Least Square Calibration for Peer Reviews

Sijun Tan[1] **Jibang Wu**[1] Xiaohui Bei[2] Haifeng Xu[1]

NeurIPS, Dec 2021

[1]University of Virginia

[2]Nanyang Technological University

Introduction

Introduction

Peer review systems are ubiquitous in a data-driven world.

Introduction

Peer review systems are ubiquitous in a data-driven world.

Peer review is also an essential part of academic research.

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
    - *stringent or lenient:* Reviewers have different standard and bias.

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.
- Typical calibration techniques:

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.
- Typical calibration techniques:
  - Averaging reviewers' scores

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.
- Typical calibration techniques:
  - Averaging reviewers' scores
  - Open discussions and expert advice (e.g. Area Chairs)

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.
- Typical calibration techniques:
  - Averaging reviewers' scores
  - Open discussions and expert advice (e.g. Area Chairs)
- Challenges:

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.
- Typical calibration techniques:
  - Averaging reviewers' scores
  - Open discussions and expert advice (e.g. Area Chairs)
- Challenges:
  - Sparsity of review data

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.
- Typical calibration techniques:
  - Averaging reviewers' scores
  - Open discussions and expert advice (e.g. Area Chairs)
- Challenges:
  - Sparsity of review data
  - Human factors

Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.
- Typical calibration techniques:
  - Averaging reviewers' scores
  - Open discussions and expert advice (e.g. Area Chairs)
- Challenges:
  - Sparsity of review data
  - Human factors
  - Inflexibility

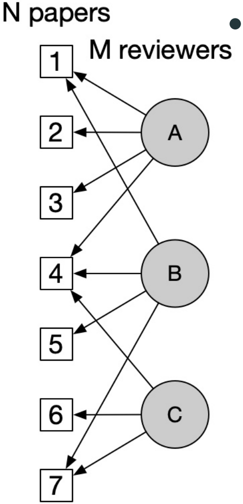
Calibration for Peer Review

"Your 2 is My 1, Your 3 is My 9." [WS18]

- *Miscalibration* is a prevalent problem.
  - *stringent or lenient:* Reviewers have different standard and bias.
  - *perception error:* We all make mistakes.
- Typical calibration techniques:
  - Averaging reviewers' scores
  - Open discussions and expert advice (e.g. Area Chairs)
- Challenges:
  - Sparsity of review data
  - Human factors
  - Inflexibility

We propose an optimization-driven framework to mitigate miscalibration.

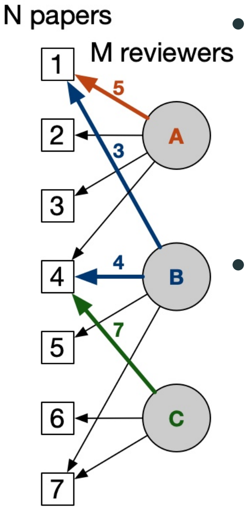
Problem Formulation



- Reviewer j reviews a subset of the papers $I_j \subseteq [N]$.

$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$

Problem Formulation



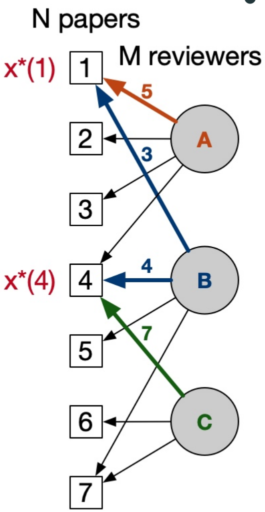
- Reviewer j reviews a subset of the papers $I_j \subseteq [N]$.

$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$

- I_j^ℓ, y_j^ℓ denotes *index*/*score* of ℓth highest paper scored by reviewer j

$$I_A^1 = I_B^2 = 1, \quad y_A^1 = 5, y_B^2 = 3$$
$$I_B^1 = I_C^1 = 4, \quad y_B^1 = 4, y_C^1 = 7$$

Problem Formulation



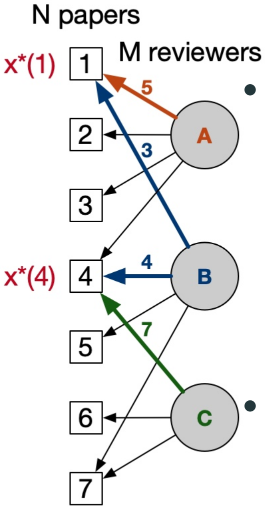
- Common hypothesis on score generation process [GWG, RRS11, BK13, MKLP17, WSWS20]

$$y_j^\ell := f_j(x^*(I_j^\ell) + \epsilon_j^\ell)$$

where ϵ_j^ℓ is independent zero-mean Gaussian noise, $x^*(i)$ is paper i's *unknown* ground-truth quality, f_j is reviewer j's scoring function.

