Choosing How to Choose Papers

Ritesh Noothigattu   Nihar B. Shah   Ariel D. Procaccia

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 — based on $L(p, q)$-norm empirical risk minimization — for learning the community's aggregate mapping. We draw on computational social choice to identify desirable values of p and q; specifically, we characterize $p = q = 1$ as the only choice 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.[1] The former editor of the British Medical Journal writes (Smith, 2006):

> *"People have a great many fantasies about peer review, and one of the most powerful is that it is a highly objective, reliable, and consistent process. I regularly received letters from authors who were upset that the BMJ rejected their paper and then published what they thought to be a much inferior paper on the same subject. Always they saw something underhand. They found it hard to accept that peer review is a subjective and, therefore, inconsistent process."*

Of course, this criticism is equally relevant to computer sci-

School of Computer Science, Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Ritesh Noothigattu <rnoothig@cs.cmu.edu>.

[1]Even papers about peer review are subject to peer review, the irony of which has not escaped us.

ence in general, and AI in particular, but AI researchers have been playing a leading role in using their craft to propose and test new approaches to peer review (Roos et al., 2011; Charlin & Zemel, 2013; Ge et al., 2013; Lawrence & Cortes, 2014; Kurokawa et al., 2015; Shah et al., 2017).

We contribute to this effort by examining one of the sources of inconsistency in peer review: the mapping from criteria scores to overall recommendation. Specifically, most conferences ask reviewers to score papers on criteria such as novelty, significance, and technical quality. But there is rarely any guidance as to how these criteria should be aggregated in order to arrive at a recommendation in a principled way. As a consequence, the acceptance decision on each paper is governed by the subjective opinions of a handful of reviewers (who review that paper) regarding the relative importance of various criteria. Furthermore, these aggregation rules are not consistent across different papers.

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). Some reviewers chose not to follow these instructions, as noted by the program chairs in an email to a subset of the program committee:

> *"Note that the reviewer forms have directions for how to assign the summary score. Ideally, this score would have been automated, but it was not. Some reviewers have not properly applied the rules."*

Although the specific issue we described can be alleviated via a revised mapping of criteria to recommendations, that

hand-crafted revised mapping is likely to have its own quirks. In summary, the AAAI-13 anecdote highlights the difficulty of devising a principled mapping from criteria scores to recommendations.

To address this problem, we propose an approach based on machine learning and social choice, which provides 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 innocuous 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 these data, we find a single monotonic mapping that minimizes a loss function (which we will discuss momentarily). We can then apply this mapping to the criteria scores associated with each review, to obtain new overall recommendations, which replace the original ones.

In the foregoing framework, the overall recommendation initially given by a reviewer only serves to express an opinion about the desired mapping in one particular case; we achieve consistency across reviews by discarding these initial recommendations and applying a single mapping, which is the output of the learning process, exclusively to criteria scores. Therefore, this process lies in the realm of *computational social choice* (Brandt et al., 2016): it can be seen as aggregating 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).

Specifically, we focus on the the family of $L(p,q)$ loss functions, for $p,q \in [1, \infty]$, which resembles the $L(p,q)$ norm of a matrix. Our question, then, is:

> For what values of p and q does our aggregation method, which chooses a monotonic mapping that minimizes the $L(p,q)$ loss on the given reviews, satisfy consensus, efficiency, and strategyproofness?

Our main theoretical result is a characterization theorem that gives a decisive answer to this question: the three properties are satisfied *if and only if* $p = q = 1$. 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 actual list of accepted papers, suggesting that our approach makes a significant difference compared to the status quo (arguably for the better).

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 *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. We do this by computing the function in $\mathcal{F}$ that minimizes the $L(p,q)$ loss on the data. In more detail, given $p, q \in [1, \infty]$, we compute

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

In words, for a function f, the $L(p,q)$ loss is the L_q norm taken over the loss associated with individual reviewers, where the latter loss is defined as the L_p norm computed on the error of f with respect to the reviewer's overall recommendations. This definition closely resembles the $L(p,q)$ norm of a matrix, which has many applications in machine learning (Ding et al., 2006; Kong et al., 2011; Nie et al., 2010). We refer to aggregation by minimizing $L(p,q)$ loss as defined in Equation (1) as $L(p,q)$ *aggregation*.

Equation (1) does not specify how to break ties between multiple minimizers. For concreteness, we select the minimizer $\widehat{f}$ with minimum empirical L_2 norm. Mathematically, let

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

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

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

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. We emphasize that although we use minimum L_2 norm for tie-breaking, all of our results hold under any reasonable tie-breaking method, such as the minimum L_k norm for any $k \in (1, \infty)$.

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 (typically this is done by taking the average). 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 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 $p, q \in [1, \infty]$, $L(p,q)$ 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 are independent of i.

