_1(z), \widehat{f}_2(z)) > G_z(\widehat{f}_1(z), 0)$, violating optimality. Therefore, $\widehat{f}_2(z) \geq 0$, completing proof of (iii).

Next, for the sake of contradiction assume that $\widehat{f}_2(z) > 1$. Then

$$\begin{aligned} G_z(\widehat{f}_1(z), \widehat{f}_2(z)) &= \left[|z - \widehat{f}_1(z)|^p + |\widehat{f}_2(z)|^p\right]^{\frac{1}{p}} + \left[|\widehat{f}_1(z)|^p + |1 - \widehat{f}_2(z)|^p\right]^{\frac{1}{p}} + \left[|\widehat{f}_1(z)|^p + |\widehat{f}_2(z)|^p\right]^{\frac{1}{p}} \\ &> \left[|z - \widehat{f}_1(z)|^p + 1\right]^{\frac{1}{p}} + \left[|\widehat{f}_1(z)|^p + 0\right]^{\frac{1}{p}} + \left[|\widehat{f}_1(z)|^p + 1\right]^{\frac{1}{p}} = G_z(\widehat{f}_1(z), 1), \end{aligned}$$

contradicting the fact that $(\widehat{f}_1(z), \widehat{f}_2(z))$ is optimal. Therefore, we also have $\widehat{f}_2(z) \leq 1$, completing proof of (iv).

Finally, we prove the more non-intuitive part, (ii). Suppose for the sake of contradiction, $\widehat{f}_1(z) > 1$. Then,

$$\begin{aligned} G_z(\widehat{f}_1(z), \widehat{f}_2(z)) &= \left[|z - \widehat{f}_1(z)|^p + |\widehat{f}_2(z)|^p\right]^{\frac{1}{p}} + \left[|\widehat{f}_1(z)|^p + |1 - \widehat{f}_2(z)|^p\right]^{\frac{1}{p}} + \left[|\widehat{f}_1(z)|^p + |\widehat{f}_2(z)|^p\right]^{\frac{1}{p}} \\ &\geq |z - \widehat{f}_1(z)| + |\widehat{f}_1(z)| + |\widehat{f}_1(z)| \\ &\geq z + |\widehat{f}_1(z)|, \end{aligned}$$

where the first inequality comes from the fact that the L_p norm of each vector is at least as high as the absolute value of its first element, and the second inequality follows from the triangle inequality. Using the assumption that $\widehat{f}_1(z) > 1$, we obtain

$$G_z(\widehat{f}_1(z), \widehat{f}_2(z)) > z + 1 = G_z(0, 0),$$

contradicting the fact that $(\widehat{f}_1(z), \widehat{f}_2(z))$ is optimal. Therefore, $\widehat{f}_1(z) \leq 1$, completing the proof. ■

Proof of Claim 3 Take any arbitrary $f_1 \in [0, 1]$ and $f_2 \in [0, 1]$. Subtracting Equations (11) and (12) we obtain

$$H_z(f_1, f_2) - H^\star(f_1, f_2) = \left[\big(z - f_1\big)^p + f_2^p\right]^{\frac{1}{p}} - \big(z - f_1\big). \tag{21}$$

Observe that since $f_2 \geq 0$, the RHS of Equation (21) is non-negative. Hence, the equation does not change on using an absolute value, i.e.,

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| = \left[\big(z - f_1\big)^p + f_2^p\right]^{\frac{1}{p}} - \big(z - f_1\big). \tag{22}$$

To prove the required result, we take a small detour and define $\phi(x) = (x^p + f_2^p)^{\frac{1}{p}} - x$. We show that $\phi(x) \to 0$ as $x \to \infty$. For this, rewrite $\phi(x)$ as follows

$$\phi(x) = x\left(1 + \frac{f_2^p}{x^p}\right)^{\frac{1}{p}} - x = \frac{\left(1 + \frac{f_2^p}{x^p}\right)^{\frac{1}{p}} - 1}{\frac{1}{x}}.$$

Taking the limit of x to infinity, we have

$$\lim_{x\to\infty} \phi(x) = \lim_{x\to\infty} \frac{\left(1 + \frac{f_2^p}{x^p}\right)^{\frac{1}{p}} - 1}{\frac{1}{x}}. \tag{23}$$

Observe that for both the numerator and denominator in the RHS of Equation (23), we have

$$\lim_{x\to\infty} \left\{\left(1 + \frac{f_2^p}{x^p}\right)^{\frac{1}{p}} - 1\right\} = 0 \quad \text{and} \quad \lim_{x\to\infty} \left\{\frac{1}{x}\right\} = 0.$$

Hence, applying L'Hospital's rule on equation (23) gives us

$$\begin{aligned}\lim_{x\to\infty}\phi(x) &= \lim_{x\to\infty}\frac{-\frac{f_2^p}{x^{p+1}}\left(1+\frac{f_2^p}{x^p}\right)^{\frac{1}{p}-1}}{-\frac{1}{x^2}}\\ &= \lim_{x\to\infty}\left\{\frac{f_2^p}{x^{p-1}}\left(1+\frac{f_2^p}{x^p}\right)^{\frac{1}{p}-1}\right\}\\ &= \left[\lim_{x\to\infty}\frac{f_2^p}{x^{p-1}}\right] * \left[\lim_{x\to\infty}\left(1+\frac{f_2^p}{x^p}\right)^{\frac{1}{p}-1}\right]\\ &= 0 * 1 = 0,\end{aligned}$$

where $\left[\lim_{x\to\infty}\frac{f_2^p}{x^{p-1}}\right] = 0$ because $p > 1$. Hence, we proved the required result, $\lim_{x\to\infty}\phi(x) = 0$. Going back to Equation (22), we rewrite it as

$$|H_z(f_1,f_2) - H^\star(f_1,f_2)| = \left[\big(z-f_1\big)^p + f_2^p\right]^{\frac{1}{p}} - \big(z-f_1\big) = \phi(z-f_1).$$