$$I_B^1 = I_C^1 = 4$$
$$y_B^1 = f_B(x^*(4) + \epsilon_B^1)$$
$$y_C^1 = f_C(x^*(4) + \epsilon_C^1)$$

Problem Formulation



- Common hypothesis on score generation process [GWG, RRS11, BK13, MKLP17, WSWS20]

$$y_j^\ell := f_j(x^*(I_j^\ell) + \epsilon_j^\ell)$$

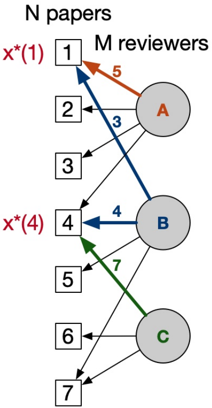
where ϵ_j^ℓ is independent zero-mean Gaussian noise, $x^*(i)$ is paper i's *unknown* ground-truth quality, f_j is reviewer j's scoring function.

- It would be an intractable *Matrix Seriation* problem, if ϵ_j^ℓ is modeled outside f_j.

Problem Formulation

Input:

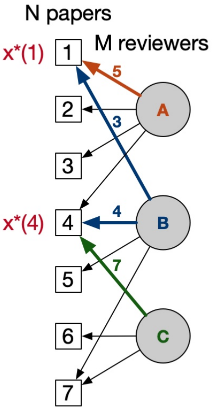
- Paper assignments $\{I_j\}_{j \in [N]}$



Problem Formulation

Input:

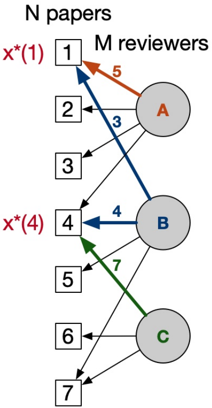
- Paper assignments $\{I_j\}_{j \in [N]}$
- Review scores $\{y_j^\ell\}_{j \in [M], \ell \in [I_j]}$



Problem Formulation

Input:

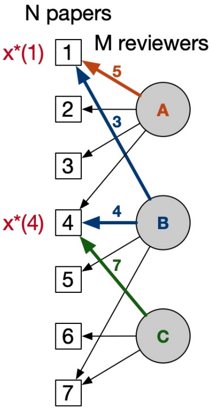
- Paper assignments $\{I_j\}_{j \in [N]}$
- Review scores $\{y_j^{\ell}\}_{j \in [M], \ell \in [I_j]}$



Problem Formulation

Input:

- Paper assignments $\{I_j\}_{j \in [N]}$
- Review scores $\{y_j^\ell\}_{j \in [M], \ell \in [I_j]}$
- Threshold parameter $n \leq N$

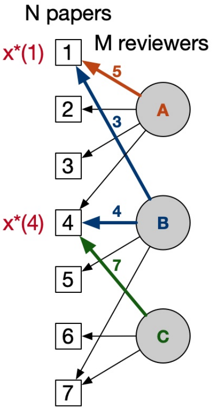


Problem Formulation

Input:

- Paper assignments $\{I_j\}_{j\in[N]}$
- Review scores $\{y_j^\ell\}_{j\in[M],\ell\in[I_j]}$
- Threshold parameter $n \leq N$

Output: a set S of n items



Problem Formulation

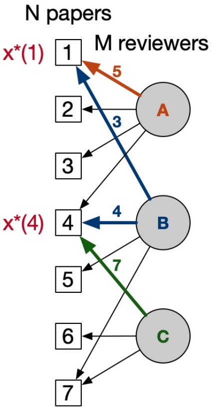
Input:

- Paper assignments $\{I_j\}_{j \in [N]}$
- Review scores $\{y_j^\ell\}_{j \in [M], \ell \in [I_j]}$
- Threshold parameter $n \leq N$

Output: a set S of n items

Objective:

S matches with ground-truth top n items based on x^*.



Problem Formulation

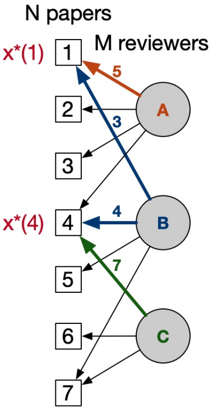
Input:

- Paper assignments $\{I_j\}_{j \in [N]}$
- Review scores $\{y_j^\ell\}_{j \in [M], \ell \in [I_j]}$
- Threshold parameter $n \leq N$

Output: a set S of n items

Objective:

S matches with ground-truth top n items based on x^*.



