Calibration with Privacy in Peer Review: A Theoretical Study

Wenxin Ding,[1] **Gautam Kamath,** [2] **Weina Wang,** [3] **Nihar B. Shah,** [3]

[1] University of Chicago
[2] University of Waterloo
[3] Carnegie Mellon University
wenxind@uchicago.edu, g@csail.mit.edu, {weinaw, nihars}@cs.cmu.edu

Abstract

Reviewers in peer review are often miscalibrated: they may be strict, lenient, extreme, moderate, etc. A number of algorithms have previously been proposed to calibrate reviews. Such attempts of calibration can however leak sensitive information about which reviewer reviewed which paper. In this paper, we identify this problem of calibration with privacy, and provide a foundational building block to address it. Specifically, we present a theoretical study of this problem under a simplified-yet-challenging model involving two reviewers, two papers, and an MAP-computing adversary. Our main results establish the Pareto frontier of the trade-off between privacy (preventing the adversary from inferring reviewer identity) and utility (accepting better papers), and design explicit computationally-efficient algorithms that we prove are Pareto optimal.

1 Introduction

It is well known that scores provided by people are frequently miscalibrated. In the application of peer review, reviewers may be strict, lenient, extreme, moderate, etc. This leads to unfairness in peer review, for instance, disadvantaging papers that happen to go to strict reviewers (Siegelman 1991): *"the existence of disparate categories of reviewers creates the potential for unfair treatment of authors. Those whose papers are sent by chance to assassins/demoters are at an unfair disadvantage, while zealots/pushovers give authors an unfair advantage."*

A number of algorithms (Flach et al. 2010; Ge, Welling, and Ghahramani 2013; Roos, Rothe, and Scheuermann 2011; Roos et al. 2012; Paul 1981; Baba and Kashima 2013; MacKay et al. 2017) are proposed in the literature to address the problem of miscalibration. There are two key challenges, however, towards any attempts of calibration using such algorithms:

Challenge #1: The calibration algorithms may leak information about which reviewer reviewed which paper. Here is an example showing how a naïve attempt at calibration can compromise privacy. Consider an adversary trying to guess the reviewer of a paper between two possibilities – reviewer X or reviewer Y. The review for the paper is lukewarm, and for simplicity suppose this is the only review. We consider the "open review" model where all submitted papers, reviews, and final decisions are public (but reviewer identities are not). Also suppose it is known that reviewer X is strict but reviewer Y is not. Then the conference will not accept the paper unless the conference performs a calibration using this information *and* the reviewer is X. The acceptance of the paper will provide the adversary with the necessary information to infer the reviewer as X.

Challenge #2: The bottleneck of a small number of samples (reviews) per reviewer. Many conferences have each reviewer reviewing only a handful papers (typically 1 to 6 papers), as well as have each paper reviewed by a handful of reviewers. As a consequence, it is often hard to decipher the miscalibration of any reviewer, particularly since human miscalibration can be quite complex (Brenner, Griffin, and Koehler 2005). Indeed, program chairs of conferences have tried to use some algorithms to calibrate reviewers' scores, but have found the outcomes to be unsatisfactory. For instance, John Langford, the program chair of the ICML 2012 conference says that *"We experimented with reviewer normalization and generally found it significantly harmful"* (Langford 2012).

Our focus on this paper is challenge #1 of privacy. Towards challenge # 2, we assume that the conference has exogenous information about the miscalibration of reviewers, such as reviewers' calibration information from other conferences where they have reviewed. (Appendix A presents simulations illustrating benefits of calibration with exogeneous information.) Tackling the problem of privacy in calibration that we identify is quite challenging in full generality. In this paper, our goal is to initiate research towards this grander goal by providing a foundational building block for it. We consider a simplified–yet highly challenging–model with two reviewers, two papers, and (exogenously) known miscalibration functions where an adversary attempts to guess the reviewer assignment based on maximum a posteriori (MAP) computation. We provide a comprehensive analysis under this model. Our contributions are summarized as follows:

- We identify the problem of privacy in calibration, and we initiate a theoretical study with the formulation of a problem that incorporates various key challenges of the more general setting.
- We provide explicit computationally-efficient algorithms

for calibration with privacy that optimally trades off the error of the conference (in terms of accepting the better paper) and the error of the adversary (in terms of guessing the reviewer).

- We establish the structure of the Pareto optimal curve between the two aforementioned desiderata. We observe that interestingly, there is a linear tradeoff between the two errors up to a certain point, after which the error of the adversary does not increase even if the conference adds more randomness in its protocols.

2 Related Work

Peer review is extensively used for evaluating scientific papers and grant proposals. However, conference peer review also incurs various challenges such as miscalibration (Flach et al. 2010; Ge, Welling, and Ghahramani 2013; Roos, Rothe, and Scheuermann 2011; Roos et al. 2012; Paul 1981; Baba and Kashima 2013; MacKay et al. 2017; Wang and Shah 2019; Shah et al. 2018), biases (Tomkins, Zhang, and Heavlin 2017; Manzoor and Shah 2021; Stelmakh, Shah, and Singh 2019; **?**), subjectivity (Lee 2015; Noothigattu, Shah, and Procaccia 2021; Mahoney 1977), dishonesty (Balietti, Goldstone, and Helbing 2016; Stelmakh, Shah, and Singh 2020; Xu et al. 2019; Littman 2021; Jecmen et al. 2020; Wu et al. 2021; Dhull et al. 2022), and others. See (Shah 2021) for a survey.

The problem of miscalibration is well recognized in the literature. A common approach to design calibration algorithms is to assume a certain model of miscalibration, and under the assumed model, estimate the calibrated scores (or the model parameters) from the scores given by reviewers. This line of literature (Flach et al. 2010; Ge, Welling, and Ghahramani 2013; Roos, Rothe, and Scheuermann 2011; Roos et al. 2012; Paul 1981; Baba and Kashima 2013; MacKay et al. 2017) assumes affine models for miscalibration: they assume that each paper has some "true" real-valued quality and that the score provided by any reviewer is some affine transform (plus noise) of this true quality. In our formulation (detailed subsequently in Section 3) we also assume papers have true qualities, and a part of our work also assumes affine miscalibrations.

A second line of literature (Mitliagkas et al. 2011; Ammar and Shah 2012; Freund et al. 2003) recognizes the problem of miscalibration, and takes the approach of using only the ranking of papers induced by the scores given by any individual reviewer, or alternatively, asking each reviewer to only provide a ranking of the papers they are reviewing. Using rankings alone thus gets rid of any miscalibrations, but on the downside, can lose some information contained in scores. Moreover, a recent work (Wang and Shah 2019) showed that under certain settings, scores can yield more information than rankings even if the miscalibration is adversarial.

Notably, these works consider addressing miscalibration using data from within the conference at hand, and moreover do not consider the issue of compromise of privacy.

We assume an "open review" model where all submitted papers and all reviews are available publicly, but where information of who reviews which paper is not. Such an open review model is gaining increasing popularity: see, for instance, openreview.net and scipost.org. This model is followed in the ICLR conference as well as other venues. In a survey (Soergel, Saunders, and McCallum 2013) at the ICLR 2013 conference, researchers felt that this open review model leads to benefits of more accountability of authors (in terms of not submitting below-par papers) as well as reviewers (in terms of giving high-quality reviews). The publicly available data has resulted in another benefit: it has yielded a rich dataset for research on peer review (Xu et al. 2019; Kang et al. 2018; Manzoor and Shah 2021; Tran et al. 2020; Bharadhwaj et al. 2020; Yuan, Liu, and Neubig 2021). A downside of the open review approach is that if a rejected paper is resubmitted elsewhere, the (publicly available) knowledge of previous rejection may bias the reviewer (Stelmakh et al. 2021).

Our work considers explicitly randomized assignments and decisions. In practice, the assignments and decision protocols are typically deterministic (although some variations naturally arise due to human involvement in various parts of the peer-review process). The assignment of reviewers to papers is done by solving a certain optimization problem (Goldsmith and Sloan 2007; Taylor 2008; Charlin and Zemel 2013; Garg et al. 2010; Stelmakh, Shah, and Singh 2021; Kobren, Saha, and McCallum 2019) involving similarities computed between each reviewer-paper pair (Mimno and McCallum 2007; Charlin and Zemel 2013; Fiez, Shah, and Ratliff 2020; Meir et al. 2020). Decisions are arrived at after discussions between the reviewers. That said, there are notable instances where randomization has been explicitly used in practice in peer review: randomization can help mitigate dishonest behavior (Jecmen et al. 2020) and can help make more fair decisions for borderline papers or grants (Liu et al. 2020; Chawla 2021). A recent survey of researchers finds support for randomized decisions (Philipps 2021). Finally, the algorithms in the theoretical work (Wang and Shah 2019) comparing scores and rankings in the context of miscalibration also employ randomization.

Issues of privacy in peer review also arise when releasing data to researchers. The program chairs of the WSDM 2017 conference performed a remarkable controlled experiment to test for biases in peer review, and in their paper (Tomkins, Zhang, and Heavlin 2017) they point out privacy-related concerns in releasing data: *"We would prefer to make available the raw data used in our study, but after some effort we have not been able to devise an anonymization scheme that will simultaneously protect the identities of the parties involved and allow accurate aggregate statistical analysis. We are familiar with the literature around privacy preserving dissemination of data for statistical analysis and feel that releasing our data is not possible using current state-of-the-art techniques."* We are aware of two past works which deal with privacy in peer review (Ding, Shah, and Wang 2020; Jecmen et al. 2020). In particular, both papers consider privacy-preserving release of peer-review data. The paper (Ding, Shah, and Wang 2020) provides an algorithm to optimize utility when releasing histograms of certain functions of the review scores. The paper (Jecmen et al. 2020)

Notation	Meaning
$i \in \{1,2\}$	Index for paper
$j \in \{1,2\}$	Index for reviewer
$\theta_i^* \in \mathbb{R}$	True quality of paper i
$\theta_i \in \mathbb{R}$	Estimated quality of paper i
A_1 and A_2	The two possible assignments
$s_i \in \mathbb{R}$	Score received by paper i; $S = [s_1, s_2]$
$\beta_j : \mathbb{R} \to \mathbb{R}$	Miscalibration function of reviewer j
$\epsilon_j \in \mathbb{R}$	Noise of reviewer j
$f_j : \mathbb{R} \to \mathbb{R}$	Marginal probability density function of score given by reviewer j, that is, distribution of $\beta_j(\theta^*)$ where $\theta^* \sim N(0,1)$
$\mathcal{E}_C \in [0,1]$	Error of the conference
$\mathcal{E}_A \in [0,1]$	Error of the adversary

Table 1: Summary of the main notation used in the paper.

uses randomized assignments to guarantee privacy of the reviewer-paper assignment when data pertaining to similarities between reviewer-paper pairs is released.

Differential privacy (Dwork et al. 2016) is a popular rigorous notion of data privacy. Roughly speaking, an algorithm is differentially private if its distribution over outputs is similar when provided with "neighboring" inputs. In our problem with two papers and two reviewers, one can consider neighboring inputs to be those that differ only in the assignment. We provide a tight characterization of the adversary's ability to determine which of the two possible assignments is the true one. Thus, it may be a useful building block towards more complex private calibration schemes. We note that our calibration algorithms are related to a form of randomized response (Warner 1965), the canonical algorithm for local differential privacy (Warner 1965; Evfimievski, Gehrke, and Srikant 2003; Kasiviswanathan et al. 2011). Though differential privacy is not the focus of our work, we further elaborate on this connection in Appendix B.

3 Problem Formulation and Preliminaries

In this section, we present the formal problem specification. We will introduce some notation in this section, and this notation is also summarized in Table 1.

Papers and reviewers. We consider a setting with two reviewers and two papers. Each paper $i \in \{1,2\}$ has some latent true quality $\theta_i^* \in \mathbb{R}$. We assume that the qualities θ_1^* and θ_2^* are drawn i.i.d. according to the standard normal distribution (and hence we have $\theta_1^* \neq \theta_2^*$ with probability 1).

Reviewer assignment. Each reviewer reviews one paper and each paper is reviewed by one reviewer. There are thus two possible assignments: we let A_1 denote the assignment of reviewer 1 to paper 1 and reviewer 2 to paper 2, and A_2 denote the assignment of reviewer 1 to paper 2 and reviewer 2 to paper 1. We assume that the assignment is chosen uniformly at random from these two possibilities. We assume that the true assignment is known (only) to the conference. We let $\mathcal{A}$ denote a random variable representing the assignment. Finally, in our exposition we will refer to the realization of $\mathcal{A}$ as the "true" assignment (and the unrealized assignment as the "wrong" assignment).

Miscalibration and reviewer scores. For each paper $i \in \{1,2\}$, we let s_i denote the score received by by paper i. Note that this notation is not indexed by the reviewer for brevity since each paper receives exactly one review. For convenience, we define the vector $S = [s_1, s_2]$. Following the popular "open review" model (https://openreview.netOpenReview.net, https://scipost.orgscipost.org), we assume that the scores s_1 and s_2 are known publicly.[1]

Following (Wang and Shah 2019), we assume that each reviewer $j \in \{1,2\}$ has a function $\beta_j : \mathbb{R} \to \mathbb{R}$ which captures their miscalibration. If reviewer $j \in \{1,2\}$ reviews paper $i \in \{1,2\}$, we assume that the reviewer provides a score $s_i \in \mathbb{R}$ given as:

$$s_i = \beta_j(\theta_i^*) + \epsilon_j,$$

where ϵ_j is a zero-mean Gaussian random variable independent of everything else. We assume that ϵ_1 and ϵ_2 are identically distributed. The value of the noise is unknown but its distribution is publicly known. We call β_j the *reviewer's miscalibration function* for reviewer j. We assume that the functions β_1 and β_2 are increasing and invertible. In one part of our work, we further make an assumption that the miscalibration functions are affine, and we detail this subsequently in the associated section. As discussed previously, our aim is to use exogenous information about the reviewer miscalibrations in order to mitigate the miscalibration, and to this end, we assume that the functions β_1 and β_2 are known publicly.

For any reviewer j, we let f_j denote the marginal probability density function of the final score given by that reviewer, that is, f_j is the distribution of $\beta_j(\theta^*)$ where $\theta^* \sim N(0,1)$.

Conference's error. The goal of the conference is to accept the paper with the higher true quality $\operatorname{argmax}_{i \in \{1,2\}} \theta_i^*$. Note that even if the noise terms were zero, simply choosing the paper with higher score (i.e., $\operatorname{argmax}_{i \in \{1,2\}} s_i$) may be erroneous due to the miscalibration of the reviewers. The conference can however calibrate the scores, that is, use the information about the miscalibration functions of the reviewers and the knowledge of the assignment to potentially make a better decision. In our analysis, we will measure the conference's performance towards its goal in terms of two types of errors:

(a) *Per-instance error*: For any given $S = [s_1, s_2]$, the per-instance error of the conference is defined as $\mathcal{E}_C([s_1, s_2]) := \Pr(\text{conference accepts lower-quality paper} \mid S = [s_1, s_2])$.

[1] Even if the conference operates in a non-open-review setting where the scores are not public, our guarantees on privacy and conference's error continue to hold. However, our algorithm may not be optimal and the suboptimality may depend on assumptions about the adversary's knowledge of the scores.