Taking the limit of z to infinity, we obtain

$$\lim_{z\to\infty}|H_z(f_1,f_2) - H^\star(f_1,f_2)| = \lim_{z\to\infty}\phi(z-f_1) = \lim_{t\to\infty}\phi(t) = 0, \tag{24}$$

where the second step follows by setting $t = z - f_1$. Equation (24) implies that

$$\lim_{z\to\infty} H_z(f_1,f_2) = H^\star(f_1,f_2).$$

■

Proof of Claim 4 In the region $[0,1]^2$, using (12), the function $H^\star$ can be written as

$$H^\star(f_1,f_2) = -f_1 + \|(0,1) - (f_1,f_2)\|_p + \|(f_1,f_2)\|_p - 1. \tag{25}$$

Observe that each term on the RHS of (25) is a convex function of $\mathbf{f}$. Hence, their sum is also convex in $\mathbf{f}$.

The proof of strict convexity closely follows the proof of claim 1. Take arbitrary $\mathbf{f}, \mathbf{g} \in (0,1]\times[0,1]$ with $\mathbf{f} \neq \mathbf{g}$, and let $\theta \in (0,1)$. We show that $H^\star(\theta\mathbf{f} + (1-\theta)\mathbf{g}) < \theta H^\star(\mathbf{f}) + (1-\theta)H^\star(\mathbf{g})$. For this, we will first show that either (i) $[(0,1) - \mathbf{f}]$ is not parallel to $[(0,1) - \mathbf{g}]$ or (ii) $\mathbf{f}$ is not parallel to $\mathbf{g}$. For the sake of contradiction, assume that this is not true. That is, assume $[(0,1) - \mathbf{f}]$ is parallel to $[(0,1) - \mathbf{g}]$, and $\mathbf{f}$ is parallel to $\mathbf{g}$. This implies that

$$\begin{bmatrix}-f_1\\1-f_2\end{bmatrix} = r\begin{bmatrix}-g_1\\1-g_2\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}f_1\\f_2\end{bmatrix} = s\begin{bmatrix}g_1\\g_2\end{bmatrix},$$

where $r, s \in \mathbb{R}$. Note that, neither r nor s can be 1 because $\mathbf{f} \neq \mathbf{g}$. The first equation tells us that $f_1 = rg_1$ and the second one tells us that $f_1 = sg_1$. And since $g_1 \neq 0$, this implies that $r = s$. The second part of the second equation now tells us that $f_2 = sg_2 = rg_2$. The second part of the first equation becomes $1 - f_2 = r - rg_2$ which implies that $r = 1$, leading to a contradiction. Therefore, at least one of (i) and (ii) is true.

Recall, L_p norm with $p \in (1, \infty)$ is a convex norm, i.e. for any $x, y \in \mathbb{R}^2$,

$$\|\theta x + (1-\theta)y\|_p \leq \theta\|x\|_p + (1-\theta)\|y\|_p. \tag{26}$$

And since $p \in (1, \infty)$, the inequality in (26) is strict if x is not parallel to y. For $H^\star$ (using Equation (25)),

$$\begin{aligned} H^\star(\theta\mathbf{f} + (1-\theta)\mathbf{g}) = &-\theta f_1 - (1-\theta)g_1 \\ &+ \big\|\theta[(0,1) - \mathbf{f}] + (1-\theta)[(0,1) - \mathbf{g}]\big\|_p \\ &+ \big\|\theta\mathbf{f} + (1-\theta)\mathbf{g}\big\|_p - 1. \end{aligned} \tag{27}$$

Because of convexity of the L_p norm, both the third and fourth term on the RHS of Equation (27) satisfy inequality (26). Further, because at least one of the pair of vectors in these two terms is not parallel (since either (i) or (ii) is true), at least one of them gives us a strict inequality. Therefore we obtain

$$H^\star(\theta\mathbf{f} + (1-\theta)\mathbf{g}) < \theta H^\star(\mathbf{f}) + (1-\theta)H^\star(\mathbf{g}).$$

■

Proof of Claim 5 To compute the minimizer of $H^\star$, we compute its gradients with respect to f_1 and f_2. Using Equation (12), the partial derivative with respect to f_1 is

$$\frac{\partial H^\star}{\partial f_1} = -1 + f_1^{p-1}\Big[f_1^p + (1-f_2)^p\Big]^{\frac{1}{p}-1} + f_1^{p-1}\Big[f_1^p + f_2^p\Big]^{\frac{1}{p}-1} \tag{28}$$

and with respect to f_2 is

$$\frac{\partial H^\star}{\partial f_2} = 0 - (1-f_2)^{p-1}\Big[f_1^p + (1-f_2)^p\Big]^{\frac{1}{p}-1} + f_2^{p-1}\Big[f_1^p + f_2^p\Big]^{\frac{1}{p}-1}. \tag{29}$$