To identify the papers with the best true qualities.

Methods

Least Square Calibration (LSC)

$$\min_{\mathbf{x},\mathbf{f},\epsilon} \quad \sum_{j=1}^{M} \sum_{\ell=1}^{|I_j|} (\epsilon_j^\ell)^2$$

$$\text{s.t.} \quad y_j^\ell = f_j\left(x(I_j^\ell) + \epsilon_j^\ell\right) \text{ and } f_j \in \mathcal{H} \qquad \forall j \in [M], \ell \leq |I_j|$$

Least Square Calibration (LSC)

$$\min_{\mathbf{x},\mathbf{f},\epsilon} \quad \sum_{j=1}^{M}\sum_{\ell=1}^{|I_j|}(\epsilon_j^\ell)^2$$

$$\text{s.t.} \quad y_j^\ell = f_j\left(x(I_j^\ell)+\epsilon_j^\ell\right) \text{ and } f_j \in \mathcal{H} \qquad \forall j \in [M], \ell \leq |I_j|$$

Interpretations:

- **Unsupervised Learning:**
  Given hypothesis class $\mathcal{H}$, find $f_1, \cdots, f_M \in \mathcal{H}$ and true qualities $\mathbf{x}$ with the least noise to match with review scores $\{y_j^\ell\}_{j\in[M],\ell\in[I_j]}$.

Least Square Calibration (LSC)

$$\min_{\mathbf{x},\mathbf{f},\epsilon} \quad \sum_{j=1}^{M}\sum_{\ell=1}^{|I_j|}(\epsilon_j^\ell)^2$$

$$\text{s.t.} \quad y_j^\ell = f_j\left(x(I_j^\ell)+\epsilon_j^\ell\right) \text{ and } f_j \in \mathcal{H} \qquad \forall j \in [M], \ell \le |I_j|$$

Interpretations:

- **Unsupervised Learning:**
  Given hypothesis class $\mathcal{H}$, find $f_1, \cdots, f_M \in \mathcal{H}$ and true qualities $\mathbf{x}$ with the least noise to match with review scores $\{y_j^\ell\}_{j\in[M],\ell\in[I_j]}$.
- **MLE:**
  Find parameters $\mathbf{x}$, $\mathbf{f}$ to maximize likelihood of observation y_j^ℓ under Gaussian noise ϵ_j^ℓ.

Least Square Calibration (LSC)

$$\min_{\mathbf{x},\mathbf{f},\epsilon} \quad \sum_{j=1}^{M}\sum_{\ell=1}^{|I_j|}(\epsilon_j^\ell)^2$$

$$\text{s.t.} \quad y_j^\ell = f_j\left(x(I_j^\ell)+\epsilon_j^\ell\right) \text{ and } f_j \in \mathcal{H} \qquad \forall j \in [M], \ell \le |I_j|$$

Interpretations:

- **Unsupervised Learning:**
  Given hypothesis class $\mathcal{H}$, find $f_1, \cdots, f_M \in \mathcal{H}$ and true qualities $\mathbf{x}$ with the least noise to match with review scores $\{y_j^\ell\}_{j\in[M],\ell\in[I_j]}$.
- **MLE:**
  Find parameters $\mathbf{x}$, $\mathbf{f}$ to maximize likelihood of observation y_j^ℓ under Gaussian noise ϵ_j^ℓ.

But *functional optimization problem* is intractable in general?

LSC under different hypothesis classes

Suppose $\mathcal{H} = \{f : f(x) = ax + b \mid a \geq 0, b \in \mathbb{R}\}$,

$$\min_{\mathbf{x},\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M} \sum_{\ell=1}^{|I_j|} (\epsilon_j^\ell)^2 \qquad \textbf{LSC (linear)}$$

$$\text{s.t.} \quad y_j^\ell = \alpha_j \cdot (x(I_j^\ell) + \epsilon_j^\ell) + \beta_j \qquad \forall j \in [M], \ell \leq |I_j|$$

LSC is reduced to a simple quadratic program.