(b) *Average-case error*: The average-case error of the conference is the per-instance error averaged over the distribution of the scores: $\int_{s_1}\int_{s_2}\mathcal{E}_C([s_1,s_2])f'_S([s_1,s_2])$ where f'_S is the p.d.f. of the joint distribution of $S=[s_1,s_2]$.

In conjunction with the goal of minimizing the error, the conference must also ensure that information about which reviewer reviewed which paper is not leaked.

Privacy. We assume that the protocols followed by the conference are public. A challenge for the conference is that performing calibration may leak information about the assignment. As a simple example, suppose that reviewer 1 is known to be strict and reviewer 2 is known to be lenient. Suppose that paper 1 is reviewed by reviewer 1 and paper 2 by reviewer 2. Suppose paper 2 receives a higher score than paper 1, but the conference decides to accept paper 1 after performing calibration. This decision leaks information that paper 2 was reviewed by the lenient reviewer, that is, by reviewer 2. Note that this issue of compromise of privacy arises whether or not the reviewer miscalibration functions are known to the conference.

To formalize the notion of privacy, we assume an adversary in the process. The goal of the adversary is to guess the assignment. In addition to knowing the scores received by both papers, the miscalibration functions of both reviewers, the noise distributions, and the final decision of the conference, the adversary also knows the calibration strategy used by the conference to make the decision.

The adversary does not know the assignment, and aims to guess the assignment. We consider an adversary with no additional information, in which case, we assume it predicts the assignment via maximum a posteriori (MAP) estimation. Formally, if the conference decides to accept paper $P\in\{1,2\}$, then the adversary computes:

$$\operatorname*{argmax}_{A\in\{A_1,A_2\}}\ \Pr(\mathcal{A}=A\mid S=[s_1,s_2],\ \text{paper } P \text{ accepted}),$$

where $\mathcal{A}$ is the random variable representing the assignment. We make no assumptions on the computational power of the adversary and aim to guarantee privacy assuming they can compute the aforementioned argmax.

As in the case of the conference's error, we also measure the error of the adversary in two ways:

(a) *Per-instance error*: For any given $S=[s_1,s_2]$, the per-instance error of the adversary is defined as $\mathcal{E}_A([s_1,s_2]) := \Pr(\text{adversary guesses wrong assignment}\mid S=[s_1,s_2])$.

(b) *Average-case error:* The average-case error of the adversary is the per-instance error averaged over the distribution of the scores: $\int_{s_1}\int_{s_2}\mathcal{E}_A([s_1,s_2])f'_S([s_1,s_2])$ where f'_S is the p.d.f. of the joint distribution of $S=[s_1,s_2]$.

Goal. Our goal is to design methods to decide which paper to accept in a manner that simultaneously minimizes the conference's error and maximizes the adversary's error. The methods will inherently rely on calibrating reviewer decisions to accept the better paper, and hence we sometimes refer to them as the calibration strategy.

The two aforementioned objectives may conflict with one another: a decision that reduces the chances of accepting the lower quality paper via calibration can also leak more information about the assignment. In this work, we thus establish the Pareto frontier of this tradeoff. We define the Pareto frontier as the set of all points of the (conference's error, adversary's error) tradeoff such that the adversary's error cannot be increased without increasing the conference's error. We call a calibration strategy Pareto optimal if for any given threshold on conference's error, it maximizes the adversary's error while ensuring that the conference's error does not exceed the given threshold.

4 Main Results

In what follows, we present results for two settings: (1) a noiseless setting, where the noise in the reviewer-provided scores is zero; and (2) a noisy setting, where the noise in a reviewer score has a positive variance. We begin by a few preliminaries which we subsequently use to derive and present our main results.

4.1 Preliminaries

We now formalize the calibration strategies that a conference can follow in a general form, and then derive a specific form that can be used without loss of optimality. Our subsequent results will then use this form of the calibration strategies.

At a high level, the calibration strategies introduce a certain amount of randomness in the acceptance decisions. In the example in the 'privacy' paragraph earlier in this section, suppose the conference does the calibration, and then tosses a coin. With probability 0.9, it accepts the paper it thinks is better and otherwise it accepts the other paper. This randomness ensures that an adversary who observes that paper 1 is accepted cannot be certain that paper 1 was reviewed by reviewer 1, due to the possibility that paper 1 was reviewed by the lenient reviewer 2 but was still accepted due to the randomness. However, due to the randomness introduced, the conference incurs an error in terms of accepting the paper which it thought was actually better. There is thus a tradeoff between the conference's error and the adversary's error, and our goal is to design calibration strategies that are optimal with respect to this tradeoff.

Let us now formalize the notion of a calibration strategy. The conference observes the scores $S=[s_1,s_2]$ and the assignment A. Given these values, a generic calibration strategy is specified by a function $g: S\times A\to[0,1]$ — the conference accepts accept paper 1 with probability $g(S,A)$ and accepts paper 2 otherwise. Note that the function g is publicly known but its realization is known only to the conference. For any function g used by the conference, the conference's error is then given by

$$\mathcal{E}_C(S,A)=\Big((1-g(S,A_1))\Pr(\mathcal{A}=A_1|\theta_1^*>\theta_2^*,S)+(1-g(S,A_2))\Pr(\mathcal{A}=A_2|\theta_1^*>\theta_2^*,S)\Big)\Pr(\theta_1^*>\theta_2^*|S)+\Big(g(S,A_1)\Pr(\mathcal{A}=A_1|\theta_1^*<\theta_2^*,S)+g(S,A_2)\Pr(\mathcal{A}=A_2|\theta_1^*<\theta_2^*,S)\Big)\Pr(\theta_1^*<\theta_2^*|S).$$

Having specified this general form of calibration strategy, we now discuss a specific variant. If one did not care about the privacy, then the conference's error can be minimized via maximum a posteriori (MAP) estimation: given scores S and the assignment A, the conference accepts paper 1 if $\Pr(\theta_1^* > \theta_2^* | S, A) > 0.5$ and accepts paper 2 otherwise (breaking ties uniformly at random). Now under our scenario also involving privacy, consider the following class of calibration strategies. The strategy is governed by a function $h : S \times A \to [0, 1]$. Given $S = [s_1, s_2]$ and the assignment A:

- With probability $h(S, A)$, the conference executes MAP estimation under scores S and the (true) assignment A,
- otherwise (with probability $1 - h(S, A)$), the conference executes MAP estimation under scores S and the **wrong** assignment $\{A_1, A_2\} \backslash A$.

As before, we assume that function h is known publicly but its realization or the random bits are not.

A calibration strategy is **Pareto optimal** if any other strategy that incurs a lower conference error must also induce a lower error of the adversary, and any other strategy that induces a higher error of the adversary must also incur a higher conference error. The **Pareto frontier** is the set of all (conference error, adversary error) pairs achieved by Pareto optimal strategies. The following proposition states that without loss of optimality, one can restrict attention to the class of strategies specified by functions h.

Proposition 4.1. *For any values of error of the conference and error of the adversary $(\mathcal{E}_C, \mathcal{E}_A)$ achieved by a calibration strategy g, there exists a function h such that under h, the error of the conference is no larger than $\mathcal{E}_C$ and the error of the adversary is no smaller than $\mathcal{E}_A$.*

The proof of this proposition is available in Appendix C.1. Hence, without loss of Pareto optimality, any generic calibration strategy g can be replaced with a strategy involving the miscalibration function h. Thus, in the sequel we restrict attention to calibration strategies using function h.

4.2 Noiseless Setting

We first study the noiseless setting where the noise in the reviewer-provided scores is zero, that is, where $\epsilon_1 = \epsilon_2 = 0$. Observe that in this setting the conference can obtain the true qualities of the papers from the scores by inverting the reviewer functions. We first explicitly characterize the Pareto frontier for per-instance errors of the conference and the adversary. Based on this characterization, we then design Pareto optimal strategies for conference calibration with respect to the per-instance error and the average-case error.

Pareto Frontier for Per-Instance Errors In the following theorem, we present the main result of this section establishing the Pareto frontier for per-instance errors in the noiseless setting.

Theorem 4.2. *Consider the peer-review system in the noiseless setting. The Pareto frontier of (per-instance error of the conference, per-instance error of the adversary) with scores $S = [s_1, s_2]$ is given as follows.*

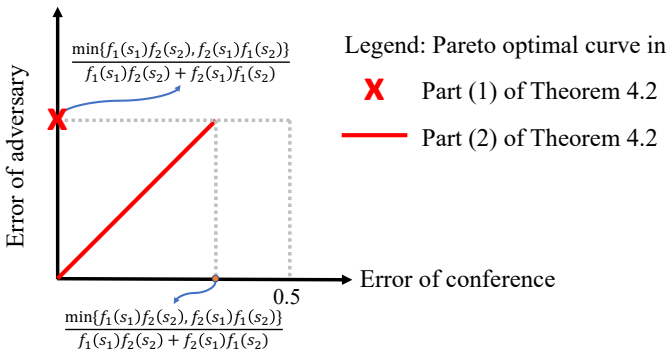


Figure 1: Pareto frontier for per-instance errors in the noiseless setting.

(1) If $s_1 \geq \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$ *or* $s_1 \leq \min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, *then the Pareto frontier consists of a single point* $(0, \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)})$.

(2) Otherwise, if $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, *then the Pareto frontier of conference error and adversary error is a line segment of slope* 1 *starting from the origin* $(0, 0)$ *to* $\left(\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\right)$.

The proof of Theorem 4.2 is provided in Appendix C.2. The Pareto frontier established in Theorem 4.2 is illustrated in Figure 1.

We now unpack the result of Theorem 4.2, beginning with part (1). Recall that in this noiseless setting, given scores $S = [s_1, s_2]$ and knowing the reviewers' miscalibration functions, the conference can estimate the qualities of papers under each assignment. We use $\theta_i \in \mathbb{R}$ to denote the estimated quality of paper i. If the conference estimates the qualities assuming that A_1 was the actual assignment, we get $\theta_1 = \beta_1^{-1}(s_1)$ and $\theta_2 = \beta_2^{-1}(s_2)$. If the conference estimates the qualities assuming that A_2 was the actual assignment, we get $\theta_1 = \beta_2^{-1}(s_1)$ and $\theta_2 = \beta_1^{-1}(s_2)$. If $s_1 \geq \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, then $\theta_1 \geq \theta_2$ under both assignments (and hence paper 1 should be accepted). Similarly, if $s_1 \leq \min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, then $\theta_1 \leq \theta_2$ under both assignments (and hence paper 2 should be accepted). Therefore, under the condition of part (1) of the theorem, the same paper has higher estimated quality under both assignments, and hence that paper will be accepted irrespective of the function h. Thus, under this condition, the Pareto optimal curve comprises just a single point where the conference has zero error, and the adversary obtains no additional information from the acceptance decision as compared to the scores $S = [s_1, s_2]$. The error of the adversary is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ when it guesses the assignment using only the scores and not the decision.

Let us now discuss part (2) of Theorem 4.2. For scores $S = [s_1, s_2]$ that do not satisfy the condition in part (1), the conference would accept different papers when performing MAP calibration under the two possible assignments. In this

case, the function h does influence the outcomes. The Pareto frontier includes the origin since the conference can ensure zero error in this noiseless setting, but this zero-error acceptance decision will also perfectly reveal the assignment to the adversary since the zero-error decisions would be different under the two assignments. Then in the proof, we find the *maximum* per-instance error of the adversary given per-instance error of the conference. We find that the adversary's error no longer increases if the conference is allowed an error greater than $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. At this value of the conference's error, the maximum per-instance error of the adversary is also $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. We further show in the proof of the theorem that the Pareto frontier is precisely the line segment joining these two points. Therefore, the Pareto frontier for scores satisfy the condition is a line segment from the origin to the point $\left(\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\right)$ as shown in Figure 1.

Optimal Calibration Strategy under Per-Instance Errors In the previous section, we characterized the fundamental tradeoff between the conference's per-instance error and the adversary's per-instance error through the Pareto frontier. In this section, we design an explicit calibration strategy that achieves per-instance errors on the Pareto frontier, and is thus optimal for per-instance errors.

Since S is a fixed realization in the analysis of per-instance errors, to simplify the notation we define

$$q_1 = h(S, A_1) \qquad \text{and} \qquad q_2 = h(S, A_2).$$

Under this notation, q_1 is the probability with which the conference calibrates under the true assignment when the true assignment is A_1, and q_2 is the probability with which the conference calibrates under the true assignment when the true assignment is A_2. Therefore, from Proposition 4.1, given the maximum allowable error of the conference $\mathcal{E}_C$, our goal is to find values of q_1 and q_2 that are Pareto optimal. We present our proposed algorithm for this setting as Algorithm 1.

Theorem 4.3. *The calibration algorithm described in Algorithm 1 ensures the maximum per-instance error of the adversary for any given value of the maximum allowable per-instance error $\mathcal{E}_C([s_1, s_2])$ for the conference, and is hence Pareto optimal.*

The proof of Theorem 4.3 is presented in Appendix C.3. If $s_1 \geq \max\{\beta_1(\beta_2^{-1}(s_2)), \beta_2(\beta_1^{-1}(s_2))\}$ or $s_1 \leq \min\{\beta_1(\beta_2^{-1}(s_2)), \beta_2(\beta_1^{-1}(s_2))\}$, we are in part (1) of Theorem 4.2. Under scores that satisfy this condition, the conference is guaranteed to accept the higher-quality paper and thus has zero error. The error of the adversary is also fixed because the adversary makes its guess based on the scores only.

Otherwise, for a Pareto optimal calibration strategy, the errors of the conference and the adversary should stay on the Pareto frontier as in Figure 1. When $\mathcal{E}_C([s_1, s_2]) \geq \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, the conference should choose q_1 and q_2 such that its per-instance error is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ since further sacrifice of accuracy cannot increase the per-instance error of the adversary as indicated by the Pareto frontier. On the other hand, if $\mathcal{E}_C([s_1, s_2]) < \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, the conference can choose q_1 and q_2 that yields the maximum allowable per-instance error. Since $\mathcal{E}_C([s_1, s_2]) < \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ and $\mathcal{E}_C([s_1, s_2]) = 1 - \frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we can conclude that $f_1(s_1)f_2(s_2)q_1 + f_2(s_1)f_1(s_2)q_2 > \max\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}$ and the adversary has the same per-instance error under this condition. Therefore, in Algorithm 1, the error of the adversary is the same as the error of the conference and the errors are always on the Pareto frontier.

Algorithm 1: Conference calibration with per-instance error in the noiseless setting

Input: scores $S = [s_1, s_2]$, maximum allowable per-instance error of the conference $\mathcal{E}_C([s_1, s_2])$

if $s_1 \geq \max\{\beta_1(\beta_2^{-1}(s_2)), \beta_2(\beta_1^{-1}(s_2))\}$ **then**
  accept paper 1
else if $s_1 \leq \min\{\beta_1(\beta_2^{-1}(s_2)), \beta_2(\beta_1^{-1}(s_2))\}$ **then**
  accept paper 2
else if $\mathcal{E}_C([s_1, s_2]) \geq \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ **then**
  choose $q_1, q_2 \in [0, 1]$ such that $f_1(s_1)f_2(s_2)q_1 + f_2(s_1)f_1(s_2)q_2 = \max\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}$
else
  choose $q_1, q_2 \in [0, 1]$ such that $\mathcal{E}_C([s_1, s_2]) = 1 - \frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$
end if

Optimal Calibration Strategy under Average-case Error In the previous section, we designed an optimal strategy under per-instance errors. In this section, we design a calibration strategy that achieves optimal average-case errors for the conference with respect to the average-case error of the adversary. Unlike for per-instance error, we do not have a closed form expression for average-case error. We present our proposed algorithm as Algorithm 2. We now present our main result of this subsection, following which we discuss more details of the algorithm this result.