Observe that at $f_2 = \frac{1}{2}$, irrespective of the value of f_1, the partial derivative (29) is

$$\left.\frac{\partial H^\star}{\partial f_2}\right|_{f_2=\frac{1}{2}} = -\frac{1}{2^{p-1}}\left[f_1^p + \frac{1}{2^p}\right]^{\frac{1}{p}-1} + \frac{1}{2^{p-1}}\left[f_1^p + \frac{1}{2^p}\right]^{\frac{1}{p}-1} = 0.$$

So, set $\widehat{v}_2 = \frac{1}{2}$. Next, we find $\widehat{v}_1$ such that the other derivative (28) is also zero at $\widehat{\mathbf{v}} = (\widehat{v}_1, \widehat{v}_2)$. Setting (28) to zero at $\widehat{\mathbf{v}}$, we obtain

$$\begin{aligned}
\left.\frac{\partial H^\star}{\partial f_1}\right|_{\mathbf{f}=\widehat{\mathbf{v}}} = 0 &= -1 + \widehat{v}_1^{p-1}\left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{\frac{1}{p}-1} + \widehat{v}_1^{p-1}\left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{\frac{1}{p}-1} \\
&\implies 1 = 2\widehat{v}_1^{p-1}\left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{\frac{1}{p}-1} \\
\implies \left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{1-\frac{1}{p}} &= 2\widehat{v}_1^{p-1} \\
\implies \left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{p-1} &= 2^p \widehat{v}_1^{p(p-1)} \\
\implies \widehat{v}_1^p + \frac{1}{2^p} &= 2^{\frac{p}{p-1}}\widehat{v}_1^p \\
\implies \frac{1}{2^p} &= \widehat{v}_1^p\left(2^{\frac{p}{p-1}} - 1\right) \\
\therefore \widehat{v}_1 &= \frac{1}{2}\left[\frac{1}{(2^{\frac{p}{p-1}} - 1)}\right]^{\frac{1}{p}}.
\end{aligned}$$

Hence, $\nabla_{\mathbf{f}} H^\star(\mathbf{f}) = \mathbf{0}$ at $\widehat{\mathbf{v}}$. And since $H^\star$ is convex in $[0,1]^2$ by Claim 4, $\widehat{\mathbf{v}}$ is the minimizer in this region. ■

Proof of Claim 6 For any $p > 1$, we know

$$\frac{p}{p-1} > 1.$$

This implies that

$$2^{\frac{p}{p-1}} - 1 > 1 \quad \text{and hence} \quad \left[\frac{1}{2^{\frac{p}{p-1}} - 1}\right]^{\frac{1}{p}} < 1.$$

Finally, using the values from Claim 5, we obtain

$$\widehat{v}_1 < \widehat{v}_2.$$

■

Proof of Claim 7 Let $z \in \mathcal{Z}_\epsilon(1,1)$. Pick an arbitrary $(f_1, f_2) \in [0,1]^2$. As in the proof of Claim 3, on subtracting Equations (11) and (12), and taking an absolute value, we obtain Equation (22), that is,

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| = \left[\big(z - f_1\big)^p + f_2^p\right]^{\frac{1}{p}} - \big(z - f_1\big). \tag{30}$$

Combining Equation (30) with the fact that $0 \le f_2 \le 1$, we obtain

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| \le \left[\big(z - f_1\big)^p + 1\right]^{\frac{1}{p}} - \big(z - f_1\big). \tag{31}$$

Now, define $\psi(x) = (x^p + 1)^{\frac{1}{p}} - x$. We show that $\psi(x)$ is a non-increasing function for $x \geq 0$. Computing the derivative, we have

$$\frac{d\psi(x)}{dx} = x^{p-1}\left(x^p + 1\right)^{\frac{1}{p}-1} - 1 = \left(\frac{x^p}{x^p+1}\right)^{\frac{p-1}{p}} - 1 \leq 0$$

for $x \geq 0$, showing that it is a non-increasing function. Going back to Equation (31), we know that $f_1 \leq 1$. Therefore, $\big(z - f_1\big) \geq \big(z - 1\big) \geq 0$. Using the fact that ψ is a non-increasing function, we obtain $\psi\big(z - f_1\big) \leq \psi\big(z - 1\big)$, which on expansion gives us

$$\Big[\big(z - f_1\big)^p + 1\Big]^{\frac{1}{p}} - \big(z - f_1\big) \leq \Big[\big(z-1\big)^p + 1\Big]^{\frac{1}{p}} - \big(z - 1\big) = |H_z(1,1) - H^\star(1,1)|. \quad (32)$$

Combining Equations (31) and (32), and the fact that $z \in \mathcal{Z}_\epsilon(1,1)$, we obtain

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| \leq |H_z(1,1) - H^\star(1,1)| < \epsilon.$$

Hence, $z \in \mathcal{Z}_\epsilon(f_1, f_2)$. ■

Proof of Claim 8 The proof follows using three facts:

1. Equation (16) for $\widehat{\mathbf{f}}(z)$ says that $H^\star(\widehat{\mathbf{f}}(z)) < H_z(\widehat{\mathbf{f}}(z)) + \epsilon$.
2. Because $\widehat{\mathbf{f}}(z)$ is the minimizer of H_z, we have $H_z(\widehat{\mathbf{f}}(z)) \leq H_z(\widehat{\mathbf{v}})$.
3. For $\widehat{\mathbf{v}}$, Equation (16) gives us $H_z(\widehat{\mathbf{v}}) < H^\star(\widehat{\mathbf{v}}) + \epsilon$.