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 final axiom is *strategyproofness*, a game-theoretic property that plays a major role in social choice theory (Moulin, 1983). Intuitively, 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.

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

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

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

The implicit 'utilities' in this axiom (3) are defined in terms of the L_2 norm. This choice is made only for concreteness, and all our results hold for any norm $L_\ell, \ell \in [1, \infty]$, in the definition (3).

3. Theoretical Results

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.

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

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

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

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

main.texwhere 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 obvious 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) = $ 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. $\square$

Next we tackle the harder 'only if' direction of Theorem 1. We do so in three steps: strategyproofness is violated by $L(p,q)$ aggregation for all $q > 1$ (Lemma 2), consensus is violated by $p = \infty$ and $q = 1$ (Lemma 3), and efficiency is violated by $p \in (1, \infty)$ 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.

Lemma 2. $L(p,q)$ *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

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

and for $q = \infty$, it is

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

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.

Lemma 3. $L(p,q)$ *aggregation with* $p = \infty$ *and* $q = 1$ *violates consensus.*

Lemma 3 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.

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

Lemma 4. $L(p,q)$ *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 we attempt to explain this phenomenon via a connection of our problem with 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.

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. Let the overall recommendations given by the reviewers be $y_{11} = z$, $y_{21} = 0$, $y_{31} = 0$ to the first paper and $y_{12} = 0$, $y_{22} = 1$ and $y_{23} = 0$ to the second paper. Under these scores, let $\widehat{f}$ denote the aggregate function that minimizes the $L(2, 1)$ loss.

The Fermat point of a triangle is a point such that the sum of its (Euclidean) distances from all three vertices is minimized. Consider a triangle in $\mathbb{R}^2$ with vertices $(z, 0)$, $(0, 1)$ and $(0, 0)$. Setting $z = 2$, one can use known algorithms to compute the Fermat point of this triangle as $(0.25, 0.30)$. More generally, when the vertex $(z, 0)$ is moved away from the rest of the triangle (by increasing z), the Fermat point paradoxically biases towards the other (second) coordinate.

Connecting back to our original problem, by definition, the Fermat point of this triangle is exactly $(\widehat{f}(\mathbf{x}_1), \widehat{f}(\mathbf{x}_2))$. When $z = 2$, paper 1 receives scores $(2, 0, 0)$ in sorted order, which dominates the sorted scores $(1, 0, 0)$ of paper 2. However the aggregate score $\widehat{f}(\mathbf{x}_1) = 0.25$ of paper 1 is strictly smaller than $\widehat{f}(\mathbf{x}_2) = 0.30$ of paper 2, thereby violating efficiency for the $L(2, 1)$ loss.

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 2 would give significant guidance on choosing an appropriate method.

4. Implementation and Experimental Results

In order to empirically analyze different aspects of our approach, we employ a dataset of reviews from the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), which was made available to us by the program chair, Carles Sierra. 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 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 $\widetilde{f}$. 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:

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

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.

4.1. Varying Number of Reviewers

In our first experiment, for each value of a parameter $k \in \{1, \dots, 5\}$, 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 1, and

# 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
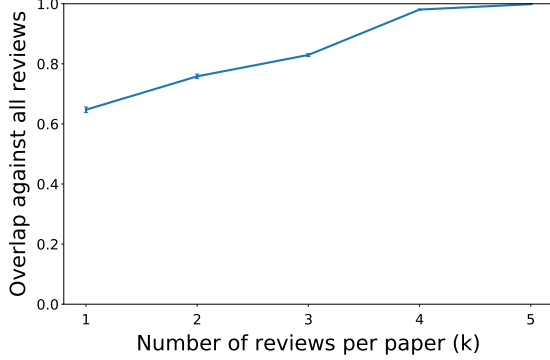
Figure 1. Fraction overlap as number of reviews per paper is restricted

Figure 2. Frequency of losses of the reviewers for $L(1,1)$ aggregation, normalized by the number of papers reviewed by the respective reviewers

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.
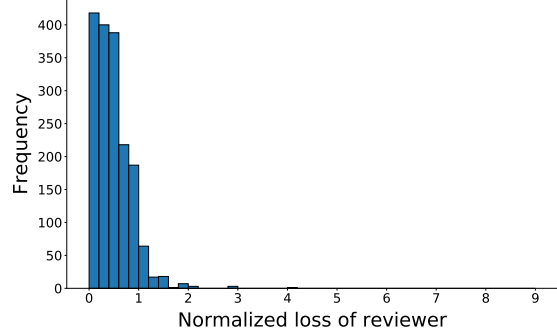
4.2. Loss Per Reviewer

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 2 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. 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 visualization and interpretation 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). We present the two most representative plots in Figure 3, and relegate the rest to Appendix B.