Theorem 4.4. *The calibration algorithm described in Algorithm 2 ensures the maximum average-case error of the adversary for any given value of the maximum allowable average-case error $\mathcal{E}_C$ for the conference, and is hence Pareto optimal.*

The proof of Theorem 4.4 is provided in Appendix C.4.

In Algorithm 2, running Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$ is a strategy that yields no error when the same paper has higher estimated quality under both assignments and otherwise, error of the conference equals error of the adversary. Moreover, both per-instance error of the conference and per-instance error of the adversary is

Algorithm 2: Conference calibration with average-case error in the noiseless setting

Input: maximum allowable average-case error of the conference $\mathcal{E}_C$
Let ζ = error of the conference for adopting Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$ for all $[s_1, s_2]$
if $\mathcal{E}_C > \zeta$ **then**
    the desired conference error is Pareto inefficient and operate at $\mathcal{E}_C = \zeta$
else if $\mathcal{E}_C = \zeta$ **then**
    run Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$
else if $\mathcal{E}_C < \zeta$ **then**
    toss a coin that has probability $\frac{\mathcal{E}_C}{\zeta}$ of head
    if coin toss outcome is head **then**
        run Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$
    else
        calibrate under true assignment
    end if
end if

$\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. That is, the maximum per-instance error for the adversary. Thus, this strategy is Pareto optimal for any score pair and is also Pareto optimal under its average-error since the error of the adversary is maximized under such average-error of the conference.

In proof for optimality of Algorithm 2, we take advantage of the fact that the Pareto frontier is either a point where the conference has no error or an increasing line with slope 1. Under this fact, the optimal average-case error of the conference is where the conference has zero error when the adversary guesses the assignment based on the scores only and has the same error as the adversary otherwise. Therefore, Algorithm 2 makes use of Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$ and on average, the error of the conference and the adversary matches the Pareto optimality for the conference.

4.3 Noisy Setting

We now study the noisy setting. We consider both reviewers' miscalibration functions β_1 and β_2 to be affine and both reviewers' noises ϵ_1 and ϵ_2 to be Gaussian. Furthermore, the distributions of the noise are the same for both reviewers with mean zero and some known variance σ^2. Formally, we assume:

$$\beta_1(\theta^*) = a_1\theta^* + b_1, \quad \beta_2(\theta^*) = a_2\theta^* + b_2,$$
$$\epsilon_1 \sim N(0, \sigma^2), \quad \text{and} \quad \epsilon_2 \sim N(0, \sigma^2).$$

As we will see below, the presence of noise makes the analysis much more complex, even when we assume affine miscalibration, as compared to the noiseless setting.

Pareto Frontier for Per-Instance Errors We begin by establishing the Pareto frontier for per-instance errors in the noisy case. Let Φ denote the cumulative distribution function of the standard normal distribution. Also define notation Φ_1 and Φ_2 as:

$$\Phi_1 = \Phi\left(\frac{a_2(a_1^2+\sigma^2)(s_2-b_2) - a_1(a_2^2+\sigma^2)(s_1-b_1)}{\sqrt{\sigma^2(a_1^2+a_2^2+2\sigma^2)(a_1^2+\sigma^2)(a_2^2+\sigma^2)}}\right) \tag{4.1a}$$

$$\Phi_2 = \Phi\left(\frac{a_1(a_2^2+\sigma^2)(s_2-b_1) - a_2(a_1^2+\sigma^2)(s_1-b_2)}{\sqrt{\sigma^2(a_1^2+a_2^2+2\sigma^2)(a_1^2+\sigma^2)(a_2^2+\sigma^2)}}\right). \tag{4.1b}$$

Theorem 4.5. *Consider the peer-review system in the noisy setting. The Pareto frontier of (per-instance error of the conference, per-instance error of the adversary) with scores $S = [s_1, s_2]$ is as follows.*

(1) If $s_1 \geq \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$ *or* $s_1 \leq \min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, *then the Pareto frontier consists of a single point.*
Specifically, when $s_1 \geq \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, *the Pareto frontier of conference error and adversary error is the point* $\left(\frac{f_1(s_1)f_2(s_2)\Phi_1+f_2(s_1)f_1(s_2)\Phi_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\right)$.
And similarly, when $s_1 \leq \min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, *the Pareto frontier is the point* $\left(\frac{f_1(s_1)f_2(s_2)(1-\Phi_1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\right)$

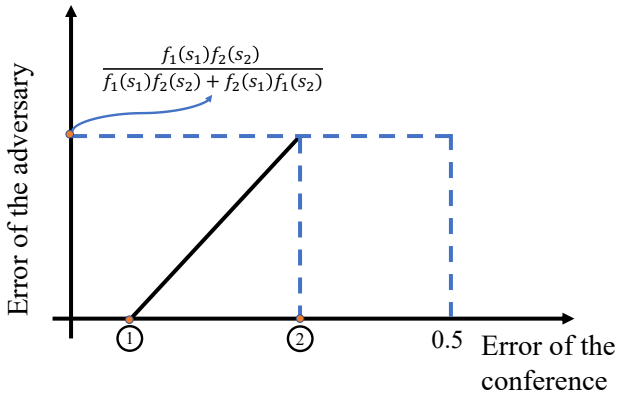
(2) If $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} < s_1 < \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, *then the Pareto frontier is an increasing line.*

The proof of Theorem 4.5 is provided in Appendix C.5. We now unpack this result and specify precisely the Pareto frontier in both parts of the theorem.

Given scores $S = [s_1, s_2]$ and knowing the reviewers' miscalibration functions, the conference can estimate the qualities of papers under each assignment. Under assignment A_1, we have $\Pr(\theta_1^* > \theta_2^* | \mathcal{A} = A_1, S = [s_1, s_2]) = 1-\Phi_1$. And under assignment A_2, we have $\Pr(\theta_1^* > \theta_2^* | \mathcal{A} = A_2, S = [s_1, s_2]) = 1 - \Phi_2$.

Let us now consider part (1) of Theorem 4.5. If the condition specified in the statement of the theorem is satisfied, then we know that either $(\Phi_1 \leq \frac{1}{2}, \Phi_2 \leq \frac{1}{2})$ in which case paper 1 has a higher estimated quality under either assignment, or $(\Phi_1 \geq \frac{1}{2}, \Phi_2 \geq \frac{1}{2})$ in which paper 2 has a higher estimated quality under either assignment. Thus, if the condition in part (1) is met, the same paper has higher estimated quality under both assignments and hence the decision does not depend on h. Thus, for such pair of scores, the Pareto optimal situation is where the conference has minimum error and the adversary guesses the assignment based on the scores alone.

Let us now move to part (2) of Theorem 4.5, and consider parameters that satisfy the condition stated therein. Under this condition, the conference would accept different papers



$$①: \frac{f_1(s_1)f_2(s_2)\Phi_1 + f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$$

$$②: \frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1) + f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$$

Figure 2: A Pareto frontier in the noisy setting of part (2) of Theorem 4.5 in the case that $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$ and $\Phi_1 < \frac{1}{2} < \Phi_2$ with $0 < \Phi_2 - \frac{1}{2} < \frac{1}{2} - \Phi_1$. The notations Φ_1 and Φ_2 are defined in (4.1).

by calibrating under different assignments, and the function h needs to be carefully designed. We study the Pareto frontier for scores in the range.

We consider a specific case where $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$, $\Phi_1 < \frac{1}{2}$ and $\Phi_2 > \frac{1}{2}$ with $0 < \Phi_2 - \frac{1}{2} < \frac{1}{2} - \Phi_1$. All other cases can be derived in a similar fashion to the proof in Appendix C.5. The Pareto frontier with these assumptions are shown in Figure 2. We first find the maximum per-instance error of the adversary given per-instance error of the conference. We find that the adversary's error no longer increases if the conference increase its error larger than $\frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ in this case. The maximum per-instance error of the adversary is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. Therefore, the Pareto frontier for scores satisfy the condition is an increasing line from $\left(\frac{f_1(s_1)f_2(s_2)\Phi_1+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, 0\right)$ to the point $\left(\frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\right)$.

We show the Pareto frontier in the case described above in Figure 2. In all other cases, the shape of the Pareto frontier is the same as Figure 2 but has different coordinates. The relationship between $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$ combining with the values of Φ_1 and Φ_2 and their distance to $\frac{1}{2}$, we have eight different combinations of these values. In all eight cases, the Pareto frontier contains either a single point or an increasing line depending on the scores. Moreover, the maximum error of the adversary is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ in all cases.

Optimal Calibration Strategy under Per-Instance Errors In the previous section, we characterized the fundamental tradeoff between the conference's per-instance error and the adversary's per-instance error through the Pareto frontier. In this section, we design a calibration strategy that achieves per-instance errors on the Pareto frontier, meaning that the strategy is optimal under per-instance errors.

Since S is a fixed realization in the analysis of per-instance errors, to simplify the notation we define (similar to Section 4.2):

$$q_1 = h(S, A_1) \qquad \text{and} \qquad q_2 = h(S, A_2).$$

Under this notation, given the maximum allowable error of the conference $\mathcal{E}_C$, our goal is to find values of q_1 and q_2 that maximize the error of the adversary $\mathcal{E}_A$. We continue to use the notations Φ_1 and Φ_2 introduced in 4.1.

We present our proposed algorithm as Algorithm 3.

Theorem 4.6. *The calibration algorithm described in Algorithm 3 ensures the maximum per-instance error of the adversary for any given value of the maximum allowable per-instance error $\mathcal{E}_C([s_1, s_2])$ for the conference, and is hence Pareto optimal.*

The proof of Theorem 4.6 is provided in Appendix C.6. For a moment, consider the case of $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$ and $\Phi_1 < \frac{1}{2} < \Phi_2$ with $0 < \Phi_2 - \frac{1}{2} < \frac{1}{2} - \Phi_1$, for a Pareto optimal calibration strategy, the error of the conference and the adversary should stay on the Pareto frontier as in Figure 2. If the required error of the conference is less than $\frac{f_1(s_1)f_2(s_2)\Phi_1+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, then due to the noise, there is no feasible calibration strategy that satisfies this requirement. Otherwise, the error of the conference and the error of the adversary adhere to the Pareto frontier.

Algorithm 3 follows directly from the Pareto frontier established in Theorem 4.5. The calibration probabilities q_1 and q_2 are chosen such that the error of the conference and the error of the adversary lie on the Pareto frontier. The first two cases of Algorithm 3 correspond to part (1) of Theorem 4.5 where the same paper has higher estimated quality under both assignments. In the noisy case, there is a minimum value for the per-instance error of the conference. Therefore, in the third case of Algorithm 3, when the maximum allowable per-instance error of the conference is too small, the conference cannot achieve such error. If $\mathcal{E}_C([s_1, s_2]) \geq \frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, the error of the conference should be $\frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ to stay Pareto optimal since further sacrifice of accuracy cannot increase error of the adversary. And for the rest of the per-instance error of the conference, we choose q_1 and q_2 such that the errors of the conference and the adversary stay on the Pareto frontier in Figure 2.

5 Discussion

Our work is only a starting point towards addressing the important problem of calibration with privacy in its full gener-

Algorithm 3: Conference calibration with per-instance error in the noisy setting

Input: scores $S = [s_1, s_2]$, maximum allowable per-instance error of the conference $\mathcal{E}_C([s_1, s_2])$

if $s_1 > \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$ **then**

accept paper 1

else if $s_1 < \min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$ **then**

accept paper 2

else if $\mathcal{E}_C([s_1, s_2]) < \frac{f_1(s_1)f_2(s_2)\Phi_1 + f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ **then**

error of the conference cannot be achieved

else if $\mathcal{E}_C([s_1, s_2]) \geq \frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ **then**

choose $q_1 = 1, q_2 = \frac{(f_2(s_1)f_1(s_2)-f_1(s_1)f_2(s_2))(1-2\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$

else

choose $q_1 = 1, q_2 = \frac{\mathcal{E}_C([s_1,s_2])\cdot(f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2))}{(1-2\Phi_2)f_2(s_1)f_1(s_2)} - \frac{f_1(s_1)f_2(s_2)(1-\Phi_1)+f_2(s_1)f_1(s_2)\Phi_2+(2\Phi_1-1)f_1(s_1)f_2(s_2)}{(1-2\Phi_2)f_2(s_1)f_1(s_2)}$

end if

ality. Several challenges need to be addressed in future work in order to design practical algorithms with guarantees for calibration and privacy. There are open problems pertaining to relaxations of assumptions made in this paper such as that of two reviewers and papers, homogeneity and knowledge of the noise variance, etc. An important open problem pertains to challenge #2 discussed in the introduction, in conjunction with challenge #1. Instead of assuming precise exogenous knowledge of the reviewers' miscalibration functions, consider having some access to data from other conferences. Then how can one obtain and use meaningful estimates of reviewer miscalibrations from past conferences while guaranteeing privacy of the current as well as past conferences ("federated learning for calibration")? In any of these endeavors, one may aim to uncover precise fundamental limits and optimal algorithms, or perhaps design algorithms that are readily applicable in practice with some basic theoretical guarantees.

Acknowledgments

This work was supported by NSF CAREER award 1942124, NSF grant CIF 1763734, an NSERC Discovery Grant, and a Google Research Scholar Award. Most of this research was done when Wenxin Ding was at Carnegie Mellon University.

References

Ammar, A.; and Shah, D. 2012. Efficient rank aggregation using partial data. In *SIGMETRICS*.

Baba, Y.; and Kashima, H. 2013. Statistical Quality Estimation for General Crowdsourcing Tasks. In *KDD*.

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.

Bharadhwaj, H.; Turpin, D.; Garg, A.; and Anderson, A. 2020. De-anonymization of authors through arXiv submissions during double-blind review. *arXiv preprint arXiv:2007.00177*.

Brenner, L.; Griffin, D.; and Koehler, D. J. 2005. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, 97(1): 64–81.

Charlin, L.; and Zemel, R. S. 2013. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*.

Chawla, D. S. 2021. Swiss funder draws lots to make grant decisions. *Nature*.

Dhull, K.; Jecmen, S.; Kothari, P.; and Shah, N. B. 2022. The Price of Strategyproofing Peer Assessment. *arXiv:2201.10631*.

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*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2016. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3): 17–51.

Evfimievski, A.; Gehrke, J.; and Srikant, R. 2003. Limiting Privacy Breaches in Privacy Preserving Data Mining. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, 211–222. New York, NY, USA: ACM.

Fiez, T.; Shah, N.; and Ratliff, L. 2020. A SUPER* Algorithm to Optimize Paper Bidding in Peer Review. In *Conference on Uncertainty in Artificial Intelligence*.

Flach, P.; Spiegler, S.; Golénia, B.; Price, S.; Guiver, J.; Herbrich, R.; Graepel, T.; and Zaki, M. 2010. Novel Tools to Streamline the Conference Review Process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*, 11(2): 63–67.