LSC under different hypothesis classes

Suppose $\mathcal{H} = \{f : f(x) = ax + b \mid a \geq 0, b \in \mathbb{R}\}$,

$$\min_{\mathbf{x},\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M} \sum_{\ell=1}^{|I_j|} (\epsilon_j^\ell)^2 \qquad \textbf{LSC (linear)}$$

$$\text{s.t.} \quad y_j^\ell = \alpha_j \cdot (x(I_j^\ell) + \epsilon_j^\ell) + \beta_j \qquad \forall j \in [M], \ell \leq |I_j|$$

LSC is reduced to a simple quadratic program.

- In fact, we can solve LSC efficiently for any monotone function, any linear scoring function, convex/concave scoring function as well as their mixture.

LSC under different hypothesis classes

$$\min_{\mathbf{x},\epsilon} \quad \sum_{j=1}^{M} \sum_{\ell=1}^{|I_j|} (\epsilon_j^\ell)^2 \qquad \textbf{LSC (mono)}$$

$$\text{s.t.} \quad \widetilde{x}_j^\ell = x(I_j^\ell) + \epsilon_j^\ell \qquad \forall j \in [M], 1 \le \ell \le |I_j|$$

$$\widetilde{x}_j^\ell - \widetilde{x}_j^{\ell-1} \ge \frac{y_j^\ell - y_j^{\ell-1}}{C} \qquad \forall j \in [M], 2 \le \ell \le |I_j|$$

- In fact, we can solve LSC efficiently for any monotone function, any linear scoring function, convex/concave scoring function as well as their mixture.

LSC under different hypothesis classes

$$\min_{\mathbf{x},\epsilon} \quad \sum_{j=1}^{M}\sum_{\ell=1}^{|I_j|}(\epsilon_j^\ell)^2 \qquad \textbf{LSC (convex)}$$

$$\text{s.t.} \quad \widetilde{x}_j^\ell - \widetilde{x}_j^{\ell-1} \geq 1 \qquad \forall j \in [M], 2 \leq \ell \leq |I_j|$$

$$\frac{\widetilde{x}_j^\ell - \widetilde{x}_j^{\ell-1}}{y_j^\ell - y_j^{\ell-1}} \leq \frac{\widetilde{x}_j^{\ell+1} - \widetilde{x}_j^\ell}{y_j^{\ell+1} - y_j^\ell} \qquad \forall j \in [M], 2 \leq \ell \leq |I_j| - 1$$

- In fact, we can solve LSC efficiently for any monotone function, any linear scoring function, convex/concave scoring function as well as their mixture.

LSC under different hypothesis classes

$$\min_{\mathbf{x},\mathbf{f},\epsilon} \sum_{j=1}^{M} \sum_{\ell=1}^{|I_j|} (\epsilon_j^\ell)^2 \qquad \textbf{LSC (mix)}$$

$$\text{s.t.} \quad y_j^\ell = f_j\left(x(I_j^\ell) + \epsilon_j^\ell\right) \qquad \forall j \in [M], \ell \le |I_j|$$

$$f_j \in \mathcal{H}^{\textbf{mono}}, f_k \in \mathcal{H}^{\textbf{linear}}, f_p \in \mathcal{H}^{\textbf{convex}}, f_q \in \mathcal{H}^{\textbf{concave}}$$

- In fact, we can solve LSC efficiently for any monotone function, any linear scoring function, convex/concave scoring function as well as their mixture.

Hence, LSC framework is *adaptive* to different levels of prior knowledge.

Linear Regression and LSC with linear scoring functions

$$\min_{\mathbf{x},\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M}\sum_{\ell=1}^{|I_j|}(\epsilon_j^\ell)^2 \qquad \textbf{LSC (linear)}$$

$$\text{s.t.} \quad y_j^\ell = \alpha_j \cdot (x(I_j^\ell) + \epsilon_j^\ell) + \beta_j \qquad \forall j \in [M], \ell \le |I_j|$$

$$\min_{\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M}(\epsilon^\ell)^2 \qquad \textbf{Ordinary Linear Regression (OLS)}$$

$$\text{s.t.} \quad y^\ell = \alpha \cdot x^\ell + \beta + \epsilon^\ell \qquad \forall \ell$$

Linear Regression and LSC with linear scoring functions

$$\min_{\mathbf{x},\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M}\sum_{\ell=1}^{|I_j|}(\epsilon_j^\ell)^2 \qquad \textbf{LSC (linear)}$$

$$\text{s.t.} \quad y_j^\ell = \alpha_j \cdot (x(I_j^\ell) + \epsilon_j^\ell) + \beta_j \qquad \forall j \in [M], \ell \leq |I_j|$$

$$\min_{\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M}(\epsilon^\ell)^2 \qquad \textbf{Ordinary Linear Regression (OLS)}$$

$$\text{s.t.} \quad y^\ell = \alpha \cdot x^\ell + \beta + \epsilon^\ell \qquad \forall \ell$$

1. $\mathbf{x}$ is known in OLS, but unknown in LSC *(unsupervised)*.