Putting these equations together:

$$H^\star(\widehat{\mathbf{f}}(z)) < H_z(\widehat{\mathbf{f}}(z)) + \epsilon \leq H_z(\widehat{\mathbf{v}}) + \epsilon < H^\star(\widehat{\mathbf{v}}) + 2\epsilon.$$

■

Proof of Claim 9 We prove the claim by contraposition. Pick an arbitrary $\mathbf{f} \in [0,1]^2$ such that $\|\mathbf{f} - \widehat{\mathbf{v}}\|_2 > \frac{\Delta}{4}$. This means that there exists $\mathbf{g} \in [0,1]^2$ on the line joining $\mathbf{f}$ and $\widehat{\mathbf{v}}$ such that $\|\mathbf{g} - \widehat{\mathbf{v}}\|_2 = \frac{\Delta}{4}$. We could alternatively write $\mathbf{g} = \theta\mathbf{f} + (1-\theta)\widehat{\mathbf{v}}$, where $\theta \in (0,1)$. By convexity of $H^\star$,

$$H^\star(\mathbf{g}) \leq \theta H^\star(\mathbf{f}) + (1-\theta)H^\star(\widehat{\mathbf{v}}). \quad (33)$$

By definition of τ in (17), we know $H^\star(\mathbf{g}) \geq \tau$. Also, we know $H^\star(\widehat{\mathbf{v}}) < \tau$. Using these in (33), we obtain

$$\tau < \theta H^\star(\mathbf{f}) + (1-\theta)\tau.$$

Therefore, we obtain $H^\star(\mathbf{f}) > \tau$. In summary, if $\|\mathbf{f} - \widehat{\mathbf{v}}\|_2 > \frac{\Delta}{4}$, then $H^\star(\mathbf{f}) > \tau$. Taking the contrapositive gives us the desired result. ■

A.2 Proof of Lemma 3

Consider $L(p, q)$ aggregation with arbitrary $q \in (1, \infty]$. We show that strategyproofness is violated. The construction for this is as follows. Suppose there is one paper a and two reviewers. The first reviewer gives the paper an overall recommendation of 1 and the second reviewer gives it an overall recommendation of 0. Let $\mathbf{x}_a$ be the (objective) criteria scores of this paper.

Let us first consider $q \in (1, \infty)$. For a function $f : \mathbb{X} \to \mathbb{Y}$, all we care about in this example is its value at $\mathbf{x}_a$. Hence, for simplicity, let f_a denote the value of function f at $\mathbf{x}_a$, i.e, $f_a := f(\mathbf{x}_a)$. Then our aggregation becomes

$$\widehat{f_a} = \operatorname*{argmin}_{f_a \in \mathbb{R}} \Big\{ |1 - f_a|^q + |f_a|^q \Big\}.$$

We claim that $f_a = 0.5$ is the unique minimizer. Observe that if $f_a = 0.5$, then the value of our objective is $0.5^q + 0.5^q < 1$ when $q \in (1, \infty)$. On the other hand, if $f_a \geq 1$ or if $f_a \leq 0$ then the value of our objective is at least 1. Hence $f_a \in (0, 1)$. By symmetry, we can restrict attention to the range $[0.5, 1)$ since if there is a minimizer in $(0, 0.5)$ then there must also be a minimizer in $(0.5, 1)$. Consequently, we rewrite the optimization problem as

$$\widehat{f_a} = \operatorname*{argmin}_{f_a \in [0.5, 1)} \Big\{ (1 - f_a)^q + f_a^q \Big\}. \tag{34}$$

Consider the function $h : [0.5, 1] \to \mathbb{R}$ defined by $h(x) = x^q$. This function is strictly convex (the second derivative is strictly positive in the domain) whenever $q \in (1, \infty)$. Hence from the definition of strict convexity, we have

$$0.5\big((1 - f_a)^q + f_a^q\big) > \big(0.5(1 - f_a + f_a)\big)^q = 0.5^q$$

whenever $f_a \in (0.5, 1)$. Consequently, the objective value of (34) is greater at $f_a \in (0.5, 1)$ than at $f_a = 0.5$. We conclude that $\widehat{f_a} = 0.5$ whenever $q \in (1, \infty)$.

When $q = \infty$, we equivalently write the optimization problem as

$$\widehat{f_a} = \operatorname*{argmin}_{f_a \in \mathbb{R}} \max\big(|1 - f_a|, |f_a|\big).$$

This objective has a value of 0.5 if $f_a = 0.5$ and strictly greater if $f_a \neq 0.5$. Hence, $\widehat{f_a} = 0.5$ for $q = \infty$ as well.