The key takeaways from this experiment are as follows. First, writing and relevance do not have a significant influence (Figure 3(a)). 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 3(b)). 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 to conference review data — for some criteria one may indeed assume a linear model, but not for all.

4.4. Overlaps with Other Aggregates

We also compute the overlap between the set of top 27.27% papers selected by $L(1,1)$ aggregation $\widetilde{f}$ with the *actual*
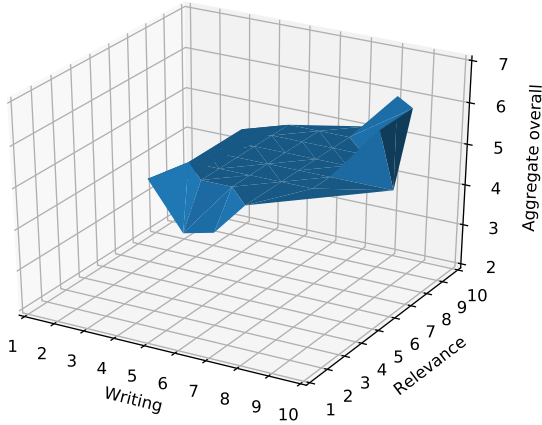
(a) Varying 'quality of writing' and 'relevance'
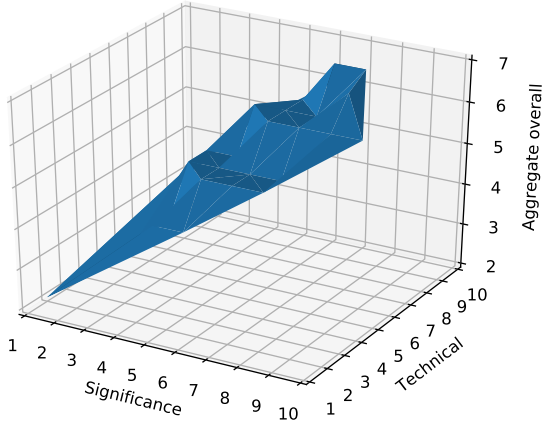
(b) Varying 'significance' and 'technical quality'

Figure 3. Influence of the different criteria under in the $L(1, 1)$ aggregate.

top 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.

Finally, 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 q has more influence than the paper-norm p; we refer the reader to Appendix B for details.

5. Discussion

A primary challenge in the analysis of peer review is the absence of ground truth. This challenge is compounded by the paucity of models for problems such as subjective mappings. We address these challenges by leveraging tools from social choice theory and taking an axiomatic viewpoint, and then making only the weak assumption that the mapping for any reviewer is monotonic.

In particular, 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 (as well as phenomena related to the data itself). Understanding this helps clear up another 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.

That said, it is of intellectual 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 that 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 can $L(1, 1)$ aggregation recover the function $f^{\star\star}$ (in the statistical consistency sense)? If so, then with what sample complexity? At a conceptual level, this non-parametric estimation problem is closely related to problems in isotonic regression (Shah et al., 2016; Gao & 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, the insights from isotonic regression suggest that the naturally occurring monotonicity assumption of our setting can yield attractive — and sometimes near-parametric (Shah et al., 2016; 2018) — rates of estimation.

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 — typically in an *ad hoc* fashion — through the lens of computational social choice.

References

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

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

Charlin, L. and Zemel, R. The Toronto paper matching system: An automated paper-reviewer assignment system. Manuscript, 2013.

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

Ding, C., Zhou, D., He, X., and Zha, H. R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 281–288, 2006.

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

Ge, H., Welling, M., and Ghahramani, Z. A Bayesian model for calibrating conference review scores. Manuscript, 2013.

Kong, D., Ding, C., and Huang, H. Robust nonnegative matrix factorization using L21-norm. In *Proceedings of the 20th International Conference on Information and Knowledge Management (CIKM)*, pp. 673–682, 2011.

Kurokawa, D., Lev, O., Morgenstern, J., and Procaccia, A. D. Impartial peer review. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 582–588, 2015.

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

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

Nie, F., Huang, H., Cai, X., and Ding, C. H. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 1813–1821, 2010.

Roos, M., Rothe, J., and Scheuermann, B. How to calibrate the scores of biased reviewers by quadratic programming. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 255–260, 2011.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., and Wainwright, M. J. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 11–20, 2016.

Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., and von Luxburg, U. Design and analysis of the NIPS 2016 review process. arXiv:1708.09794, 2017.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. Low permutation-rank matrices: Structural properties and noisy completion. arXiv:1709.00127, 2018.

Smith, R. Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99:178–182, 2006.