Linear Regression and LSC with linear scoring functions

LSC (linear)

$$\min_{\mathbf{x},\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M}\sum_{\ell=1}^{|I_j|}(\epsilon_j^\ell)^2$$

$$\text{s.t.} \quad y_j^\ell = \alpha_j \cdot (x(I_j^\ell) + \epsilon_j^\ell) + \beta_j \qquad \forall j \in [M], \ell \le |I_j|$$

Ordinary Linear Regression (OLS)

$$\min_{\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M}(\epsilon^\ell)^2$$

$$\text{s.t.} \quad y^\ell = \alpha \cdot x^\ell + \beta + \epsilon^\ell \qquad \forall \ell$$

1. $\mathbf{x}$ is known in OLS, but unknown in LSC *(unsupervised)*.
2. LSC models the extra structure in the paper assignments:
   Paper i have consistent x_i; Reviewer j have consistent f_j.

Linear Regression and LSC with linear scoring functions

$$\min_{\mathbf{x},\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M} \sum_{\ell=1}^{|I_j|} (\epsilon_j^\ell)^2 \qquad \textbf{LSC (linear)}$$

$$\text{s.t.} \quad y_j^\ell = \alpha_j \cdot (x(I_j^\ell) + \epsilon_j^\ell) + \beta_j \qquad \forall j \in [M], \ell \leq |I_j|$$

$$\min_{\alpha,\beta,\epsilon} \quad \sum_{j=1}^{M} (\epsilon^\ell)^2 \qquad \textbf{Ordinary Linear Regression (OLS)}$$

$$\text{s.t.} \quad y^\ell = \alpha \cdot x^\ell + \beta + \epsilon^\ell \qquad \forall \ell$$

1. $\mathbf{x}$ is known in OLS, but unknown in LSC *(unsupervised)*.
2. LSC models the extra structure in the paper assignments:
   Paper i have consistent x_i; Reviewer j have consistent f_j.

How does LSC guarantee that $\mathbf{x}$, $\mathbf{f}$ is necessarily ground-truth $\mathbf{x}^*$, $\mathbf{f}^*$?

When/Why LSC works?

Without perception noise, LSC is reduced to a linear feasibility problem:

When/Why LSC works?

Without perception noise, LSC is reduced to a linear feasibility problem:

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & 0 \\
\text{s.t.} \quad & x(I_j^{\ell}) - x(I_j^{\ell-1}) \geq 1 & \forall j \in [M], \forall 2 \leq \ell \leq |I_j| \\
& \frac{x(I_j^{\ell}) - x(I_j^{\ell-1})}{y_j^{\ell} - y_j^{\ell-1}} = \frac{x(I_j^{\ell+1}) - x(I_j^{\ell})}{y_j^{\ell+1} - y_j^{\ell}} & \forall j \in [M], 2 \leq \ell \leq |I_j| - 1
\end{aligned}
$$

When/Why LSC works?

Without perception noise, LSC is reduced to a linear feasibility problem:

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & 0 \\
\text{s.t.} \quad & x(I_j^{\ell}) - x(I_j^{\ell-1}) \geq 1 && \forall j \in [M], \forall 2 \leq \ell \leq |I_j| \\
& \frac{x(I_j^{\ell}) - x(I_j^{\ell-1})}{y_j^{\ell} - y_j^{\ell-1}} = \frac{x(I_j^{\ell+1}) - x(I_j^{\ell})}{y_j^{\ell+1} - y_j^{\ell}} && \forall j \in [M], 2 \leq \ell \leq |I_j| - 1
\end{aligned}
$$

The effectiveness of calibration depends on the assignment:

- Reviewer j reviewed only one paper that are also reviewed by others.

When/Why LSC works?

Without perception noise, LSC is reduced to a linear feasibility problem:

$$\begin{aligned}
\min_{\mathbf{x}} \quad & 0 \\
\text{s.t.} \quad & x(I_j^{\ell}) - x(I_j^{\ell-1}) \geq 1 && \forall j \in [M], \forall 2 \leq \ell \leq |I_j| \\
& \frac{x(I_j^{\ell}) - x(I_j^{\ell-1})}{y_j^{\ell} - y_j^{\ell-1}} = \frac{x(I_j^{\ell+1}) - x(I_j^{\ell})}{y_j^{\ell+1} - y_j^{\ell}} && \forall j \in [M], 2 \leq \ell \leq |I_j| - 1
\end{aligned}$$

The effectiveness of calibration depends on the assignment:

- Reviewer j reviewed only one paper that are also reviewed by others.
- Even if we know the ground-truth quality of this paper,

When/Why LSC works?