Freund, Y.; Iyer, R. D.; Schapire, R. E.; and Singer, Y. 2003. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, 4: 933–969.

Garg, N.; Kavitha, T.; Kumar, A.; Mehlhorn, K.; and Mestre, J. 2010. Assigning Papers to Referees. *Algorithmica*, 58(1): 119–136.

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

Goldsmith, J.; and Sloan, R. H. 2007. The AI conference paper assignment problem. In *Proc. AAAI Workshop on Preference Handling for Artificial Intelligence, Vancouver*, 53–57.

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 *NeurIPS*.

Kang, D.; Ammar, W.; Dalvi, B.; van Zuylen, M.; Kohlmeier, S.; Hovy, E.; and Schwartz, R. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*.

Kasiviswanathan, S. P.; Lee, H. K.; Nissim, K.; Raskhodnikova, S.; and Smith, A. 2011. What Can We Learn Privately? *SIAM Journal on Computing*, 40(3): 793–826.

Kobren, A.; Saha, B.; and McCallum, A. 2019. Paper Matching with Local Fairness Constraints. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Langford, J. 2012. ICML acceptance statistics. http://hunch.net/?p=2517 [Online; accessed 14-May-2021].

Lee, C. J. 2015. Commensuration bias in peer review. *Philosophy of Science*, 82(5): 1272–1283.

Littman, M. L. 2021. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6): 43–44.

Liu, M.; Choy, V.; Clarke, P.; Barnett, A.; Blakely, T.; and Pomeroy, L. 2020. The acceptability of using a lottery to allocate research funding: A survey of applicants. *Research integrity and peer review*, 5(1): 1–7.

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 *AAAI*.

Meir, R.; Lang, J.; Lesca, J.; Kaminsky, N.; and Mattei, N. 2020. A market-inspired bidding scheme for peer review paper assignment. In *Games, Agents, and Incentives Workshop at AAMAS*.

Mimno, D.; and McCallum, A. 2007. Expertise Modeling for Matching Papers with Reviewers. In *KDD*.

Mitliagkas, I.; Gopalan, A.; Caramanis, C.; and Vishwanath, S. 2011. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference*.

Noothigattu, R.; Shah, N.; and Procaccia, A. 2021. Loss Functions, Axioms, and Peer Review. *Journal of Artificial Intelligence Research*.

Paul, S. R. 1981. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34(2): 213–223.

Philipps, A. 2021. Research funding randomly allocated? A survey of scientists' views on peer review and lottery. *Science and Public Policy*.

Roos, M.; Rothe, J.; Rudolph, J.; Scheuermann, B.; and Stoyan, D. 2012. A statistical approach to calibrating the scores of biased reviewers: The linear vs. the nonlinear model. In *Multidisciplinary Workshop on Advances in Preference Handling*.

Roos, M.; Rothe, J.; and Scheuermann, B. 2011. How to Calibrate the Scores of Biased Reviewers by Quadratic Programming. In *AAAI Conference on Artificial Intelligence*.

Shah, N.; Tabibian, B.; Muandet, K.; Guyon, I.; and Von Luxburg, U. 2018. Design and Analysis of the NIPS 2016 Review Process. *JMLR*, 19(1): 1913–1946.

Shah, N. B. 2021. Systemic Challenges and Solutions on Bias and Unfairness in Peer Review. Preprint http://www.cs.cmu.edu/~nihars/preprints/Shah_Survey_PeerReview.pdf.

Siegelman, S. S. 1991. Assassins and zealots: Variations in peer review. *Radiology*, 178(3): 637–642.

Soergel, D.; Saunders, A.; and McCallum, A. 2013. Open Scholarship and Peer Review: A Time for Experimentation.

Stelmakh, I.; Shah, N.; and Singh, A. 2019. On Testing for Biases in Peer Review. In *NeurIPS*.

Stelmakh, I.; Shah, N.; and Singh, A. 2020. Catch Me if I Can: Detecting Strategic Behaviour in Peer Assessment. *arXiv*.

Stelmakh, I.; Shah, N.; and Singh, A. 2021. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. *JMLR*.

Stelmakh, I.; Shah, N.; Singh, A.; and Daumé III, H. 2021. Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review. In *CSCW*.

Taylor, C. J. 2008. On the optimal assignment of conference papers to reviewers.

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.

Tran, D.; Valtchanov, A.; Ganapathy, K.; Feng, R.; Slud, E.; Goldblum, M.; and Goldstein, T. 2020. An Open Review of OpenReview: A Critical Analysis of the Machine Learning Conference Review Process. *arXiv preprint arXiv:2010.05137*.

Wang, J.; and Shah, N. B. 2019. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In *AAMAS*.

Warner, S. L. 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309): 63–69.

Wu, R.; Guo, C.; Wu, F.; Kidambi, R.; van der Maaten, L.; and Weinberger, K. 2021. Making Paper Reviewing Robust to Bid Manipulation Attacks. *arXiv:2102.06020*.

Xu, Y.; Zhao, H.; Shi, X.; and Shah, N. 2019. On Strategyproof Conference Review. In *IJCAI*.

Yuan, W.; Liu, P.; and Neubig, G. 2021. Can We Automate Scientific Reviewing? *arXiv preprint arXiv:2102.00176*.

Appendices

A Simulations: Correcting miscalibration with and without exogeneous information

In the introduction section in the main text, we discussed two ways of reducing miscalibration: one where only the current conferences' data is used and another where miscalibration parameters of reviewers are obtained exogenously (e.g., from previous conferences). In this section, we conduct a simulation-based study to understand the performance of these approaches: What is the reduction in error if correcting for miscalibration? What if the reviewer-calibration parameters are known?

Our main results in the main text was focused on privacy and considered a setting with two reviewers and two papers. In this section we consider a larger number of reviewers and papers. The methods we simulate in this more general setting do not consider privacy.

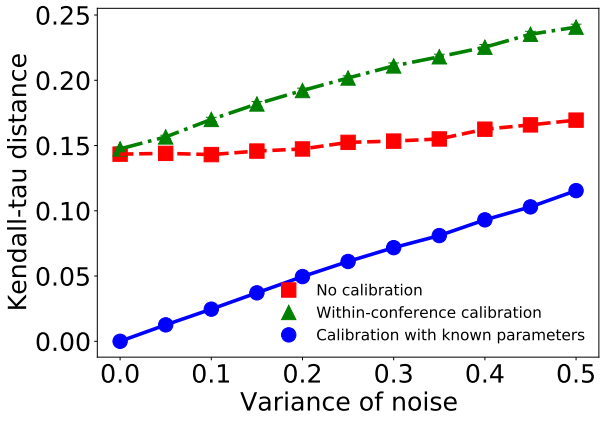
We first describe the simulation setting, and then discuss the results. The code for the simulations is available here: https://github.com/wenxind/calibration-with-privacy-in-peer-review.

Conference review setup: We consider 100 reviewers and 100 papers. We assign reviewers to papers uniformly at random with 3 reviewers per paper and 3 papers per reviewer.

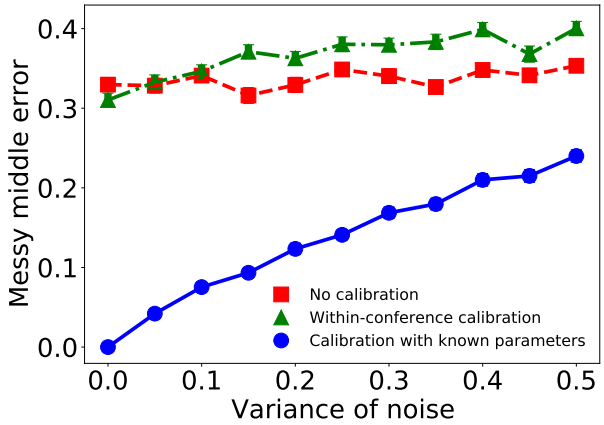
Miscalibration model: We assume each reviewer has a linear miscalibration function: the miscalibration function h_j of reviewer j is given by $h_j(\theta^*) = a_j\theta^* + b_j$ where θ^* is the true quality of the paper being reviewed. For every paper i, its true quality θ_i^* is drawn from a Gaussian distribution with mean 0 and variance 1 independent of all else. The scalars a_j in the reviewers' miscalibration functions are i.i.d. exponential random variables with rate 1. The biases b_j are i.i.d. Gaussian random variables with mean 0 and variance 0.5. The score given by any reviewer j to any paper i is then given as θ_i^* is $a_j\theta_i^* + b_j + \epsilon_{ij}$ where ϵ_{ij} is a Gaussian random variable with mean 0, whose variance is varied in the plots.

Calibration methods: We consider the following three methods to calibrate the decisions.

- No calibration: The score for each paper is the mean score of the 3 review scores.
- Within-conference calibration: For each reviewer, we compute the mean score and standard deviation of the 3 review scores given by the reviewer and normalize the 3 scores by subtracting mean and dividing standard deviation. Then the score for each paper is the mean score of the 3 normalized review scores.
- Calibration with known parameters: We assume that the miscalibration parameters of the reviewers are known (exogeneously). Then we estimate the quality of each paper via maximum likelihood estimation as follows. For any paper i, let $\mathcal{R}_i$ denote the set of reviewers for paper i. Then the estimate of the score for any paper i is $\frac{\sum_{j\in\mathcal{R}_i} a_j(s_{ji}-b_j)}{\sum_{j\in\mathcal{R}_i} a_j^2}$ where s_{ji} denotes the score given by reviewer j to paper i.



(a) Kendall-τ distance.



(b) Messy middle error.

Figure 3: A simulation of the review process where the reviewers are miscalibrated.

For each paper, we then take a mean of the calibrated scores across all its reviewers. The papers are then ranked according to these mean scores; we call this the estimated ranking.

Error metrics: We consider two ways of measuring the error between the ranking of the papers in terms of their true scores and the ranking of the papers in terms of their estimated scores.

- Kendall tau distance: Given two rankings of the papers, the Kendall tau distance between the two ranking is $\frac{\text{number of discordant pairs}}{\text{total number of pairs}}$.
- Messy middle error: Given two rankings of the 100 papers, suppose that the conference wishes to accept the top 25 papers. Then we consider papers 11-40 as those that are marginal accepts, and we measure the error as the fraction of these papers which are (erroneously) rejected. In other words, the messy middle error equals $\frac{\text{number of papers whose true ranking is between 11–40 that are wrongly accepted/rejected}}{30}$.

Results: The results of the simulations are shown in Fig-

ure 3. Each point depicts the mean from 100 iterations of these simulations. The error bars are too small to be visible. We see that correcting for miscalibration even without access to the parameters can lead to significant reduction in the error as compared to not correcting for the miscalibration. Furthermore, if the parameters were known (e.g., from other conferences) then it can lead to multi-fold further reductions in the error.

B Connection to Local Differential Privacy

In this section, we discuss the connections between our algorithm and differential privacy (DP). We recall the definition of DP:

Definition B.1 ((Dwork et al. 2016))**.** An algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is ϵ-differentially private (DP) if, for all $X, X' \in \mathcal{X}^n$ which differ in one entry (often called *neighboring databases*) and $\mathcal{S} \subseteq Y$, we have that

$$\Pr[M(X) \in \mathcal{S}] \leq e^{\epsilon} \Pr[M(X') \in \mathcal{S}].$$

Roughly speaking, a procedure involving n users is ϵ-*locally differentially private* if each user applies an ϵ-DP algorithm to their single datapoint and shares only the result with other users or a data curator. The most familiar LDP algorithm is (binary) randomized response.

Definition B.2. Binary randomized response with parameter γ is an algorithm $M : \{0,1\} \to \{0,1\}$, which, given input x, outputs x with probability $\frac{1}{1+\gamma}$, and outputs $1-x$ with probability $\frac{\gamma}{1+\gamma}$.

The following claim is immediate from the definition of differential privacy and randomized response.

Proposition B.3. *Binary randomized response with parameter e^{ϵ} is ϵ-DP.*

We now relate randomized response to the algorithms proposed in our setting. The private information for our calibration problem consists solely of the reviewer assignment A, which takes one of two different values (i.e., reviewer 1 is assigned to paper 1 and reviewer 2 is assigned to paper 2, or vice versa). These two reviewer assignments can be considered to be "neighboring" datasets, as mentioned in Definition B.1. All other information (paper scores S and reviewer's miscalibration functions) are assumed to be public.

As argued in Proposition 4.1, it is without loss of optimality to solely consider strategies of the form h, in which the conference calibrates according to the true assignment with probability $h(S, A)$ and according to the false assignment otherwise. This can be rephrased into the language of randomized response by considering the assignment A to be the input bit to binary randomized response, in which it is preserved with probability $h(S, A)$ (in the language of Definition B.2, $h(S, A) = \frac{1}{1+\gamma}$) and flipped otherwise, and then the conference calibrates with the resulting assignment.

We caution that this connection does *not* directly imply that our algorithms are differentially private. This is because the probability $h(S, A)$ is selected in a data-dependent way, whereas differential privacy requires it to be data independent. Nevertheless, our work provides tight guarantees on the probability that an MAP adversary can determine the true assignment.

C Appendix: Proofs

In the appendix, we present complete proofs of the results claimed in the main text.

C.1 Proof of Proposition 4.1

The most generic calibration strategy can be represented using a function g such that for any given score and assignment, g outputs a probability for accepting paper 1. In other words, the conference accepts paper 1 with probability $g(S, A)$ and accepts paper 2 with probability $1 - g(S, A)$. We propose a calibration strategy using function h instead of g, where h outputs a probability that the conference calibrates under the true assignment by accepting the paper with higher estimated quality under the true assignment.

Calibrating using the calibration strategy of function h differs from calibrating using the calibration strategy of function g only when the same paper has higher estimated quality under both assignments by the MAP. Since otherwise, when calibrating under the true assignment and calibrating assuming the wrong assignment lead to accepting different papers, either paper can have arbitrary non-zero probability of being accepted (their probabilities sum to 1) by adjusting the output of $h(S, A_1)$ and $h(S, A_2)$. Then it is the same calibration strategy as using function g.