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

Appendix

A. Proof Of Theorem 1

Recall that the proof of Theorem 1 includes three lemmas for the 'only if' direction. We give the proofs of these lemmas below.

A.1. Proof of Lemma 2

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

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

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

$$\widehat{f}_a = \operatorname*{argmin}_{f_a \in [0.5, 1)} \left\{ (1 - f_a)^q + f_a^q \right\}. \tag{9}$$

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 (9) 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.2. Proof of Lemma 3

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:

$$(\widehat{f}_1, \widehat{f}_2) \in \underset{(f_1, f_2) \in \mathbb{R}^2}{\operatorname{argmin}} \left\{ \left\| (0, 1) - (f_1, f_2) \right\|_\infty + \left\| (2, 1) - (f_1, f_2) \right\|_\infty \right\}. \tag{10}$$

We do not have any monotonicity constraints in (10) because the two papers have incomparable criteria scores. Denote the objective function of (10) by $G(f_1, f_2)$. We can simplify this objective to

$$G(f_1, f_2) = \max(|f_1|, |f_2 - 1|) + \max(|2 - f_1|, |f_2 - 1|). \tag{11}$$

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.[2] □

COMPLETE PICTURE OF MINIMIZERS.

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

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

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 $(1, 2)$.
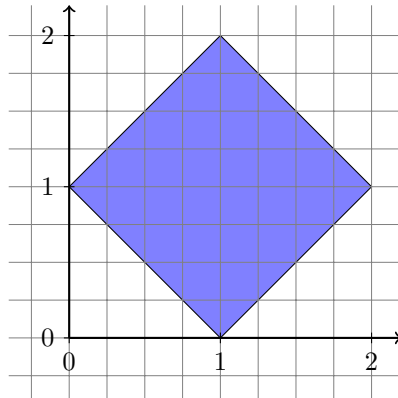

Figure 4. The shaded region depicts the set of all minimizers of (10). f_1 is on the x-axis and f_2 is on the y-axis.

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.

---

[2]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.

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 $f_1 + f_2 = 1$ $(0 \le f_1 \le 1)$ 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.

A.3. Proof of Lemma 4

Consider $L(p, 1)$ aggregation with an arbitrary $p \in (1, \infty)$. We show that efficiency is violated. The construction for this is as follows. 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:

$$(\widehat{f}_1(z), \widehat{f}_2(z)) \in \operatorname*{argmin}_{(f_1, f_2) \in \mathbb{R}^2} \left\{ \left\| (z, 0) - (f_1, f_2) \right\|_p + \left\| (0, 1) - (f_1, f_2) \right\|_p + \left\| (f_1, f_2) \right\|_p \right\}. \tag{12}$$

We do not have any monotonicity constraints in (12) 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 (12) by $G_z(\mathbf{f})$. That is,

$$G_z(f_1, f_2) = \left[|z - f_1|^p + |f_2|^p \right]^{\frac{1}{p}} + \left[|f_1|^p + |1 - f_2|^p \right]^{\frac{1}{p}} + \left[|f_1|^p + |f_2|^p \right]^{\frac{1}{p}}. \tag{13}$$

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 (13). Hence, for the rest of the proof, we only consider the space $[0, 1]^2$. In this region, the optimization problem (12) can be rewritten as

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

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.,

$$H_z(f_1, f_2) = \left[(z - f_1)^p + f_2^p \right]^{\frac{1}{p}} - z + \left[f_1^p + (1 - f_2)^p \right]^{\frac{1}{p}} + \left[f_1^p + f_2^p \right]^{\frac{1}{p}} - 1. \tag{14}$$

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

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

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

$$H^\star(f_1, f_2) = -f_1 + \left[f_1^p + (1 - f_2)^p \right]^{\frac{1}{p}} + \left[f_1^p + f_2^p \right]^{\frac{1}{p}} - 1. \tag{15}$$

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*

$$\widehat{v}_1 = \frac{1}{2} \left[\frac{1}{(2^{\frac{p}{p-1}} - 1)} \right]^{\frac{1}{p}}, \qquad \widehat{v}_2 = \frac{1}{2}. \tag{16}$$

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

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| < \epsilon. \tag{17}$$

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 (17) 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 (17) 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

$$\widetilde{z}_\epsilon := z_\epsilon(1,1) + 1. \tag{18}$$

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 (17) 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

$$H^\star(f_1, f_2) - \epsilon < H_z(f_1, f_2) < H^\star(f_1, f_2) + \epsilon, \tag{19}$$

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

$$\tau := \min_{\mathbf{f} \in [0,1]^2 : \|\mathbf{f} - \widehat{\mathbf{v}}\|_2 = \frac{\Delta}{4}} H^\star(\mathbf{f}). \tag{20}$$

Observe that a minimum exists (infimum is not required) for the minimization in (20) 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 (16) 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 (20) 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:

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

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

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 $\|\widehat{\mathbf{f}}(\widetilde{z}_{\epsilon_o}) - \widehat{\mathbf{v}}\|_2 \leq \frac{\Delta}{4}$, completing part (i).