Without perception noise, LSC is reduced to a linear feasibility problem:

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & 0 \\
\text{s.t.} \quad & x(I_j^{\ell}) - x(I_j^{\ell-1}) \geq 1 && \forall j \in [M], \forall 2 \leq \ell \leq |I_j| \\
& \frac{x(I_j^{\ell}) - x(I_j^{\ell-1})}{y_j^{\ell} - y_j^{\ell-1}} = \frac{x(I_j^{\ell+1}) - x(I_j^{\ell})}{y_j^{\ell+1} - y_j^{\ell}} && \forall j \in [M], 2 \leq \ell \leq |I_j| - 1
\end{aligned}
$$

The effectiveness of calibration depends on the assignment:

- Reviewer j reviewed only one paper that are also reviewed by others.
- Even if we know the ground-truth quality of this paper,
- there are infinitely many feasible $x(I_j^{\ell}), f_j$ to match observation y_j^{ℓ}.

When/Why LSC works?

Without perception noise, LSC is reduced to a linear feasibility problem:

$$\begin{aligned}
\min_{\mathbf{x}} \quad & 0 \\
\text{s.t.} \quad & x(I_j^\ell) - x(I_j^{\ell-1}) \geq 1 && \forall j \in [M], \forall 2 \leq \ell \leq |I_j| \\
& \frac{x(I_j^\ell) - x(I_j^{\ell-1})}{y_j^\ell - y_j^{\ell-1}} = \frac{x(I_j^{\ell+1}) - x(I_j^\ell)}{y_j^{\ell+1} - y_j^\ell} && \forall j \in [M], 2 \leq \ell \leq |I_j| - 1
\end{aligned}$$

The effectiveness of calibration depends on the assignment:

- Reviewer j reviewed only one paper that are also reviewed by others.
- Even if we know the ground-truth quality of this paper,
- there are infinitely many feasible $x(I_j^\ell), f_j$ to match observation y_j^ℓ.

What assignment rule do we need?

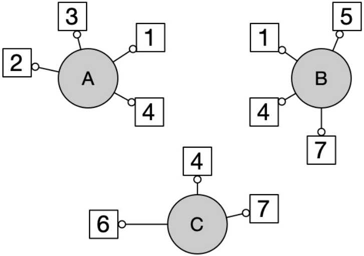
Perfect Recovery and Doubly-Connected Review Graph

$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$

Review Graph.

Perfect Recovery and Doubly-Connected Review Graph

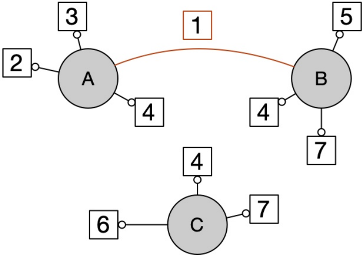
$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$



Review Graph. reviewer as vertex,

Perfect Recovery and Doubly-Connected Review Graph

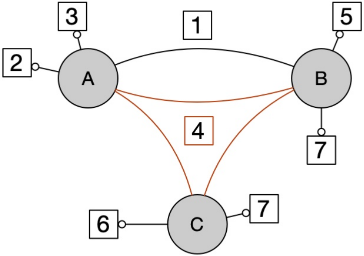
$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$



Review Graph. reviewer as vertex, commonly reviewed paper as edge.

Perfect Recovery and Doubly-Connected Review Graph

$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$

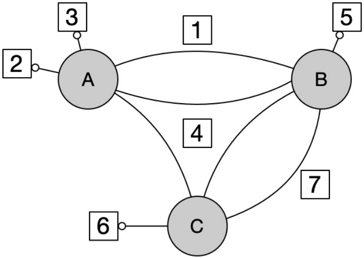


Review Graph. reviewer as vertex, commonly reviewed paper as edge.

Perfect Recovery and Doubly-Connected Review Graph

$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$

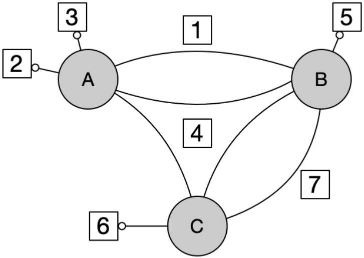
Review Graph. reviewer as vertex, commonly reviewed paper as edge.