Note that the adversary makes its guess using the MAP $\operatorname{argmax}_{\{A=A_1 \text{or} A=A_2\}} \Pr(\mathcal{A} = A | \boldsymbol{D} = P, S = [s_1, s_2])$ where $\boldsymbol{D}$ is the random variable for the decision made by the conference (acceptance of paper) and P is the paper being accepted. By expanding the probability expression, we have that

$$\begin{aligned}
&\operatorname*{argmax}_{A \in \{A_1, A_2\}} \Pr(\mathcal{A} = A | \boldsymbol{D} = P, S = [s_1, s_2]) \\
&= \operatorname*{argmax}_{A \in \{A_1, A_2\}} \frac{\Pr(\mathcal{A} = A, \boldsymbol{D} = P | S = [s_1, s_2])}{\Pr(\boldsymbol{D} = P | S = [s_1, s_2])} \\
&= \operatorname*{argmax}_{A \in \{A_1, A_2\}} \frac{\Pr(\boldsymbol{D} = P | \mathcal{A} = A, S = [s_1, s_2]) \Pr(\mathcal{A} = A | S = [s_1, s_2])}{\Pr(\boldsymbol{D} = P | S = [s_1, s_2])} \\
&= \operatorname*{argmax}_{A \in \{A_1, A_2\}} \Pr(\boldsymbol{D} = P | \mathcal{A} = A, S = [s_1, s_2]) \Pr(\mathcal{A} = A | S = [s_1, s_2]).
\end{aligned}$$

If the same paper has higher estimated quality under both assignments, and the conference accepts the believed higher-quality paper, then the adversary guesses the assignment based on the scores only. Because the adversary knows the calibration strategy used by the conference, if P is the paper that has higher quality under both assignments, then $\Pr(\boldsymbol{D} = P | \mathcal{A} = A, S = [s_1, s_2]) = 1$ for both $\mathcal{A} = A_1$ and $\mathcal{A} = A_2$. Therefore, the MAP used by the adversary simplifies to $\operatorname{argmax}_{A=A_1 \text{or} A=A_2} \Pr(\mathcal{A} = A | S = [s_1, s_2])$. In this case, the conference does not have extra privacy leakage by accepting P since the adversary is making its guess based on the information that is already public (the scores). In addition, if the conference has non-zero probability of accepting the other paper, its utility decreases by accepting the lower-quality paper. Even if the conference accepts the lower-quality paper, the error of the adversary remains unchanged as it can use the scores to guess the assignment without being affected by the conference decision. Thus,

there is no need for the conference to have non-zero probability for accepting the paper that has lower-quality under both assignments.

In conclusion, calibrating using the calibration strategy of function h instead of the calibration strategy of function g does not reduce the optimally of the conference. Therefore, we consider the calibration strategy with function h in our analysis.

C.2 Proof of Theorem 4.2

To find the Pareto frontier of per-instance error of the adversary against per-instance error of the conference, we first derive expressions for per-instance error of the conference and the adversary. We find a fixed expression for the error of the conference and calculate the maximum per-instance error of the adversary in different cases. We then analyze the relation between the errors and complete plots for maximum per-instance error of the adversary for any per-instance error of conference. Finally, we derive the Pareto frontier from the plots.

In the noiseless setting, the conference uses the inverse functions of reviewers' miscalibration functions and the scores to exactly compute the quality of the papers. If the conference estimates the qualities assuming that A_1 was the actual assignment, we get $\theta_1 = \beta_1^{-1}(s_1)$ and $\theta_2 = \beta_2^{-1}(s_2)$. If the conference estimates the qualities assuming that A_2 was the actual assignment, we get $\theta_1 = \beta_2^{-1}(s_1)$ and $\theta_2 = \beta_1^{-1}(s_2)$. If $s_1 > \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, then $\theta_1 > \theta_2$ under both assignments (and hence paper 1 should be accepted). Similarly, if $s_1 < \min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, then $\theta_1 < \theta_2$ under both assignments and hence paper 2 should be accepted. Therefore, when $s_1 > \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$ or $s_1 < \min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, which is a subset of part (1) of Theorem 4.2, the same paper has higher estimated quality under both assignments, and hence that paper will be accepted irrespective of the function h. Thus, under this condition, the Pareto optimal curve comprises just a single point where the conference has zero error, and the adversary obtains no additional information from the acceptance decision as compared to the scores $S = [s_1, s_2]$. The error of the adversary is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ when it guesses the assignment using only the scores and not the decision.

For the rest scores, the conference uses function h to decide acceptance of paper. Since S is a fixed realization in the analysis, we simplify the calibration strategy for the conference as

$$q_1 = h(S, A_1)$$
$$q_2 = h(S, A_2).$$

We now consider the rest scores in part (1) of the theorem. If $s_1 = \beta_2(\beta_1^{-1}(s_2))$, the conference accepts each paper uniform at random if calibrating under A_1 and accepts paper 1 if calibrating under A_2. Since paper 1 has higher or equal quality than paper 2, the conference only has error when paper 2 is accepted and $\mathcal{A} = A_2$.

$$\begin{aligned}
&\Pr(\text{conference accepts lower-quality paper}|S = [s_1, s_2]) \\
=&\Pr(\text{conference accepts lower-quality paper}|S = [s_1, s_2], D = P_1) \\
&\cdot \Pr(D = P_1|S = [s_1, s_2]) \\
&+ \Pr(\text{conference accepts lower-quality paper}|S = [s_1, s_2], D = P_2) \\
&\cdot \Pr(D = P_2|S = [s_1, s_2]) \\
=&\Pr(\text{conference accepts lower-quality paper}|S = [s_1, s_2], D = P_2) \\
&\cdot \Pr(D = P_2|S = [s_1, s_2]).
\end{aligned}$$

Given that the conference accepts P_2, the probability of the conference making error is the probability $\Pr(\mathcal{A} = A_2|S = [s_1, s_2])$.

$$\begin{aligned}
&\Pr(D = P_2|S = [s_1, s_2]) \\
=&\Pr(D = P_2|S = [s_1, s_2], \mathcal{A} = A_1)\Pr(\mathcal{A} = A_1|S = [s_1, s_2]) \\
&+ \Pr(D = P_2|S = [s_1, s_2], \mathcal{A} = A_2)\Pr(\mathcal{A} = A_2|S = [s_1, s_2]) \\
=&\frac{1}{2}q_1 \Pr(\mathcal{A} = A_1|S = [s_1, s_2]) + \frac{1}{2}(1 - q_2)\Pr(\mathcal{A} = A_2|S = [s_1, s_2]).
\end{aligned}$$

For the adversary, if paper 1 is accepted, it gains no information on the assignment other than the scores so its error is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. Otherwise, it guesses $\mathcal{A} = A_1$ and its error is $\Pr(\mathcal{A} = A_2|S = [s_1, s_2])$. Note that error of the adversary does not exceed $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ since in the worst case for the adversary, it guesses the assignment solely based on the scores and ignore the conference decision.

$$\begin{aligned}
&\Pr(\text{adversary guesses assignment wrong}|S = [s_1, s_2]) \\
=&\Pr(\text{adversary guesses assignment wrong}|S = [s_1, s_2], D = P_1) \\
&\cdot \Pr(D = P_1|S = [s_1, s_2]) \\
&+ \Pr(\text{adversary guesses assignment wrong}|S = [s_1, s_2], D = P_2) \\
&\cdot \Pr(D = P_2|S = [s_1, s_2]) \\
=&\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2) + f_2(s_1)f_1(s_2)} \\
&\cdot \Big((1 - \frac{1}{2}q_1)\Pr(\mathcal{A} = A_1|S = [s_1, s_2]) \\
&+ (\frac{1}{2} + \frac{1}{2}q_2)\Pr(\mathcal{A} = A_2|S = [s_1, s_2])\Big) \\
&+ \Pr(\mathcal{A} = A_2|S = [s_1, s_2]) \\
&\cdot \Big(\frac{1}{2}q_1 \Pr(\mathcal{A} = A_1|S = [s_1, s_2]) \\
&+ \frac{1}{2}(1 - q_2)\Pr(\mathcal{A} = A_2|S = [s_1, s_2])\Big)
\end{aligned}$$

Therefore, we can minimize the error of the conference to 0 by choosing $q_1 = 0$ and $q_2 = 1$, which results in the conference always accepts paper 1. Then error of the adversary is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, which is maximized. Further increase of error of the conference cannot increase error of the adversary. So the Pareto optimal point is $(0, \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)})$. The same argument works when $s_1 = \beta_1(\beta_2^{-1}(s_2))$.

In the noiseless setting where $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, which is part (2) of Theorem 4.2, we first find the maximum per-instance error of the adversary given per-instance error of the conference in this range. We will show the proof with the assumptions that $\beta_2(\beta_1^{-1}(s_2)) > \beta_1(\beta_2^{-1}(s_2))$ and $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$. The proof follows the same procedure for other values of $\beta_2(\beta_1^{-1}(s_2))$, $\beta_1(\beta_2^{-1}(s_2))$, $f_1(s_1)f_2(s_2)$, and $f_2(s_1)f_1(s_2)$.

When the scores satisfy $\beta_1(\beta_2^{-1}(s_2)) < s_1 < \beta_2(\beta_1^{-1}(s_2))$, the conference always accepts the higher-quality paper if it calibrates under the true assignment, and the conference always accepts the lower-quality paper if it calibrates assuming the wrong assignment. But the conference can calibrate assuming the wrong assignment for the purpose of misleading the adversary. We use $\mathcal{A}$ to denote the random variable for the assignment, $\boldsymbol{D}$ to denote the random variable for the conference decision and S is the scores. In addition, we use $\boldsymbol{C}$ to denote the calibration status. If the conference calibrates under the true assignment then $\boldsymbol{C} = T$. Otherwise, $\boldsymbol{C} = F$.

Therefore, the error of the conference is computed as

$$
\begin{aligned}
&\Pr(\text{conference accepts lower-quality paper}|S=[s_1,s_2])\\
=&\Pr(\boldsymbol{C}=F|S=[s_1,s_2])\\
=&\Pr(\boldsymbol{C}=F,\mathcal{A}=A_1|S=[s_1,s_2])\\
&+\Pr(\boldsymbol{C}=F,\mathcal{A}=A_2|S=[s_1,s_2])\\
=&\Pr(\boldsymbol{C}=F|\mathcal{A}=A_1,S=[s_1,s_2])\Pr(\mathcal{A}=A_1|S=[s_1,s_2])\\
&+\Pr(\boldsymbol{C}=F|\mathcal{A}=A_2,S=[s_1,s_2])P(\mathcal{A}=A_2|S=[s_1,s_2])\\
=&(1-q_1)\cdot\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\\
&+(1-q_2)\cdot\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\\
=&1-\frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}.
\end{aligned}
$$

The adversary uses MAP to guess the assignment. If the two assignments have the same a posteriori probability, then the adversary makes a random guess between the assignments where either assignment has probability $\frac{1}{2}$ of being guessed. When making a guess, the adversary observes the scores and the conference decision. So the adversary finds $\operatorname{argmax}_{\{A=A_1 \text{or} A=A_2\}}\Pr(\mathcal{A}=A|\boldsymbol{D}=P,S=[s_1,s_2])$ where P is the paper being accepted. Following Section C.1, the adversary finds

$$
\begin{aligned}
&\operatorname*{argmax}_{A\in\{A_1,A_2\}}\Pr(\mathcal{A}=A|\boldsymbol{D}=P,S=[s_1,s_2])\\
=&\operatorname*{argmax}_{A\in\{A_1,A_2\}}\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2])\Pr(\mathcal{A}=A|S=[s_1,s_2])\\
=&\operatorname*{argmax}_{A\in\{A_1,A_2\}}\Big(\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2],\boldsymbol{C}=T)\\
&\cdot\Pr(\boldsymbol{C}=T|\mathcal{A}=A,S=[s_1,s_2])\\
&+\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2]),\boldsymbol{C}=F)\\
&\cdot\Pr(\boldsymbol{C}=F|\mathcal{A}=A,S=[s_1,s_2])\Big)\\
&\cdot\Pr(\mathcal{A}=A|S=[s_1,s_2])\\
=&\operatorname*{argmax}_{A\in\{A_1,A_2\}}\Big(\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2],\boldsymbol{C}=T)\\
&\cdot h(S=[s_1,s_2],\mathcal{A}=A)\\
&+\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2],\boldsymbol{C}=F)\\
&\cdot(1-h(S=[s_1,s_2],\mathcal{A}=A))\Big)\cdot\Pr(\mathcal{A}=A|S=[s_1,s_2])
\end{aligned}
$$

Under our assumptions of $\beta_2(\beta_1^{-1}(s_2)) > \beta_1(\beta_2^{-1}(s_2))$ and $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$, paper 1 has higher estimated quality under A_1 and paper 2 has higher estimated quality under A_2. Suppose paper 1 is accepted, i.e., $\boldsymbol{D}=P_1$. The value of the above expression under $\mathcal{A}=A_1$ is

$$
\begin{aligned}
&(\Pr(\boldsymbol{D}=P_1|\mathcal{A}=A_1,S=[s_1,s_2],\boldsymbol{C}=T)\\
&\cdot h(S=[s_1,s_2],\mathcal{A}=A_1)\\
&+\Pr(\boldsymbol{D}=P_1|\mathcal{A}=A_1,S=[s_1,s_2],\boldsymbol{C}=F)\\
&\cdot(1-h(S=[s_1,s_2],\mathcal{A}=A_1)))\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])\\
=&(q_1+0)\cdot f_1(s_1)f_2(s_2)\\
=&f_1(s_1)f_2(s_2)q_1.
\end{aligned}
$$

On the other hand, suppose paper 1 is accepted, the value of the above expression under $\mathcal{A}=A_2$ is

$$
\begin{aligned}
&(\Pr(\boldsymbol{D}=P_1|\mathcal{A}=A_2,S=[s_1,s_2],\boldsymbol{C}=T)\\
&\cdot h(S=[s_1,s_2],\mathcal{A}=A_2)\\
&+\Pr(\boldsymbol{D}=P_1|\mathcal{A}=A_2,S=[s_1,s_2],\boldsymbol{C}=F)\\
&\cdot(1-h(S=[s_1,s_2],\mathcal{A}=A_2)))\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])\\
=&(0+(1-q_2))\cdot f_1(s_1)f_2(s_2)\\
=&f_2(s_1)f_1(s_2)(1-q_2).
\end{aligned}
$$

Therefore, when the conference accepts paper 1, if $f_1(s_1)f_2(s_2)q_1 > f_2(s_1)f_1(s_2)(1-q_2)$, then the adversary guesses $\mathcal{A}=A_1$. Otherwise, it guesses $\mathcal{A}=A_2$ except that when $f_1(s_1)f_2(s_2)q_1 = f_2(s_1)f_1(s_2)(1-q_2)$, it makes a random guess assigning probability $\frac{1}{2}$ to each assignment. Similarly, if paper 2 is accepted, the adversary compares $f_1(s_1)f_2(s_2)(1-q_1)$ and $f_2(s_1)f_1(s_2)q_2$ where it guesses $\mathcal{A}=A_1$ when $f_1(s_1)f_2(s_2)(1-q_1) > f_2(s_1)f_1(s_2)q_2$. There are 2 papers and 2 possible assignments, so we have 4 scenarios combining decisions and assignments.

1. Scenario 1: $\mathcal{A} = A_1$ and $\boldsymbol{D} = P_1$
   This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_1(s_1)f_2(s_2)q_1}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. In this scenario, the adversary guesses wrong if $f_1(s_1)f_2(s_2)q_1 < f_2(s_1)f_1(s_2)(1-q_2)$.
2. Scenario 2: $\mathcal{A} = A_1$ and $\boldsymbol{D} = P_2$
   This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_2 | S = [s_1, s_2]) = \frac{f_1(s_1)f_2(s_2)(1-q_1)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. In this scenario, the adversary guesses wrong if $f_1(s_1)f_2(s_2)(1-q_1) < f_2(s_1)f_1(s_2)q_2$.
3. Scenario 3: $\mathcal{A} = A_2$ and $\boldsymbol{D} = P_1$
   This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_2(s_1)f_1(s_2)(1-q_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. In this scenario, the adversary guesses wrong if $f_1(s_1)f_2(s_2)q_1 > f_2(s_1)f_1(s_2)(1-q_2)$.
4. Scenario 4: $\mathcal{A} = A_2$ and $\boldsymbol{D} = P_2$
   This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_2 | S = [s_1, s_2]) = \frac{f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. In this scenario, the adversary guesses wrong if $f_1(s_1)f_2(s_2)(1-q_1) > f_2(s_1)f_1(s_2)q_2$.