The true overall recommendation of reviewer 1 differs from the aggregate $\widehat{f_a}$ by 0.5 (in every L_ℓ norm). However, if reviewer 1 reported an overall recommendation of 2, then an argument identical to that above shows that the minimizer is $\widehat{g}_a = 1$. Reviewer 1 has thus successfully brought down the difference between her own true overall recommendation and the aggregate $\widehat{g}_a$ to 0. We conclude that strategyproofness is violated whenever $q \in (1, \infty]$. ■

A.3 Proof of Lemma 4

The construction showing that $L(\infty, 1)$ aggregation violates consensus is as follows. Suppose there are two papers, two reviewers and both reviewers review both papers. Assume that the

papers have objective criteria scores $\mathbf{x}_1$ and $\mathbf{x}_2$, and that neither of these scores is pointwise greater than or equal to the other. Let the overall recommendations of the reviewers for the papers be given by the matrix

$$\mathbf{y} = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix},$$

where y_{ia} denotes the overall recommendation of reviewer i for paper a. Since both reviewers give the same overall recommendation of 1 to paper 2, any aggregation method that satisfies consensus must also give paper 2 an aggregate score of 1. We show that this is not the case under $L(\infty, 1)$ aggregation.

Let f_i denote the value of function f on paper i, i.e. $f_i := f(\mathbf{x}_i)$. And let $\widehat{f}_i$ denote the aggregate score of paper i. Since we are minimizing $L(\infty, 1)$ loss, the aggregate function satisfies:

$$(\widehat{f}_1, \widehat{f}_2) \in \operatorname*{argmin}_{(f_1,f_2)\in\mathbb{R}^2} \Big\{ \big\|(0,1) - (f_1, f_2)\big\|_\infty + \big\|(2,1) - (f_1, f_2)\big\|_\infty \Big\}. \tag{35}$$

We do not have any monotonicity constraints in (35) because the two papers have incomparable criteria scores. Denote the objective function of (35) by $G(f_1, f_2)$. We can simplify this objective to

$$G(f_1, f_2) = \max(|f_1|, |f_2 - 1|) + \max(|2 - f_1|, |f_2 - 1|). \tag{36}$$

We claim that $(0.5, 0.5)$ is a minimizer of G. The objective function value at this point is

$$G(0.5, 0.5) = \max(0.5, 0.5) + \max(1.5, 0.5) = 0.5 + 1.5 = 2.$$

For arbitrary $(f_1, f_2) \in \mathbb{R}^2$, we have

$$\begin{aligned} G(f_1, f_2) &= \max(|f_1|, |f_2 - 1|) + \max(|2 - f_1|, |f_2 - 1|) \\ &\geq |f_1| + |2 - f_1| \\ &\geq 2 = G(0.5, 0.5), \end{aligned}$$

where the first inequality holds because the maximum of two elements is always larger than the first, and the second inequality holds by the triangle inequality. Therefore, $(0.5, 0.5)$ is a minimizer of G. The L_2 norm of this minimizer is $0.5\sqrt{2} < 1$. On the other hand, any minimizer $(\widehat{f}_1, \widehat{f}_2)$ with $\widehat{f}_2 = 1$ would have an L_2 norm of at least 1. It follows that such a minimizer will not be selected. In other words, $L(\infty, 1)$ aggregation would select a minimizer for which the aggregate score of paper 2 is not 1, violating consensus.[7] ■

Complete picture of minimizers For completeness, we look at the set of all minimizers of G. This is given by

$$\widehat{F} = \big\{(f_1, f_2) \mid f_1 \in [0, 2], f_2 \in [1 - \min(f_1, 2 - f_1), 1 + \min(f_1, 2 - f_1)]\big\}.$$

7. Observe that even if we used any L_k norm with $k \in (1, \infty)$ for tie-breaking, the L_k norm of $(0.5, 0.5)$ would be $0.5\sqrt[k]{2} < 1$, while the L_k norm of any minimizer $(\widehat{f}_1, 1)$ would still be at least 1, violating consensus.

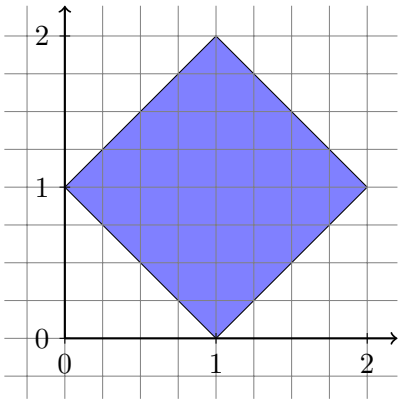


Figure 4: The shaded region depicts the set of all minimizers of (35). f_1 is on the x-axis and f_2 is on the y-axis.

Pictorially, this set is given by the shaded square in Figure 4. It is the square with vertices at $(0, 1)$, $(1, 0)$, $(2, 1)$ and $(1, 2)$.

This shows that almost all minimizers violate consensus. For the specific tie-breaking considered, the minimizer chosen is the one with minimum L_2 norm, i.e., the projection of $(0, 0)$ onto this square. This gives us $(0.5, 0.5)$, violating consensus.

Observe that tie-breaking using minimum L_k norm, for $k \in (1, \infty]$, also chooses $(0.5, 0.5)$ as the aggregate function, violating consensus. For $k = 1$, all points on the line segment $f_1 + f_2 = 1$ $(0 \leq f_1 \leq 1)$ would be tied winners, almost all of which violate consensus. Further, even if one uses other reasonable tie-breaking schemes like maximum L_k norm, they suffer from the same issue, i.e., there is a tied winner which violates consensus.

Appendix B. Additional Empirical Results

We present some more empirical results in addition to those provided in the main text.