Perfect Recovery and Doubly-Connected Review Graph

$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$

Review Graph. reviewer as vertex, commonly reviewed paper as edge.



Theorem (Informal)

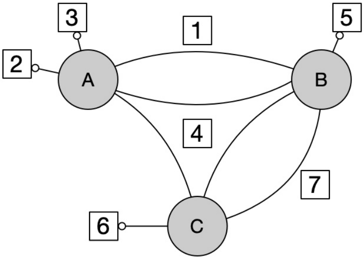
LSC perfectly recovers a review graph G iff. G has a doubly-connected component S that covers all papers, i.e., $\bigcup_{i \in S} I_i = [N]$.

Perfect Recovery and Doubly-Connected Review Graph

$I_A = \{1, 2, 3, 4\}$

$I_B = \{1, 4, 5, 7\}$

$I_C = \{4, 6, 7\}$



Review Graph. reviewer as vertex, commonly reviewed paper as edge.

This instance forms a doubly-connected graph.

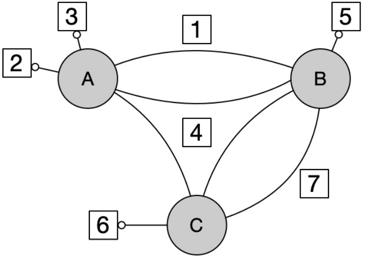
Theorem (Informal)

LSC perfectly recovers a review graph G iff. G has a doubly-connected component S that covers all papers, i.e., $\bigcup_{i \in S} I_i = [N]$.

The notion of *double-connectivity* generalizes from single-connectivity.

Perfect Recovery and Doubly-Connected Review Graph

$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$



Review Graph. reviewer as vertex, commonly reviewed paper as edge.

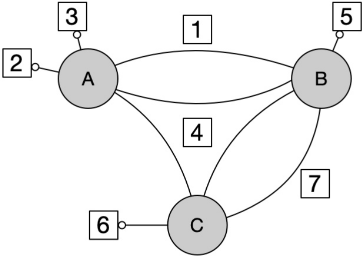
Theorem (Informal)

LSC perfectly recovers a review graph G iff. G has a doubly-connected component S that covers all papers, i.e., $\bigcup_{i \in S} I_i = [N]$.

Remark. Paper assignment matters for successful calibration.

Perfect Recovery and Doubly-Connected Review Graph

$$I_A = \{1, 2, 3, 4\}$$
$$I_B = \{1, 4, 5, 7\}$$
$$I_C = \{4, 6, 7\}$$



Review Graph. reviewer as vertex, commonly reviewed paper as edge.

Theorem (Informal)

LSC perfectly recovers a review graph G iff. G has a doubly-connected component S that covers all papers, i.e., $\bigcup_{i \in S} I_i = [N]$.

Remark. Paper assignment matters for successful calibration.

A justification for the extra reviews in post-rebuttal discussions!

Experiments

Experiment Setup

Datasets:

- Synthesized Conference Review Data (due to lack of $\mathbf{x}^*$)

Experiment Setup

Datasets:

- Synthesized Conference Review Data (due to lack of $\mathbf{x}^*$)
- Peer-Grading Dataset [SAvL16]

Experiment Setup

Datasets:

- Synthesized Conference Review Data (due to lack of $\mathbf{x}^*$)
- Peer-Grading Dataset [SAvL16]

Baselines:

- **Average**, the most common heuristic

Experiment Setup

Datasets:

- Synthesized Conference Review Data (due to lack of $\mathbf{x}^*$)
- Peer-Grading Dataset [SAvL16]

Baselines:

- **Average**, the most common heuristic
- The quadratic program (**QP**) proposed by [RRS11]
- The bayesian model (**Bayesian**) proposed by [GWG]

Experiment Setup

Datasets:

- Synthesized Conference Review Data (due to lack of $\mathbf{x}^*$)
- Peer-Grading Dataset [SAvL16]

Baselines:

- **Average**, the most common heuristic
- The quadratic program (**QP**) proposed by [RRS11]
- The bayesian model (**Bayesian**) proposed by [GWG]

Metrics:

- Precision, *the percentage of selected ground-truth top papers*
- Ranking-based metrics such as NDCG

Experiment Results

Table 1: Performance on Conference Review (**L**) and Peer-Grading (**R**) dataset

Model \ Metric	Pre. (%)	NDCG (%)	Pre. (%)	NDCG (%)
Average	39.2	45.8	0.80	0.34
QP	69.2	68.9	0.80	0.82
Bayesian	71.5	71.4	0.78	0.71
LSC (mono)	75.9	79.2	0.78	0.81
LSC (linear)	**80.1**	**84.7**	**0.82**	**0.85**

Conference review data is generated with random linear scoring function with perception noisy ($\sigma = 0.5$).