To compute the error of the adversary, we need to compare $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$. So as in our assumptions, $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$. From the above 4 scenarios, 2 of them compare $f_1(s_1)f_2(s_2)q_1$ with $f_2(s_1)f_1(s_2)(1-q_2)$ and 2 of them compare $f_1(s_1)f_2(s_2)q_1$ with $f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$. To analyze the error of the adversary, we consider 5 cases of the value of $f_1(s_1)f_2(s_2)q_1$ separated by $f_2(s_1)f_1(s_2)(1-q_2)$ and $f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$. For each case, we refer to the 4 scenarios of $(\mathcal{A}, \boldsymbol{D})$ above. Also note that $\mathcal{E}_C([s_1, s_2]) = 1 - \frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ as computed above.

- If $f_1(s_1)f_2(s_2)q_1 < f_2(s_1)f_1(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary guesses wrong in scenarios 1 and 4.
  Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. Since $\mathcal{E}_C([s_1, s_2]) = 1 - \frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, the relation between error of the adversary and error of the conference is $\mathcal{E}_A([s_1, s_2]) = 1 - \mathcal{E}_C([s_1, s_2])$. For $0 \le f_1(s_1)f_2(s_2)q_1 < f_2(s_1)f_1(s_2) - f_2(s_1)f_1(s_2)q_2$, $\mathcal{E}_C([s_1, s_2]) \in (\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, 1]$.
- If $f_1(s_1)f_2(s_2)q_1 = f_2(s_1)f_1(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary makes random guess in scenarios 1 and 3 and guesses wrong in scenario 4.
  Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ and error of the conference $\mathcal{E}_C([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $f_2(s_1)f_1(s_2) - f_2(s_1)f_1(s_2)q_2 < f_1(s_1)f_2(s_2)q_1 < f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary guesses wrong in scenarios 3 and 4.
  Error of the the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, which is constant. In this case, since error of the conference $\mathcal{E}_C([s_1, s_2]) = 1 - \frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we can find that $\mathcal{E}_C([s_1, s_2])$ ranges from $(\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ to $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)})$.
- If $f_1(s_1)f_2(s_2)q_1 = f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary makes random guess in scenarios 2 and 4 and guesses wrong in scenario 3.
  Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ and error of the conference $\mathcal{E}_C([s_1, s_2])$ is also $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $f_1(s_1)f_2(s_2)q_1 > f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary guesses wrong in scenarios 2 and 3.
  Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $1 - \frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, which is the same as the error of the conference $\mathcal{E}_C([s_1, s_2])$. The relation between error of the adversary and error of the conference is $\mathcal{E}_A([s_1, s_2]) = \mathcal{E}_C([s_1, s_2])$. For $1 \ge f_1(s_1)f_2(s_2)q_1 > f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$, $\mathcal{E}_C([s_1, s_2]) \in [0, \frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)})$.

Therefore, the relation between error of the adversary and error of the conference when $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$ is of the shape of a trapezoid in $[0, 1]$ with the three line segments of the slope +1, 0, and -1 as in Figure 4a. Note that the relation between the per-instance errors does not change with the relation between values of $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$. So Figure 4a is the relation between the errors when $u > v$. Similarly, Figure 4b is the relation between the errors when $u \le v$.

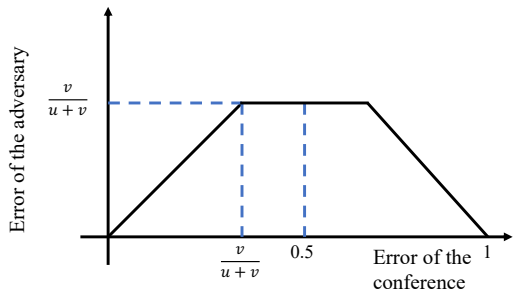
From Figure 4 we see that the conference should keep its per-instance error less than $\frac{\min\{u,v\}}{u+v}$ to stay optimal. Because if the error of the conference is greater than $\frac{\min\{u,v\}}{u+v}$, increasing its error does not increase the error of the adversary and thus is not optimal. Thus, the Pareto frontier of per-instance error of the adversary against error of the conference is the first line segment with slope 1 in both Figure 4a and Figure 4b when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$.

C.3 Proof of Theorem 4.3

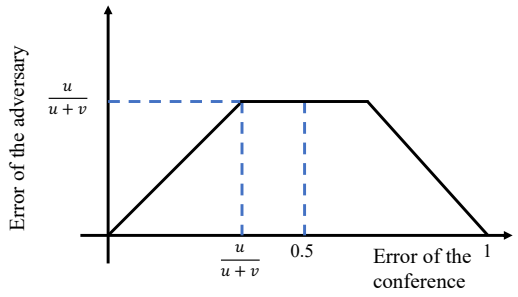
We prove that Algorithm 1 is optimal for each instance of scores $S = [s_1, s_2]$ with desired error of the conference $\mathcal{E}_C([s_1, s_2])$ in the noiseless setting.

From Theorem 4.2 we know that if a paper has higher estimated quality under both assignments, the conference should accept the paper. This is the optimal calibration strategy for the conference.

Otherwise when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, we use the Pareto frontier in Theorem 4.2 to explain the optimality of our algorithm. Suppose $f_1(s_1)f_2(s_2) \le f_2(s_1)f_1(s_2)$, then the endpoint on the Pareto frontier has both error of the conference and error of the adversary being $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. If $\mathcal{E}_C([s_1, s_2]) < \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we maximize the



(a) Maximum per-instance error of the adversary given per-instance error of the conference when $u > v$.



(b) Maximum per-instance error of the adversary given per-instance error of the conference when $u \leq v$

Figure 4: Relation between error of the adversary and error of the conference with $u = f_1(s_1)f_2(s_2)$ and $v = f_2(s_1)f_1(s_2)$.

error of the adversary by operating on the Pareto frontier. If $\mathcal{E}_C([s_1, s_2]) \geq \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we operate at the endpoint where error of the adversary is maximum and error of the conference is no larger than the desired $\mathcal{E}_C([s_1, s_2])$. The endpoint is the point with minimum error of the conference such that error of the adversary is maximum. Therefore, it is optimal for the conference.

Similarly, if $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$, the algorithm is also optimal by maximizing error of the adversary under desired error of the conference following the Pareto frontier. Algorithm 1 follows the procedure by choosing the corresponding q_1 and q_2 for each point on the Pareto frontier and thus is optimal for the conference.

C.4 Proof of Theorem 4.4

Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$ operates on the endpoint of the Pareto frontier when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$. We use ζ to denote the error of running Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$ for all $[s_1, s_2]$. Then we have Algorithm 2 that has a maximum allowable average-case error of the conference $\mathcal{E}_C$ as input.

If $\mathcal{E}_C \geq \zeta$, we operate at $\mathcal{E}_C = \zeta$ by running Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$. From Theorem 4.3 we know that Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$ is Pareto optimal for all score pairs such that the error of the adversary is maximized and no smaller error of the conference can achieve the same privacy guarantee. Increasing the error of the conference will not increase the error of the adversary. Thus, it is Pareto optimal for the allowable average-case error of the conference.

If $\mathcal{E}_C < \zeta$, the coin toss ensures that the average-case error of the conference is $\zeta \cdot \frac{\mathcal{E}_C}{\zeta} + 0 \cdot (1 - \frac{\mathcal{E}_C}{\zeta}) = \mathcal{E}_C$. If we use η to denote the average-case error of the adversary when the conference always calibrates under the true assignment, then the average-case error of the adversary when the conference runs Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$ is $\zeta + \eta$. Because when the conference always calibrates under the true assignment, the adversary only makes error when $s_1 \leq \min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$ or $s_1 \geq \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$. And if the conference adopts Algorithm 1 with $\mathcal{E}_C([s_1, s_2]) = 1$, the adversary has error η when $s_1 \leq \min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$ or $s_1 \geq \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$ and has error ζ when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$. Therefore, the average-case error of the adversary is $(\eta+\zeta)\cdot\frac{\mathcal{E}_C}{\zeta}+\eta\cdot(1-\frac{\mathcal{E}_C}{\zeta}) = \mathcal{E}_C+\eta$. From Theorem 4.3 we know that when the conference has per-instance error $\mathcal{E}_C$, the maximum per-instance error of the adversary is $\mathcal{E}_C$ if $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$. In addition, a Pareto optimal strategy when $s_1 \leq \min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$ or $s_1 \geq \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$ has error of the adversary being η and error of the conference being 0. Therefore, for the average-case error of the conference being $\mathcal{E}_C$, the average-case error of the adversary is no larger than $\mathcal{E}_C + \eta$. Therefore Algorithm 2 is Pareto optimal.

C.5 Proof of Theorem 4.5

To find the Pareto frontier of per-instance error of the adversary against per-instance error of the conference in the noisy setting, we first find the maximum per-instance error of the adversary given per-instance error of the conference in this range.

Prior to computing the errors, we compute the posterior distribution of the quality of the papers given the assignment and scores. We have $\theta_1^*|\{S = [s_1, s_2], \mathcal{A} = A_1\} \sim N\left(\frac{a_1(s_1-b_1)}{a_1^2+\sigma^2}, \frac{\sigma^2}{a_1^2+\sigma^2}\right)$ and $\theta_1^*|\{S = [s_1, s_2], \mathcal{A} = A_2\} \sim N\left(\frac{a_2(s_1-b_2)}{a_2^2+\sigma^2}, \frac{\sigma^2}{a_2^2+\sigma^2}\right)$. Similarly, $\theta_2^*|\{S = [s_1, s_2], \mathcal{A} = A_1\} \sim N\left(\frac{a_2(s_2-b_2)}{a_2^2+\sigma^2}, \frac{\sigma^2}{a_2^2+\sigma^2}\right)$ and $\theta_2^*|\{S = [s_1, s_2], \mathcal{A} = A_2\} \sim N\left(\frac{a_1(s_2-b_1)}{a_1^2+\sigma^2}, \frac{\sigma^2}{a_1^2+\sigma^2}\right)$. We show calculation for one of the posterior distribution. Note that in continuous space, the probability is taken as the density of the corresponding distribution.

$$\begin{aligned}&\Pr(\theta_1^* = t|S = [s_1, s_2], \mathcal{A} = A_1)\\ =&\frac{\Pr(S = [s_1, s_2]|\theta_1^* = t, \mathcal{A} = A_1) \cdot \Pr(\theta_1^* = t|\mathcal{A} = A_1)}{\Pr(S = [s_1, s_2]|\mathcal{A} = A_1)}\end{aligned}$$

Then we separately compute each term in the equation above. Note that $s_1|\{\theta_1^* = t, \mathcal{A} = A_1\} \sim N(a_1 t + b_1, \sigma^2)$ and $s_1|\mathcal{A} = A_1 \sim N(b_1, a_1^2 + \sigma^2)$. Since s_2 is independent of θ_1^* given that $\mathcal{A} = A_1$, $s_2|\{\theta_1^* = t, \mathcal{A} = A_1\}$ and $s_2|\mathcal{A} = A_1$ have the same distribution. In addition, θ^* and A are independent.

$$
\begin{aligned}
&\Pr(S = [s_1, s_2]|\theta_1^* = t, \mathcal{A} = A_1) \\
=&\Pr(S[1] = s_1|\theta_1^* = t, \mathcal{A} = A_1) \cdot \Pr(S[2] = s_2|\mathcal{A} = A_1) \\
=&\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{s_1 - (a_1 t + b_1)}{\sigma}\right)^2} \cdot \Pr(S[2] = s_2|\mathcal{A} = A_1)
\end{aligned}
$$

$$
\begin{aligned}
&\Pr(\theta_2^* = t|\mathcal{A} = A_1) \\
=&\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}
\end{aligned}
$$

$$
\begin{aligned}
&\Pr(S = [s_1, s_2]|\mathcal{A} = A_1) \\
=&\Pr(S[1] = s_1|\mathcal{A} = A_1) \cdot \Pr(S[2] = s_2|\mathcal{A} = A_1) \\
=&\frac{1}{\sqrt{2\pi}\sqrt{a_1^2 + \sigma^2}} e^{-\frac{1}{2}\left(\frac{s_1 - b_1}{\sqrt{a_1^2 + \sigma^2}}\right)^2} \cdot \Pr(S[2] = s_2|\mathcal{A} = A_1)
\end{aligned}
$$

Therefore, combining the terms we get

$$
\begin{aligned}
&\Pr(\theta_1^* = t|S = [s_1, s_2], \mathcal{A} = A_1) \\
=&\frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{s_1 - (a_1 t + b_1)}{\sigma}\right)^2} \cdot \Pr(S[2] = s_2|\mathcal{A} = A_1) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}}{\frac{1}{\sqrt{2\pi}\sqrt{a_1^2 + \sigma^2}} e^{-\frac{1}{2}\left(\frac{s_1 - b_1}{\sqrt{a_1^2 + \sigma^2}}\right)^2} \cdot \Pr(S[2] = s_2|\mathcal{A} = A_1)} \\
=&\frac{1}{\sqrt{2\pi}} \sqrt{\frac{a_1^2 + \sigma^2}{\sigma^2}} e^{-\frac{1}{2}\left(\left(\frac{s_1 - (a_1 t + b_1)}{\sigma}\right)^2 + t^2 - \left(\frac{s_1 - b_1}{\sqrt{a_1^2 + \sigma^2}}\right)^2\right)} \\
=&\frac{1}{\sqrt{2\pi}} \sqrt{\frac{a_1^2 + \sigma^2}{\sigma^2}} e^{-\frac{1}{2}\left(t - \frac{a_1(s_1 - b_1)}{a_1^2 + \sigma^2}\right)^2 \cdot \frac{a_1^2 + \sigma^2}{\sigma^2}}.
\end{aligned}
$$

The other three posteriors are computed in a similar fashion.