B.1 Influence of Varying the Hyperparameters

Although our theoretical results identify $L(1, 1)$ aggregation as the most desirable, we would like to paint a broader picture by determining how much impact the choice of p and q actually has on selected papers. To this end, we compute the overlap between the papers selected by $L(p, q)$ aggregation, for $p, q \in \{1, 2, 3\}$ (although in general p and q need not be integral, they can be real as well as ∞). Table 2 shows the overlap between papers selected by $L(p_1, q_1)$ and $L(p_2, q_2)$, where the rows represent (p_1, q_1) and columns represent (p_2, q_2). Note that the table is symmetric. The results suggest that q has a more significant impact than p on $L(p, q)$ aggregation. For instance, $L(1, 1)$ behaves more similarly to $L(2, 1)$ and $L(3, 1)$ than to $L(1, 2)$ and $L(1, 3)$.

	1,1	**1,2**	**1,3**	**2,1**	**2,2**	**2,3**	**3,1**	**3,2**	**3,3**
1,1	100.0	87.5	82.7	96.1	88.0	82.6	92.3	87.5	82.1
1,2	87.5	100.0	94.5	88.3	94.9	93.1	87.7	94.6	92.3
1,3	82.7	94.5	100.0	84.0	92.1	95.2	83.5	91.8	94.0
2,1	96.1	88.3	84.0	100.0	89.8	84.4	95.7	89.5	84.0
2,2	88.0	94.9	92.1	89.8	100.0	94.1	89.8	98.8	93.7
2,3	82.6	93.1	95.2	84.4	94.1	100.0	84.4	94.1	98.6
3,1	92.3	87.7	83.5	95.7	89.8	84.4	100.0	89.7	84.0
3,2	87.5	94.6	91.8	89.5	98.8	94.1	89.7	100.0	93.8
3,3	82.1	92.3	94.0	84.0	93.7	98.6	84.0	93.8	100.0

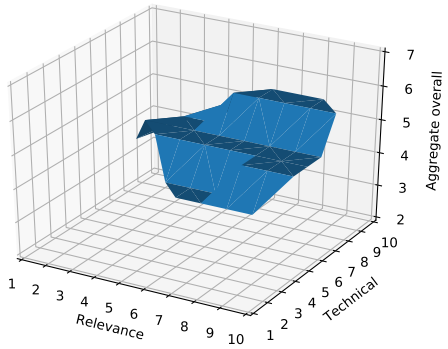
Table 2: Percentage of overlap (in selected papers) between different $L(p, q)$ aggregation methods

B.2 Visualizing the Community Aggregate Mapping

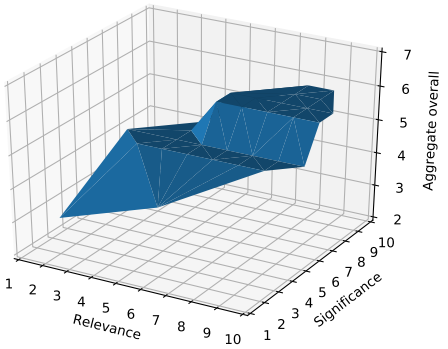
Our framework is not only useful for computing an aggregate mapping to help in acceptance decisions, but also for understanding the preferences of the community for use in subsequent modeling and research. We illustrate this application by providing some visualizations and interpretations of the aggregate function $\widetilde{f}$ obtained from $L(1, 1)$ aggregation on the IJCAI review data.

The function $\widetilde{f}$ lives in a 5-dimensional space, making it hard to visualize the entire aggregate function. Instead, we fix the values of 3 criteria at a time and plot the function in terms of the remaining two criteria. In all of the visualizations below, the fixed criteria are set to their respective (marginal) modes: For 'quality of writing' the mode is 7 (715 reviews), for 'originality' it is 6 (826 reviews), for 'relevance' it is 8 (888 reviews), for 'significance' it is 5 (800 reviews), and for 'technical quality' it is 6 (702 reviews).

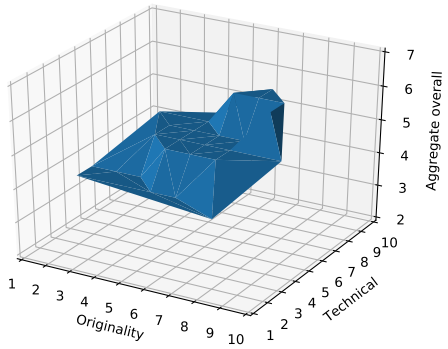
The key takeaways from this experiment are as follows. First, writing and relevance do not have a significant influence (Figure 5(e)). Really bad writing or relevance is a significant downside, excellent writing or relevance is appreciated, but everything else in between in irrelevant. Second, technical quality and significance exert a high influence (Figure 5(f)). Moreover, the influence is approximately linear. Third, linear models (i.e., models that are linear in the criteria) are quite popular in machine learning, and our empirical observations reveal that linear models are partially applicable for the mapping — for some criteria one may indeed assume a linear model, but not for all.