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

$$\left|\widehat{f}_1(\widetilde{z}_{\epsilon_o}) - \widehat{v}_1\right| \leq \frac{\Delta}{4} \quad \text{and} \quad \left|\widehat{f}_2(\widetilde{z}_{\epsilon_o}) - \widehat{v}_2\right| \leq \frac{\Delta}{4}. \tag{21}$$

Using these properties, we have

$$\widehat{f}_1(\widetilde{z}_{\epsilon_o}) \leq \widehat{v}_1 + \frac{\Delta}{4}$$
$$= \widehat{v}_2 - \Delta + \frac{\Delta}{4}$$
$$\leq \widehat{f}_2(\widetilde{z}_{\epsilon_o}) + \frac{\Delta}{4} - \Delta + \frac{\Delta}{4} = \widehat{f}_2(\widetilde{z}_{\epsilon_o}) - \frac{\Delta}{2},$$

where the first inequality holds because of the first part of (21), the equality holds because $\Delta = \widehat{v}_2 - \widehat{v}_1$ and the second inequality holds because of the second part of (21). 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) $[(z, 0) - \mathbf{f}]$ is not parallel to $[(z, 0) - \mathbf{g}]$, (ii) $[(0, 1) - \mathbf{f}]$ is not parallel to $[(0, 1) - \mathbf{g}]$ or (iii) $\mathbf{f}$ is not parallel to $\mathbf{g}$. 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

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

where $r, s, t \in \mathbb{R}$ [3]. 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.

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

$$\|\theta x + (1-\theta)y\|_p \leq \theta\|x\|_p + (1-\theta)\|y\|_p. \tag{22}$$

---

[3] 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 $\mathbf{f} = \mathbf{g}$.

Further, since $p \in (1, \infty)$, the inequality in (22) is strict if x is not parallel to y. For our objective (in Equation (12)),

$$
\begin{aligned}
G_z(\theta \mathbf{f} + (1 - \theta)\mathbf{g}) &= \left\| \theta[(z, 0) - \mathbf{f}] + (1 - \theta)[(z, 0) - \mathbf{g}] \right\|_p \\
&+ \left\| \theta[(0, 1) - \mathbf{f}] + (1 - \theta)[(0, 1) - \mathbf{g}] \right\|_p \\
&+ \left\| \theta \mathbf{f} + (1 - \theta)\mathbf{g} \right\|_p.
\end{aligned}
\tag{23}
$$

Because of convexity of the L_p norm, each of the three terms on the RHS of Equation (23) satisfies inequality (22). 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

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

$\qquad\square$

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. $\qquad\square$

Proof of Claim 3. Take any arbitrary $f_1 \in [0, 1]$ and $f_2 \in [0, 1]$. Subtracting Equations (14) and (15) we obtain

$$
H_z(f_1, f_2) - H^\star(f_1, f_2) = \left[(z - f_1)^p + f_2^p\right]^{\frac{1}{p}} - (z - f_1).
\tag{24}
$$

Observe that since $f_2 \geq 0$, the RHS of Equation (24) is non-negative. Hence, the equation does not change on using an absolute value, i.e.,

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| = \left[(z - f_1)^p + f_2^p\right]^{\frac{1}{p}} - (z - f_1). \tag{25}$$

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

$$\lim_{x \to \infty} \phi(x) = \lim_{x \to \infty} \frac{\left(1 + \frac{f_2^p}{x^p}\right)^{\frac{1}{p}} - 1}{\frac{1}{x}}. \tag{26}$$

Observe that for both the numerator and denominator in the RHS of Equation (26), 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 (26) 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 (25), we rewrite it as

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

Taking the limit of z to infinity, we obtain

$$\lim_{z \to \infty} |H_z(f_1, f_2) - H^\star(f_1, f_2)| = \lim_{z \to \infty} \phi(z - f_1) = \lim_{t \to \infty} \phi(t) = 0, \tag{27}$$

where the second step follows by setting $t = z - f_1$. Equation (27) implies that

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

$\square$

Proof of Claim 4. In the region $[0, 1]^2$, using (15), the function $H^\star$ can be written as

$$H^\star(f_1, f_2) = -f_1 + \|(0, 1) - (f_1, f_2)\|_p + \|(f_1, f_2)\|_p - 1. \tag{28}$$