Given the posterior distribution of the qualities, we can compute the posterior probability that one paper has higher quality than the other.

$$
\begin{aligned}
&\Pr(\theta_1^* > \theta_2^*|\mathcal{A} = A_1, S = [s_1, s_2]) \\
=&\Pr\left(N\left(\frac{a_1(s_1 - b_1)}{a_1^2 + \sigma^2}, \frac{\sigma^2}{a_1^2 + \sigma^2}\right)\right. \\
&\left. > N\left(\frac{a_2(s_2 - b_2)}{a_2^2 + \sigma^2}, \frac{\sigma^2}{a_2^2 + \sigma^2}\right)\right) \\
=&\Pr\left(N\left(\frac{a_1(s_1 - b_1)}{a_1^2 + \sigma^2} - \frac{a_2(s_2 - b_2)}{a_2^2 + \sigma^2}, \frac{\sigma^2}{a_1^2 + \sigma^2} + \frac{\sigma^2}{a_2^2 + \sigma^2}\right) > 0\right) \\
=&\Pr\left(\frac{a_1(s_1 - b_1)}{a_1^2 + \sigma^2} - \frac{a_2(s_2 - b_2)}{a_2^2 + \sigma^2}\right. \\
&\left. + \sqrt{\frac{\sigma^2}{a_1^2 + \sigma^2} + \frac{\sigma^2}{a_2^2 + \sigma^2}} N(0,1) > 0\right) \\
=&\Pr\left(N(0,1) > \frac{a_2(a_1^2 + \sigma^2)(s_2 - b_2) - a_1(a_2^2 + \sigma^2)(s_1 - b_1)}{\sqrt{\sigma^2(a_1^2 + a_2^2 + 2\sigma^2)(a_1^2 + \sigma^2)(a_2^2 + \sigma^2)}}\right) \\
=&1 - \Phi\left(\frac{a_2(a_1^2 + \sigma^2)(s_2 - b_2) - a_1(a_2^2 + \sigma^2)(s_1 - b_1)}{\sqrt{\sigma^2(a_1^2 + a_2^2 + 2\sigma^2)(a_1^2 + \sigma^2)(a_2^2 + \sigma^2)}}\right)
\end{aligned}
$$

We use Φ to denote the cumulative distribution function of standard Gaussian distribution. Similarly, we can compute that

$$
\begin{aligned}
&\Pr(\theta_1^* \leq \theta_2^*|\mathcal{A} = A_1, S = [s_1, s_2]) \\
=&\Phi\left(\frac{a_2(a_1^2 + \sigma^2)(s_2 - b_2) - a_1(a_2^2 + \sigma^2)(s_1 - b_1)}{\sqrt{\sigma^2(a_1^2 + a_2^2 + 2\sigma^2)(a_1^2 + \sigma^2)(a_2^2 + \sigma^2)}}\right) \\
&\Pr(\theta_1^* > \theta_2^*|\mathcal{A} = A_2, S = [s_1, s_2]) \\
=&1 - \Phi\left(\frac{a_1(a_2^2 + \sigma^2)(s_2 - b_1) - a_2(a_1^2 + \sigma^2)(s_1 - b_2)}{\sqrt{\sigma^2(a_1^2 + a_2^2 + 2\sigma^2)(a_1^2 + \sigma^2)(a_2^2 + \sigma^2)}}\right) \\
&\Pr(\theta_1^* \leq \theta_2^*|\mathcal{A} = A_2, S = [s_1, s_2]) \\
=&\Phi\left(\frac{a_1(a_2^2 + \sigma^2)(s_2 - b_1) - a_2(a_1^2 + \sigma^2)(s_1 - b_2)}{\sqrt{\sigma^2(a_1^2 + a_2^2 + 2\sigma^2)(a_1^2 + \sigma^2)(a_2^2 + \sigma^2)}}\right)
\end{aligned}
$$

For simplicity, let $\Phi_1 = \Phi\left(\frac{a_2(a_1^2 + \sigma^2)(s_2 - b_2) - a_1(a_2^2 + \sigma^2)(s_1 - b_1)}{\sqrt{\sigma^2(a_1^2 + a_2^2 + 2\sigma^2)(a_1^2 + \sigma^2)(a_2^2 + \sigma^2)}}\right)$ and $\Phi_2 = \Phi\left(\frac{a_1(a_2^2 + \sigma^2)(s_2 - b_1) - a_2(a_1^2 + \sigma^2)(s_1 - b_2)}{\sqrt{\sigma^2(a_1^2 + a_2^2 + 2\sigma^2)(a_1^2 + \sigma^2)(a_2^2 + \sigma^2)}}\right)$. Since the conference does calibration using the posterior probabilities, the values of Φ_1 and Φ_2 determines the conference decision. By Proposition 4.1, we know that the conference should accept the paper with higher estimated quality under both assignments without any calibration. Therefore, if Φ_1 and Φ_2 are both less than $\frac{1}{2}$, the conference should accept paper 1. Similarly, if Φ_1 and Φ_2 are both greater than $\frac{1}{2}$, the conference should accept paper 2. Otherwise, when $\Phi_1 - \frac{1}{2}$ and $\Phi_2 - \frac{1}{2}$ have different signs, the conference should do calibration with function h. As before, since S is a fixed realization in the analysis, we simplify the calibration

strategy for the conference as

$$q_1 = h(S, A_1)$$
$$q_2 = h(S, A_2).$$

We first consider part (1) of the theorem. If $s_1 > \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, which is when $\Phi_1 \le \frac{1}{2}$ and $\Phi_2 \le \frac{1}{2}$, the conference accepts paper 1 and the adversary guesses the assignment based on the scores only. Then the error of the conference is the probability that paper 2 has higher quality.

$$\begin{aligned}
&\Pr(\theta_1^* < \theta_2^* | S = [s_1, s_2]) \\
=&\Pr(\theta_1^* < \theta_2^* | \mathcal{A} = A_1, S = [s_1, s_2]) \\
&\cdot \Pr(\mathcal{A} = A_1 | S = [s_1, s_2]) \\
&+ \Pr(\theta_1^* < \theta_2^* | \mathcal{A} = A_2, S = [s_1, s_2]) \\
&\cdot \Pr(\mathcal{A} = A_2 | S = [s_1, s_2]) \\
=&\Phi_1 \cdot \frac{f_1(s_1) f_2(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)} \\
&+ \Phi_2 \cdot \frac{f_2(s_1) f_1(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}
\end{aligned}$$

Similarly, if $s_1 < \min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, which is when $\Phi_1 \ge \frac{1}{2}$ and $\Phi_2 \ge \frac{1}{2}$, the conference accepts paper 2 and the error of the conference is the probability that paper 1 has higher quality.

$$\begin{aligned}
&\Pr(\theta_1^* > \theta_2^* | S = [s_1, s_2]) \\
=&\Pr(\theta_1^* > \theta_2^* | \mathcal{A} = A_1, S = [s_1, s_2]) \\
&\cdot \Pr(\mathcal{A} = A_1 | S = [s_1, s_2]) \\
&+ \Pr(\theta_1^* > \theta_2^* | \mathcal{A} = A_2, S = [s_1, s_2]) \\
&\cdot \Pr(\mathcal{A} = A_2 | S = [s_1, s_2]) \\
=&(1 - \Phi_1) \cdot \frac{f_1(s_1) f_2(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)} \\
&+ (1 - \Phi_2) \cdot \frac{f_2(s_1) f_1(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}
\end{aligned}$$

In both cases, error of the adversary is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, which is the error when the adversary guesses the assignment based on scores only.

We now consider the rest scores in part (1) of the theorem. If $s_1 = \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, without loss of generality, we assume $\max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} = \beta_2(\beta_1^{-1}(s_2))$, then the conference accepts each paper uniform at random if calibrating under A_1 and accepts paper 1 if calibrating under A_2. Since paper 1 has higher or equal quality than paper 2, the conference only has error when paper 2 is accepted and $\mathcal{A} = A_2$.

$$\begin{aligned}
&\Pr(\text{conference accepts lower-quality paper} | S = [s_1, s_2]) \\
=&\Pr(\text{conference accepts lower-quality paper} | S = [s_1, s_2], D = P_1) \\
&\cdot \Pr(D = P_1 | S = [s_1, s_2]) \\
&+ \Pr(\text{conference accepts lower-quality paper} | S = [s_1, s_2], D = P_2) \\
&\cdot \Pr(D = P_2 | S = [s_1, s_2]) \\
=&\Pr(\theta_1^* < \theta_2^* | S = [s_1, s_2]) \Pr(D = P_1 | S = [s_1, s_2]) \\
&+ \Pr(\theta_1^* > \theta_2^* | S = [s_1, s_2]) \Pr(D = P_2 | S = [s_1, s_2]).
\end{aligned}$$

Note that in this case, $\Pr(\theta_1^* < \theta_2^* | S = [s_1, s_2]) < \Pr(\theta_1^* > \theta_2^* | S = [s_1, s_2])$. By similar calculation as in Appendix C.3, we have

$$\begin{aligned}
&\Pr(D = P_1 | S = [s_1, s_2]) \\
=&\frac{1}{2}(1 - q_1) \Pr(\mathcal{A} = A_1 | S = [s_1, s_2]) + \frac{1}{2} q_2 \Pr(\mathcal{A} = A_2 | S = [s_1, s_2]) \\
&\Pr(D = P_2 | S = [s_1, s_2]) \\
=&\frac{1}{2} q_1 \Pr(\mathcal{A} = A_1 | S = [s_1, s_2]) + \frac{1}{2}(1 - q_2) \Pr(\mathcal{A} = A_2 | S = [s_1, s_2]).
\end{aligned}$$

Error of the conference is then a convex combination of $\Pr(\theta_1^* < \theta_2^* | S = [s_1, s_2])$ and $\Pr(\theta_1^* > \theta_2^* | S = [s_1, s_2])$ and is minimized when the weight of $\Pr(\theta_1^* > \theta_2^* | S = [s_1, s_2])$ is 0.

For the adversary, if paper 1 is accepted, it gains no information on the assignment other than the scores so its error is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. Otherwise, it guesses $\mathcal{A} = A_1$ and its error is $\Pr(\mathcal{A} = A_2 | S = [s_1, s_2])$. Note that error of the adversary does not exceed $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ since in the worst case for the adversary, it guesses the assignment solely based on the scores and ignore the conference decision.

$$\begin{aligned}
&\Pr(\text{adversary guesses assignment wrong} | S = [s_1, s_2]) \\
=&\Pr(\text{adversary guesses assignment wrong} | S = [s_1, s_2], D = P_1) \\
&\cdot \Pr(D = P_1 | S = [s_1, s_2]) \\
&+ \Pr(\text{adversary guesses assignment wrong} | S = [s_1, s_2], D = P_2) \\
&\cdot \Pr(D = P_2 | S = [s_1, s_2]) \\
=&\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2) + f_2(s_1)f_1(s_2)} \\
&\cdot \Big((1 - \frac{1}{2} q_1) \Pr(\mathcal{A} = A_1 | S = [s_1, s_2]) \\
&+ (\frac{1}{2} + \frac{1}{2} q_2) \Pr(\mathcal{A} = A_2 | S = [s_1, s_2])\Big) \\
&+ \Pr(\mathcal{A} = A_2 | S = [s_1, s_2]) \\
&\cdot \Big(\frac{1}{2} q_1 \Pr(\mathcal{A} = A_1 | S = [s_1, s_2]) \\
&+ \frac{1}{2}(1 - q_2) \Pr(\mathcal{A} = A_2 | S = [s_1, s_2])\Big)
\end{aligned}$$

Therefore, we can minimize the error of the conference to 0 by choosing $q_1 = 0$ and $q_2 = 1$, which results in the conference always accepts paper 1. Then error of

the adversary is $\frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, which is maximized. Further increase of error of the conference cannot increase error of the adversary. So the Pareto optimal point is $(\Phi_1 \cdot \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \Phi_2 \cdot \frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)})$. The same argument follows when $s_1 = \min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$.

We then look at part (2) of the theorem where the scores lie in the region $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} < s_1 < \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$. We will then show the proof with the assumptions that $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$ and $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ with $0 < \varphi_2 < \varphi_1$. The analysis is of the same procedure for different assumptions on the values of $f_1(s_1)f_2(s_2)$, $f_2(s_1)f_1(s_2)$, Φ_1 and Φ_2 with $\Phi_1 - \frac{1}{2}$ and $\Phi_2 - \frac{1}{2}$ having different signs. The notations are of the same meaning as in Section C.3. In the noisy setting, even if the conference calibrates under the true assignment, there is still possibility to accept the lower-quality paper due to the noise in the scores given by the reviewers. Note that with the assumptions and when $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} < s_1 < \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, the conference accepts paper 1 if calibrates under A_1 and accepts paper 2 if calibrates under A_2 by the assumptions on Φ_1 and Φ_2. So we have

$$
\begin{aligned}
&\Pr(\text{conference accepts lower-quality paper}|S=[s_1,s_2])\\
=&\Pr(\text{conference accepts } P_1, \theta_1^*<\theta_2^*|S=[s_1,s_2])\\
&+\Pr(\text{conference accepts } P_2, \theta_1^*>\theta_2^*|S=[s_1,s_2])\\
=&\Pr(\text{conference accepts } P_1|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&\cdot\Pr(\theta_1^*<\theta_2^*|S=[s_1,s_2])\\
&+\Pr(\text{conference accepts } P_2|\theta_1^*>\theta_2^*, S=[s_1,s_2])\\
&\cdot\Pr(\theta_1^*>\theta_2^*|S=[s_1,s_2]).
\end{aligned}
$$

We then expand each of the two terms.

$$
\begin{aligned}
&\Pr(\text{conference accepts } P_1|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
=&\Pr(\text{conference accepts } P_1, \mathcal{A}=A_1|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&+\Pr(\text{conference accepts } P_1, \mathcal{A}=A_2|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
=&\Pr(\text{conference accepts } P_1|\mathcal{A}=A_1, \theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&\cdot P(\mathcal{A}=A_1|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&+\Pr(\text{conference accepts } P_1|\mathcal{A}=A_2, \theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&\cdot\Pr(\mathcal{A}=A_2|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
=&\Pr(\boldsymbol{C}=T|\mathcal{A}=A_1, \theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&\cdot\Pr(\mathcal{A}=A_1|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&+\Pr(\boldsymbol{C}=F|\mathcal{A}=A_2, \theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&\cdot\Pr(\mathcal{A}=A_2|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
=&q_1\Pr(\mathcal{A}=A_1|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
&+(1-q_2)\Pr(\mathcal{A}=A_2|\theta_1^*<\theta_2^*, S=[s_1,s_2])\\
=&q_1\frac{\Pr(\mathcal{A}=A_1, \theta_1^*<\theta_2^*|S=[s_1,s_2])}{\Pr(\theta_1^*<\theta_2^*|S=[s_1,s_2])}\\
&+(1-q_2)\frac{\Pr(\mathcal{A}=A_2, \theta_1^*<\theta_2^*|S=[s_1,s_2])}{\Pr(\theta_1^*<\theta_2^*|S=[s_1,s_2])}\\
=&q_1\\
&\cdot\frac{\Pr(\theta_1^*<\theta_2^*|\mathcal{A}=A_1, S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])}{\Pr(\theta_1^*<\theta_2^*|S=[s_1,s_2])}\\
&+(1-q_2)\\
&\cdot\frac{\Pr(\theta_1^*<\theta_2^*|\mathcal{A}=A_2, S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])}{\Pr(\theta_1^*<\theta_2^*|S=[s_1,s_2])}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\Pr(\text{conference accepts } P_2|\theta_1^*>\theta_2^*, S=[s_1,s_2])\\
=&(1-q_1)\\
&\cdot\frac{\Pr(\theta_1^*>\theta_2^*|\mathcal{A}=A_1, S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])}{\Pr(\theta_1^*>\theta_2^*|S=[s_1,s_2])}\\
&+q_2\\
&\cdot\frac{\Pr(\theta_1^*>\theta_2^*|\mathcal{A}=A_2, S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])}{\Pr(\theta_1^*>\theta_2^*|S=[s_1,s_2])}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\Pr(\text{conference accepts lower-quality paper}|S=[s_1,s_2])\\
=&q_1\Pr(\theta_1^*<\theta_2^*|\mathcal{A}=A_1, S=[s_1,s_2])\\
&\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])\\
&+(1-q_2)\Pr(\theta_1^*<\theta_2^*|\mathcal{A}=A_2, S=[s_1,s_2])\\
&\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])\\
&+(1-q_1)\Pr(\theta_1^*>\theta_2^*|\mathcal{A}=A_1, S=[s_1,s_2])\\
&\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])\\
&+q_2\Pr(\theta_1^*>\theta_2^*|\mathcal{A}=A_2, S=[s_1,s_2])\\
&\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])\\
=&\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}(q_1\Phi_1+(1-q_1)(1-\Phi_1))\\
&+\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\cdot((1-q_2)\Phi_2+q_2(1-\Phi_2)).
\end{aligned}
$$

Under the assumptions that $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ where $0 < \varphi_2 < \varphi_1$ and $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$, we analyze the per-instance error of the adversary similar to the procedure in Section C.2. There are 4 scenarios combining the decision and the true assignment.