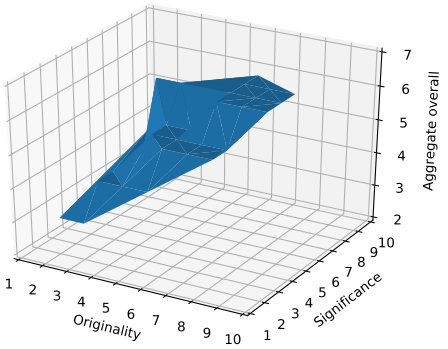
(a) Varying 'relevance' and 'technical quality'



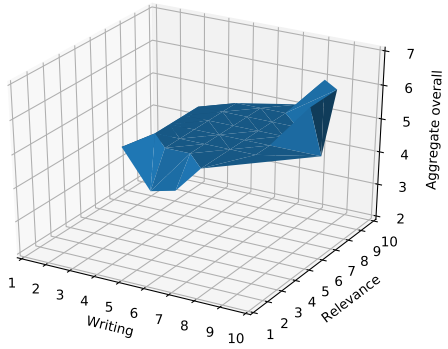
(b) Varying 'relevance' and 'significance'



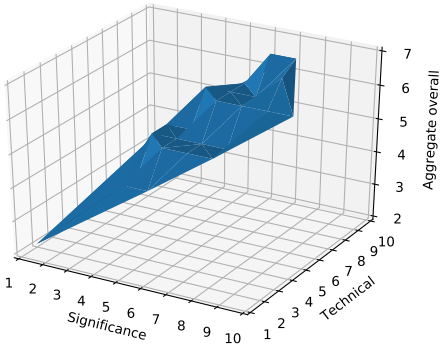
(c) Varying 'originality' and 'technical quality'



(d) Varying 'originality' and 'significance'

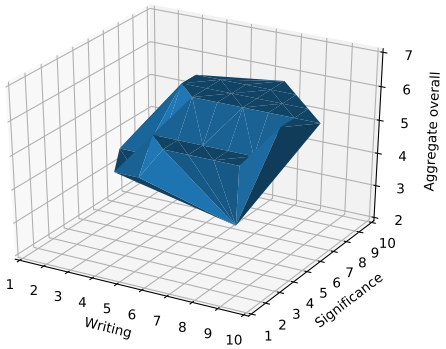


(e) Varying 'quality of writing' and 'relevance'

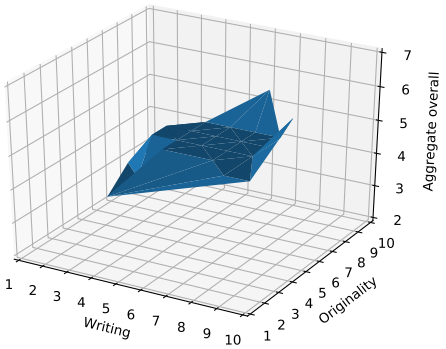


(f) Varying 'significance' and 'technical quality'

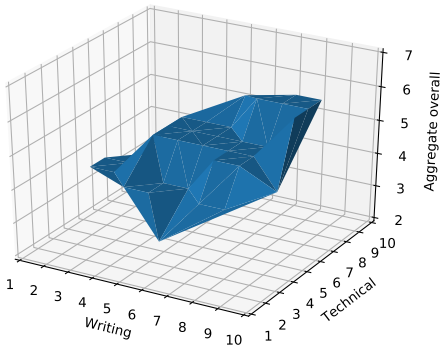
Figure 5: Impact of varying different criteria under $L(1, 1)$ aggregation



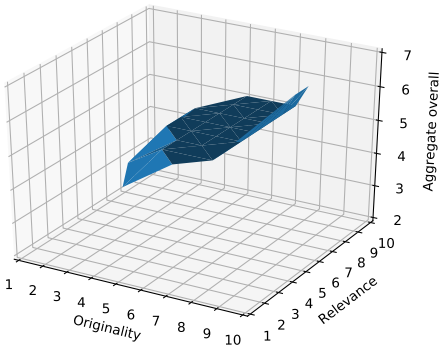
(a) Varying 'quality of writing' and 'significance'



(b) Varying 'quality of writing' and 'originality'



(c) Varying 'quality of writing' and 'technical quality'



(d) Varying 'originality' and 'relevance'

Figure 6: Impact of varying different criteria under $L(1,1)$ aggregation (continued)

References

Alon, N., Fischer, F., Procaccia, A., and Tennenholtz, M. (2011). Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 101–110. ACM.

Arrow, K. (1951). *Social Choice and Individual Values*. Wiley.

Aziz, H., Lev, O., Mattei, N., Rosenschein, J. S., and Walsh, T. (2019). Strategyproof peer selection using randomization, partitioning, and apportionment. *Artificial Intelligence*.

Bakanic, V., McPhail, C., and Simon, R. J. (1987). The manuscript review and decision-making process. *American Sociological Review*, 52(5):631–642.

Balietti, S., Goldstone, R. L., and Helbing, D. (2016). Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 113(30):8414–8419.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors (2016). *Handbook of Computational Social Choice*. Cambridge University Press.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ-1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Chatterjee, S., Guntuboyina, A., and Sen, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli*, 24(2):1072–1100.

Church, K. (2005). Reviewing the reviewers. *Computational Linguistics*, 31(4):575–578.

Ding, C., Zhou, D., He, X., and Zha, H. (2006). R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 281–288.

Ding, W., Shah, N. B., and Wang, W. (2020). On the privacy-utility tradeoff in peer-review data analysis. In *AAAI Privacy-Preserving Artificial Intelligence (PPAI-21) workshop*.

Gao, F. and Wellner, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764.

Ge, H., Welling, M., and Ghahramani, Z. (2013). A Bayesian model for calibrating conference review scores. Manuscript. Available online `http://mlg.eng.cam.ac.uk/hong/unpublished/nips-review-model.pdf` Last accessed: April 4, 2021.