Observe that each term on the RHS of (28) 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 $x, y \in \mathbb{R}^2$,

$$\|\theta x + (1 - \theta)y\|_p \leq \theta\|x\|_p + (1 - \theta)\|y\|_p. \tag{29}$$

And since $p \in (1, \infty)$, the inequality in (29) is strict if x is not parallel to y. For $H^{\star}$ (using Equation (28)),

$$\begin{aligned}
H^{\star}(\theta \mathbf{f} + (1 - \theta)\mathbf{g}) = & -\theta f_1 - (1 - \theta)g_1 \\
& + \big\|\theta[(0, 1) - \mathbf{f}] + (1 - \theta)[(0, 1) - \mathbf{g}]\big\|_p \\
& + \big\|\theta\mathbf{f} + (1 - \theta)\mathbf{g}\big\|_p - 1.
\end{aligned} \tag{30}$$

Because of convexity of the L_p norm, both the third and fourth term on the RHS of Equation (30) satisfy inequality (29). 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

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

$\square$

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 (15), the partial derivative with respect to f_1 is

$$\frac{\partial H^{\star}}{\partial f_1} = -1 + f_1^{p-1}\left[f_1^p + (1 - f_2)^p\right]^{\frac{1}{p} - 1} + f_1^{p-1}\left[f_1^p + f_2^p\right]^{\frac{1}{p} - 1} \tag{31}$$

and with respect to f_2 is

$$\frac{\partial H^{\star}}{\partial f_2} = 0 - (1 - f_2)^{p-1}\left[f_1^p + (1 - f_2)^p\right]^{\frac{1}{p} - 1} + f_2^{p-1}\left[f_1^p + f_2^p\right]^{\frac{1}{p} - 1}. \tag{32}$$

Observe that at $f_2 = \frac{1}{2}$, irrespective of the value of f_1, the partial derivative (32) is

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

So, set $\widehat{v}_2 = \frac{1}{2}$. Next, we find $\widehat{v}_1$ such that the other derivative (31) is also zero at $\widehat{\mathbf{v}} = (\widehat{v}_1, \widehat{v}_2)$. Setting (31) to zero at $\widehat{\mathbf{v}}$, we obtain

$$\frac{\partial H^{\star}}{\partial f_1}\bigg|_{\mathbf{f} = \widehat{\mathbf{v}}} = 0 = -1 + \widehat{v}_1^{p-1}\left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{\frac{1}{p} - 1} + \widehat{v}_1^{p-1}\left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{\frac{1}{p} - 1}$$

$$\implies 1 = 2\widehat{v}_1^{p-1}\left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{\frac{1}{p} - 1}$$

$$\implies \left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{1-\frac{1}{p}} = 2\widehat{v}_1^{p-1}$$

$$\implies \left[\widehat{v}_1^p + \frac{1}{2^p}\right]^{p-1} = 2^p\widehat{v}_1^{p(p-1)}$$

$$\implies \widehat{v}_1^p + \frac{1}{2^p} = 2^{\frac{p}{p-1}}\widehat{v}_1^p$$

$$\implies \frac{1}{2^p} = \widehat{v}_1^p\left(2^{\frac{p}{p-1}} - 1\right)$$

$$\therefore \widehat{v}_1 = \frac{1}{2}\left[\frac{1}{\left(2^{\frac{p}{p-1}} - 1\right)}\right]^{\frac{1}{p}}.$$

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. $\qquad\square$

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

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

$\qquad\square$

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 (14) and (15), and taking an absolute value, we obtain Equation (25), that is,

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| = \left[(z - f_1)^p + f_2^p\right]^{\frac{1}{p}} - (z - f_1). \tag{33}$$

Combining Equation (33) with the fact that $0 \le f_2 \le 1$, we obtain

$$|H_z(f_1, f_2) - H^\star(f_1, f_2)| \le \left[(z - f_1)^p + 1\right]^{\frac{1}{p}} - (z - f_1). \tag{34}$$

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

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

for $x \ge 0$, showing that it is a non-increasing function. Going back to Equation (34), we know that $f_1 \le 1$. Therefore, $(z - f_1) \ge (z - 1) \ge 0$. Using the fact that ψ is a non-increasing function, we obtain $\psi(z - f_1) \le \psi(z - 1)$, which on expansion gives us

$$\left[(z - f_1)^p + 1\right]^{\frac{1}{p}} - (z - f_1) \le \left[(z - 1)^p + 1\right]^{\frac{1}{p}} - (z - 1) = |H_z(1,1) - H^\star(1,1)|. \tag{35}$$

Combining Equations (34) and (35), and the fact that $z \in \mathcal{Z}_\epsilon(1,1)$, we obtain

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

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

Proof of Claim 8. The proof follows using three facts:

1. Equation (19) 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 (19) gives us $H_z(\widehat{\mathbf{v}}) < H^\star(\widehat{\mathbf{v}}) + \epsilon$.