1. Scenario 1: $\mathcal{A} = A_1$ and $\boldsymbol{D} = P_1$
   This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_1(s_1)f_2(s_2)q_1}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. In this scenario, the adversary guesses wrong if $q_1 f_1(s_1)f_2(s_2) < (1-q_2)f_2(s_1)f_1(s_2)$.
2. Scenario 2: $\mathcal{A} = A_1$ and $\boldsymbol{D} = P_2$
   This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_1(s_1)f_2(s_2)(1-q_1)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. In this scenario, the adversary guesses wrong if $(1-q_1)f_1(s_1)f_2(s_2) < q_2 f_2(s_1)f_1(s_2)$.
3. Scenario 3: $\mathcal{A} = A_2$ and $\boldsymbol{D} = P_1$
   This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_2(s_1)f_1(s_2)(1-q_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. In this scenario, the adversary guesses wrong if $q_1 f_1(s_1)f_2(s_2) > (1-q_2)f_2(s_1)f_1(s_2)$.
4. Scenario 4: $\mathcal{A} = A_2$ and $\boldsymbol{D} = P_2$
   This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. In this scenario, the adversary guesses wrong if $(1-q_1)f_1(s_1)f_2(s_2) > q_2 f_2(s_1)f_1(s_2)$.

To compute the error of the adversary, we need to compare $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$. So we suppose $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$. From the above 4 scenarios, 2 of them compare $f_1(s_1)f_2(s_2)q_1$ with $f_2(s_1)f_1(s_2)(1-q_2)$ and 2 of them compare $f_1(s_1)f_2(s_2)q_1$ with $f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$. To analyze the error of the adversary, we consider 5 cases of the value of $f_1(s_1)f_2(s_2)q_1$ separated by $f_2(s_1)f_1(s_2)(1-q_2)$ and $f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$. We refer to the 4 scenarios of $(\mathcal{A}, \boldsymbol{D})$ above.

- If $q_1 f_1(s_1)f_2(s_2) < f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary guesses wrong in scenarios 1 and 4. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $q_1 f_1(s_1)f_2(s_2) = f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary makes random guess in scenarios 2 and 4 and guesses wrong in scenario 1. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{q_1 f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{1}{2}(\frac{(1-q_1)f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}) = \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2) < q_1 f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary guesses wrong in scenarios 1 and 2. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $q_1 f_1(s_1)f_2(s_2) = f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary makes random guess in scenarios 1 and 3 and guesses wrong in scenario 2. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{(1-q_1)f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{1}{2}(\frac{q_1 f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{(1-q_2)f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}) = \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $q_1 f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary guesses wrong in scenarios 2 and 3. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $1 - \frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

To find the maximum error of the adversary given error of the conference, we solve an optimization problem. In order to formulate the optimization problem, we can combine the 5 cases above into 3 cases for simplicity.

- If $q_1 f_1(s_1)f_2(s_2) \le f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2)$, error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2) \le q_1 f_1(s_1)f_2(s_2) \le f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $q_1 f_1(s_1)f_2(s_2) \ge f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $1 - \frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

We let $T(\mathcal{E}_C) = \mathcal{E}_C(u+v) - u \cdot (1-\Phi_1) - v \cdot \Phi_2$ to be a function that takes the error of the conference as input.

- Maximize $\frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ subject to $\mathcal{E}_C([s_1, s_2])(f_1(s_1)f_2(s_2) + f_2(s_1)f_1(s_2)) - f_1(s_1)f_2(s_2) \cdot (1-\Phi_1) - f_2(s_1)f_1(s_2) \cdot \Phi_2 = f_1(s_1)f_2(s_2)(2\Phi_1 - 1)q_1 + f_2(s_1)f_1(s_2) \cdot (1-2\Phi_2)q_2$ and $q_1 f_1(s_1)f_2(s_2) \le f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2)$.
  The maximum occurs at $q_1 f_1(s_1)f_2(s_2) = f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2)$. Then the intersection of the two lines is $q_1 = 1 - \frac{(2\Phi_1-1)u - T(\mathcal{E}_C([s_1,s_2]))}{(2\Phi_1+2\Phi_2-2)u}$ and $q_2 = \frac{(2\Phi_1-1)u - T(\mathcal{E}_C([s_1,s_2]))}{(2\Phi_1+2\Phi_2-2)v}$.
  - If the intersection point can be reached, $q_1, q_2 \in [0,1]$, $(2\Phi_1 - 1)u \le T(\mathcal{E}_C([s_1, s_2])) \le (1-2\Phi_2)u$, then error of the conference $\mathcal{E}_C([s_1, s_2])$ ranges from $\frac{f_1(s_1)f_2(s_2)\Phi_1}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{f_2(s_1)f_1(s_2)\Phi_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ to $\frac{f_1(s_1)f_2(s_2)(2-\Phi_1-2\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{f_2(s_1)f_1(s_2)\Phi_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
    Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
  - If the intersection point can not be reached and $T(\mathcal{E}_C([s_1, s_2])) < (2\Phi_1 - 1)u$, then no q_1, q_2 are qualified for the constraints.
  - If the intersection point can not be reached and $T(\mathcal{E}_C([s_1, s_2])) > (1-2\Phi_2)u$.
    * If $(1-2\Phi_2)u < T(\mathcal{E}_C([s_1, s_2])) \le 0$ then the maximum is reached when $q_1 = 0$ and $q_2 = \frac{T(\mathcal{E}_C([s_1,s_2]))}{(1-2\Phi_2)v}$. Error of the conference $\mathcal{E}_C([s_1, s_2])$ ranges from $\frac{(2-\Phi_1-2\Phi_2)u+\Phi_2 v}{u+v}$ (when $T(\mathcal{E}_C([s_1, s_2])) = (1-2\Phi_2)u$) to $\frac{(1-\Phi_1)u+\Phi_2 v}{u+v}$ (when $T(\mathcal{E}_C([s_1, s_2])) = 0$).

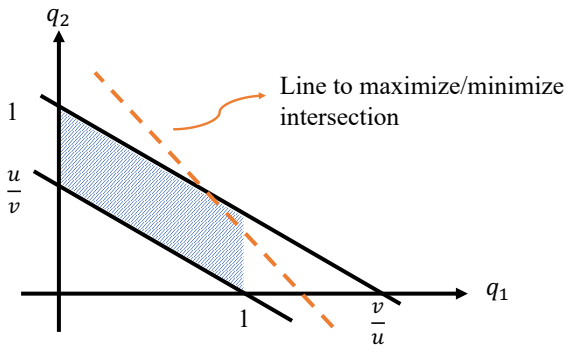


Figure 5: A diagram illustrates the optimization problem in this case.

Error of the adversary $\mathcal{E}_A([s_1,s_2])$ is $\frac{T(\mathcal{E}_C([s_1,s_2]))}{(1-2\Phi_2)(u+v)}$, ranges from $\frac{u}{u+v}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (1-2\Phi_2)u$) to 0 (when $T(\mathcal{E}_C([s_1,s_2])) = 0$).

* If $T(\mathcal{E}_C([s_1,s_2])) > 0$ then no q_1, q_2 are qualified for the constraints.

- Error of the adversary $\mathcal{E}_A([s_1,s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ subject to $f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2) \le q_1 f_1(s_1)f_2(s_2) \le f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$.

  From Figure 5 we can see that error of the conference $\mathcal{E}_C([s_1,s_2])$ has its extremes at $q_1 = 0, q_2 = \frac{u}{v}$ and $q_1 = 1, q_2 = 1 - \frac{u}{v}$. Therefore, error of the conference ranges from $\frac{(2-\Phi_1-2\Phi_2)u+\Phi_2 v}{u+v}$ to $\frac{(\Phi_1+2\Phi_2-1)u+(1-\Phi_2)v}{u+v}$.

- Maximize $1 - \frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ subject to $\mathcal{E}_C([s_1,s_2])(f_1(s_1)f_2(s_2) + f_2(s_1)f_1(s_2)) - f_1(s_1)f_2(s_2)\cdot(1-\Phi_1) - f_2(s_1)f_1(s_2)\cdot\Phi_2 = f_1(s_1)f_2(s_2)(2\Phi_1-1)q_1 + f_2(s_1)f_1(s_2)\cdot(1-2\Phi_2)q_2$ and $q_1 f_1(s_1)f_2(s_2) \ge f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$.

  The maximum occurs at $q_1 f_1(s_1)f_2(s_2) = f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$. Then the intersection of the two lines is $q_1 = \frac{(1-2\Phi_2)v - T(\mathcal{E}_C([s_1,s_2]))}{(2-2\Phi_1-2\Phi_2)u}$ and $q_2 = \frac{T(\mathcal{E}_C([s_1,s_2]))-(2\Phi_1-1)v}{(2-2\Phi_1-2\Phi_2)v}$.

  – If the intersection point can be reached, $q_1, q_2 \in [0,1]$, $(1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u \le T(\mathcal{E}_C([s_1,s_2])) \le (1-2\Phi_2)v$, then error of the conference $\mathcal{E}_C([s_1,s_2])$ ranges from $\frac{f_1(s_1)f_2(s_2)(1-\Phi_1)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (1-2\Phi_2)v$) to $\frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$).

  Error of the adversary $\mathcal{E}_A([s_1,s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

  – If the intersection point can not be reached and $T(\mathcal{E}_C([s_1,s_2])) > (1-2\Phi_2)v$, then no q_1, q_2 are qualified for the constraints.

  – If the intersection point can not be reached and $T(\mathcal{E}_C([s_1,s_2])) < (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$.

  * If $(2\Phi_1-1)u + (1-2\Phi_2)v \le T(\mathcal{E}_C([s_1,s_2])) < (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$ then the maximum is reached when $q_1 = 1$ and $q_2 = \frac{T(\mathcal{E}_C([s_1,s_2]))-(2\Phi_1-1)u}{(1-2\Phi_2)v}$.

    Error of the conference $\mathcal{E}_C([s_1,s_2])$ ranges from $\frac{(\Phi_1+2\Phi_2-1)u+(1-\Phi_2)v}{u+v}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$) to $\frac{\Phi_1 u+(1-\Phi_2)v}{u+v}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (2\Phi_1-1)u + (1-2\Phi_2)v$).

    Error of the adversary $\mathcal{E}_A([s_1,s_2])$ is $1 - \frac{T(\mathcal{E}_C([s_1,s_2]))+(2-2\Phi_1-2\Phi_2)u}{(1-2\Phi_2)(u+v)}$, ranges from $\frac{u}{u+v}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$) to 0 (when $T(\mathcal{E}_C([s_1,s_2])) = (2\Phi_1-1)u + (1-2\Phi_2)v$).

  * If $T(\mathcal{E}_C([s_1,s_2])) < (2\Phi_1-1)u + (1-2\Phi_2)v$ then no q_1, q_2 are qualified for the constraints.

Therefore, the relation between error of the adversary and error of the conference when $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ where $0 < \varphi_2 < \varphi_1$ and $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$ is of the shape of a trapezoid in $[0,1]$ as in Figure 6. Note that the relation between the per-instance errors does not change with the relation between values of $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$ or with the values of Φ_1 and Φ_2.

From Figure 6 we see that the conference should keep its per-instance error between $\frac{u\Phi_1+v(1-\Phi_2)}{u+v}$ and $\frac{u(\Phi_1+2\Phi_2-1)+v(1-\Phi_2)}{u+v}$ to stay optimal. The conference cannot have its error less than $\frac{u\Phi_1+v(1-\Phi_2)}{u+v}$ due to the reviewers' noise. If error of the conference is greater than $\frac{u(\Phi_1+2\Phi_2-1)+v(1-\Phi_2)}{u+v}$, increasing its error does not increase error the adversary and thus is not optimal. Thus, the Pareto frontier of per-instance error of the adversary against error of the conference is the first line segment with positive slope in Figure 6 when $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} < s_1 < \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$.

C.6 Proof of Theorem 4.6

We prove that Algorithm 3 is optimal for each instance of scores $S = [s_1,s_2]$ with desired error of the conference $\mathcal{E}_C([s_1,s_2])$ in the noisy setting. We carry the assumptions from Section C.5 that $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ where $0 < \varphi_2 < \varphi_1$ and $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$.

From Proposition 4.1 we know that if a paper has higher estimated quality under both assignments, the conference should accept the paper. This is the optimal calibration strategy for the conference.

Otherwise when the scores are in the region $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} <$

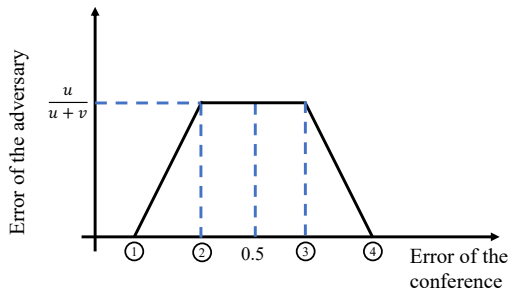


Figure 6: Maximum per-instance error of the adversary given per-instance error of the conference when $u < v$, $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ with $0 < \varphi_2 < \varphi_1$. The coordinates in the plot are: ① = $\frac{u\Phi_1+v(1-\Phi_2)}{u+v}$, ② = $\frac{u(\Phi_1+2\Phi_2-1)+v(1-\Phi_2)}{u+v}$, ③ = $\frac{u(2-\Phi_1-2\Phi_2)+v\Phi_2}{u+v}$, ④ = $\frac{u(1-\Phi_1)+v\Phi_2}{u+v}$.

$s_1 < \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, we use the Pareto frontiers analyze the optimality of our algorithm. Theorem 4.5 shows that the Pareto frontier in the noiseless setting within this region. The analysis is similar to the one in the noiseless setting in Section C.3.

Suppose $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$, then the endpoint on the Pareto frontier has error of the adversary being $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ and error of the conference being $\frac{f_1(s_1)f_2(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2+f_2(s_1)f_1(s_2)}$. If $\frac{f_1(s_1)f_2(s_2)\Phi_1+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} \leq \mathcal{E}_C([s_1,s_2]) < \frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we maximize the error of the adversary by operating on the Pareto frontier. If $\mathcal{E}_C([s_1,s_2]) \geq \frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we operate at the endpoint where error of the adversary is maximum and error of the conference is no larger than the desired $\mathcal{E}_C([s_1,s_2])$. The endpoint is the point with minimum error of the conference such that error of the adversary is maximum. Therefore, it is optimal for the conference.

Algorithm 3 follows the procedure by choosing the corresponding q_1 and q_2 for each point on the Pareto frontier and thus is optimal for the conference.