Hojat, M., Gonnella, J. S., and Caelleigh, A. S. (2003). Impartial judgment by the "gatekeepers" of science: Fallibility and accountability in the peer review process. *Advances in Health Sciences Education*, 8(1):75–96.

Hua, X., Nikolov, M., Badugu, N., and Wang, L. (2019). Argument mining for understanding peer reviews. *arXiv preprint arXiv:1903.10104*.

Jecmen, S., Zhang, H., Liu, R., Shah, N. B., Conitzer, V., and Fang, F. (2020). Mitigating manipulation in peer review via randomized reviewer assignments. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Kahng, A., Kotturi, Y., Kulkarni, C., Kurokawa, D., and Procaccia, A. (2018). Ranking wily people who rank each other. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Kashlak, A. B. and Kong, L. (2021). Nonasymptotic support recovery for high-dimensional sparse covariance matrices. *Stat*, 10(1):e316.

Kerr, S., Tolliver, J., and Petree, D. (1977). Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal*, 20(1):132–141.

Kong, D., Ding, C., and Huang, H. (2011). Robust nonnegative matrix factorization using L21-norm. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*.

Kowalski, M. (2009). Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324.

Kurokawa, D., Lev, O., Morgenstern, J., and Procaccia, A. D. (2015). Impartial peer review. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Lamont, M. (2009). *How Professors Think*. Harvard University Press.

Lawrence, N. and Cortes, C. (2014). The NIPS Experiment. `http://inverseprobability.com/2014/12/16/the-nips-experiment`.

Lee, C. (2015). Commensuration bias in peer review. *Philosophy of Science*, 82:1272–1283.

MacKay, R. S., Kenna, R., Low, R. J., and Parker, S. (2017). Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*, 4(2).

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 1(2):161–175.

Manzoor, E. and Shah, N. B. (2021). Uncovering latent biases in text: Method and application to peer review. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Masnadi-Shirazi, H. and Vasconcelos, N. (2008). On the design of loss functions for classification: Theory, robustness to outliers, and SavageBoost. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Mei, S., Bai, Y., and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *Annals of Statistics*, 46(6A):2747–2774.

Mogul, J. C. and Anderson, T. (2008). Before and after wowcs: A literature survey, a list of papers we wish had been submitted. In *Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems (WOWCS)*.

Moulin, H. (1983). *The Strategy of Social Choice*, volume 18 of *Advanced Textbooks in Economics*. North-Holland.

Nie, F., Huang, H., Cai, X., and Ding, C. H. (2010). Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Nielsen, M. W., Baker, C. F., Brady, E., Petersen, M. B., and Andersen, J. P. (2021). Meta-research: Weak evidence of country-and institution-related status bias in the peer review of abstracts. *Elife*.

Rahimpour, A., Taalimi, A., and Qi, H. (2017). Feature encoding in band-limited distributed surveillance systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Roos, M., Rothe, J., and Scheuermann, B. (2011). How to calibrate the scores of biased reviewers by quadratic programming. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M., and Verri, A. (2004). Are loss functions all the same? *Neural Computation*, 16(5):1063–1076.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., and Wainwright, M. J. (2017). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory*, 63(2):934–959.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. (2019). Low permutation-rank matrices: Structural properties and noisy completion. *Journal of Machine Learning Research*.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. (2020). A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*.

Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., and Von Luxburg, U. (2018). Design and analysis of the NIPS 2016 review process. *The Journal of Machine Learning Research*, 19(1):1913–1946.

Spain, P. G. (1996). The Fermat point of a triangle. *Mathematics Magazine*, 69(2):131–133.

Stelmakh, I., Rastogi, C., Shah, N. B., Singh, A., and Daumé III, H. (2020). A large scale randomized controlled trial on herding in peer-review discussions. *arXiv preprint arXiv:2011.15083*.

Stelmakh, I., Shah, N., and Singh, A. (2019a). PeerReview4All: Fair and accurate reviewer assignment in peer review. In *Proceedings of the Conference on Algorithmic Learning Theory*.

Stelmakh, I., Shah, N., and Singh, A. (2021a). Catch me if i can: Detecting strategic behaviour in peer assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Stelmakh, I., Shah, N., Singh, A., and Daum III, H. (2021b). A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Stelmakh, I., Shah, N., Singh, A., and Daum III, H. (2021c). Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. In *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.

Stelmakh, I., Shah, N. B., and Singh, A. (2019b). On testing for biases in peer review. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Tomkins, A., Zhang, M., and Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713.

Vijaykumar, T. N. (2020a). Potential organized fraud in ACM/IEEE computer architecture conferences. Online `https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences-ccd61169370d` Last accessed April 4, 2021.

Vijaykumar, T. N. (2020b). Potential organized fraud in ongoing ASPLOS reviews. Online `https://medium.com/@tnvijayk/potential-organized-fraud-in-on-going-asplos-reviews-874ce14a3ebe` Last accessed April 4, 2021.

Wang, J. and Shah, N. B. (2019). Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Xu, Y., Zhao, H., Shi, X., and Shah, N. (2019). On strategyproof conference review. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Zhaoshui, H. and Cichocki, A. (2008). Sparse representation of L-order Markov signals. In *International Symposium on Nonlinear Theory and its Applications*.