Putting these equations together:

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

$\square$

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 $H^\star$,

$$H^\star(\mathbf{g}) \leq \theta H^\star(\mathbf{f}) + (1-\theta)H^\star(\widehat{\mathbf{v}}). \tag{36}$$

By definition of τ in (20), we know $H^\star(\mathbf{g}) \geq \tau$. Also, we know $H^\star(\widehat{\mathbf{v}}) < \tau$. Using these in (36), 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. $\square$

B. Additional Empirical Results

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

B.1. Influence of Varying p, q

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). Since the table is symmetric, we show only the upper triangle to avoid clutter. 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		100.0	94.5	88.3	94.9	93.1	87.7	94.6	92.3
1,3			100.0	84.0	92.1	95.2	83.5	91.8	94.0
2,1				100.0	89.8	84.4	95.7	89.5	84.0
2,2					100.0	94.1	89.8	98.8	93.7
2,3						100.0	84.4	94.1	98.6
3,1							100.0	89.7	84.0
3,2								100.0	93.8
3,3									100.0

Table 2. Percentage of overlap (in selected papers) between different $L(p,q)$ aggregation methods
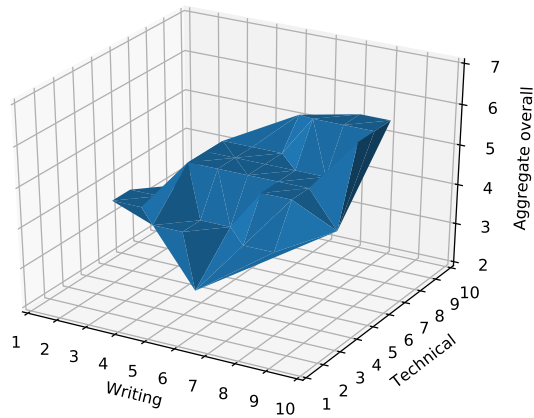
B.2. Influence of Varying Criteria

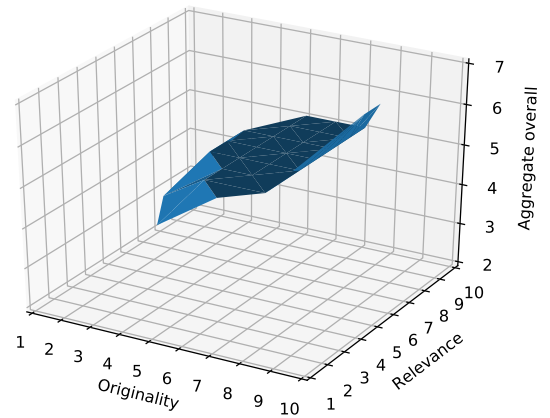Figures 5 and 6 present the visualizations of $\widehat{f}$ as all pairs of criteria are varied.
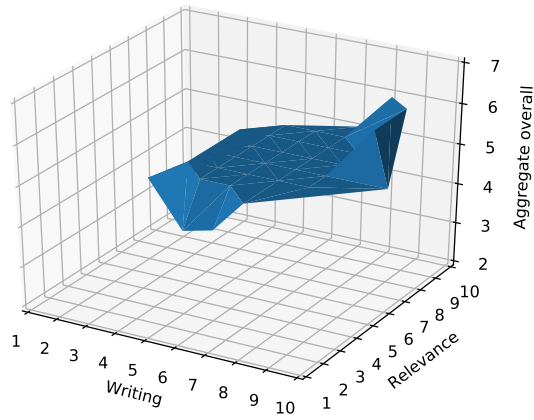
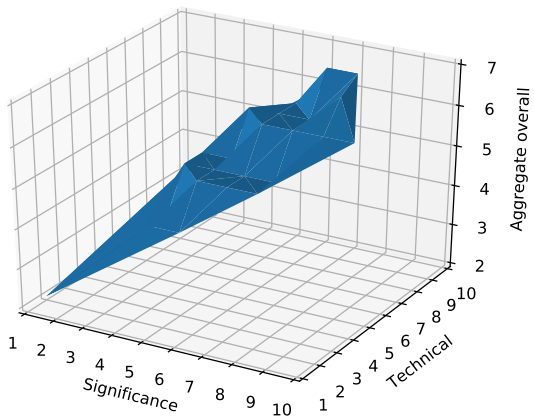(a) Varying 'quality of writing' and 'relevance'
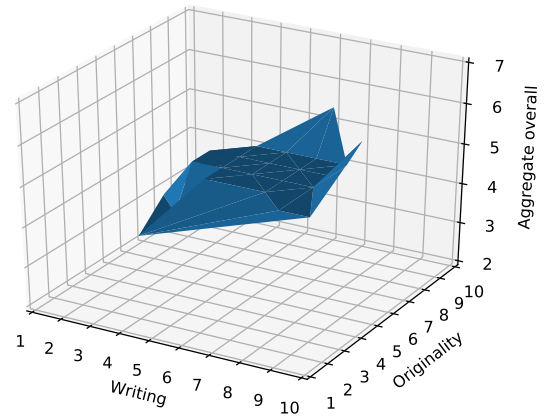
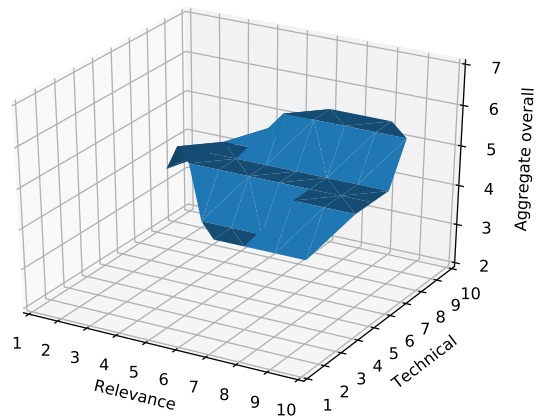(b) Varying 'quality of writing' and 'originality'

(c) Varying 'quality of writing' and 'technical quality'
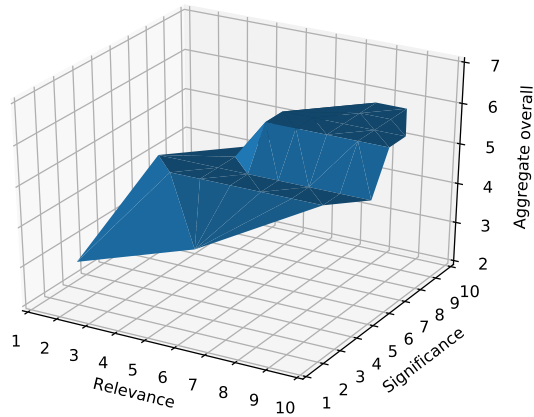
(d) Varying 'originality' and 'relevance'

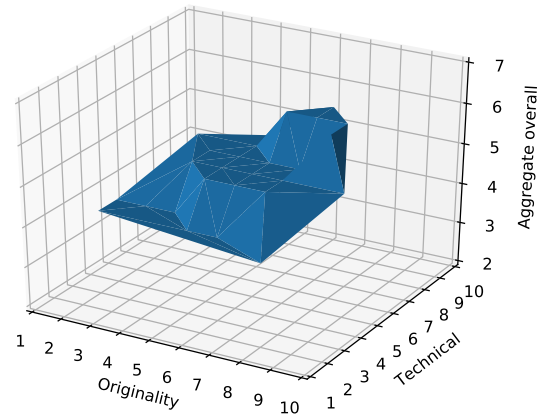(e) Varying 'significance' and 'technical quality'

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

Figure 5. Impact of varying different criteria under $L(1, 1)$ aggregation
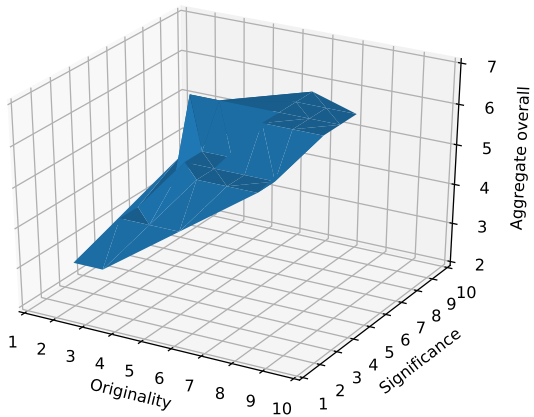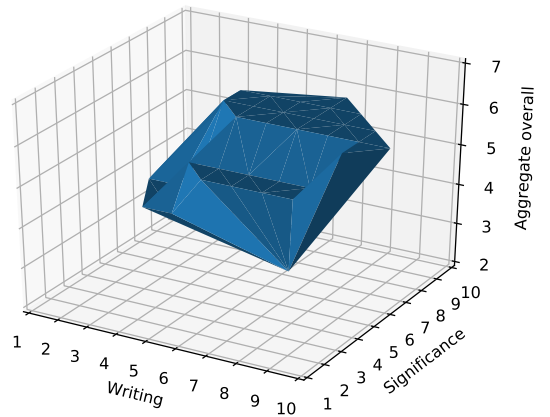
(a) Varying 'relevance' and 'significance'

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

(c) Varying 'originality' and 'significance'

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

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