Experiment Results

Table 1: Performance on Conference Review (**L**) and Peer-Grading (**R**) dataset

Model \ Metric	Pre. (%)	NDCG (%)	Pre. (%)	NDCG (%)
Average	39.2	45.8	0.80	0.34
QP	69.2	68.9	0.80	0.82
Bayesian	71.5	71.4	0.78	0.71
LSC (mono)	75.9	79.2	0.78	0.81
LSC (linear)	**80.1**	**84.7**	**0.82**	**0.85**

We use the TA's grade as the ground truth quality x^*.

Experiment Results

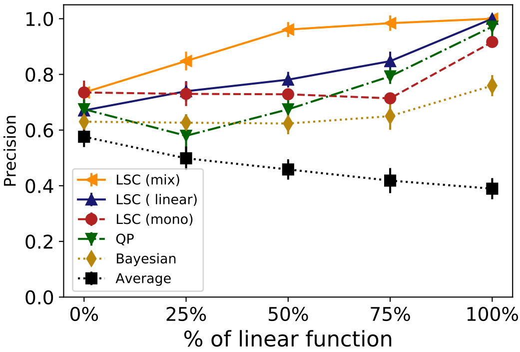
Table 1: Performance on Conference Review (**L**) and Peer-Grading (**R**) dataset

Model \ Metric	Pre. (%)	NDCG (%)	Pre. (%)	NDCG (%)
Average	39.2	45.8	0.80	0.34
QP	69.2	68.9	0.80	0.82
Bayesian	71.5	71.4	0.78	0.71
LSC (mono)	75.9	79.2	0.78	0.81
LSC (linear)	**80.1**	**84.7**	**0.82**	**0.85**

LSC consistently outperforms other baselines on both datasets.

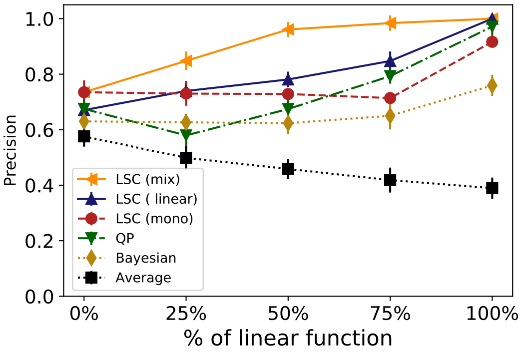
Robustness to Mis-Specified Prior Knowledge

Figure 1: Performance comparisons in mixed setups



Robustness to Mis-Specified Prior Knowledge

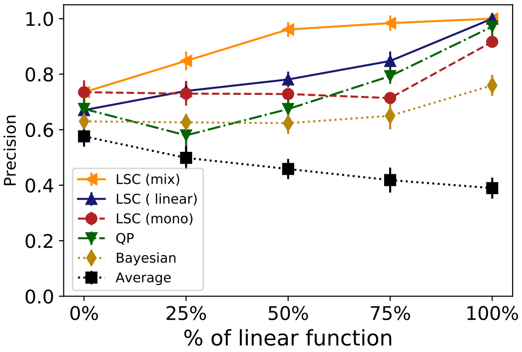
Figure 1: Performance comparisons in mixed setups



- LSC (mix) has the best performances with full prior knowledge.

Robustness to Mis-Specified Prior Knowledge

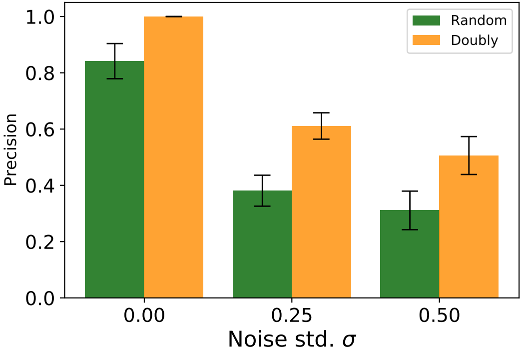
Figure 1: Performance comparisons in mixed setups



- LSC (mix) has the best performances with full prior knowledge.
- LSC (linear) is robust under mis-specified prior knowledge.

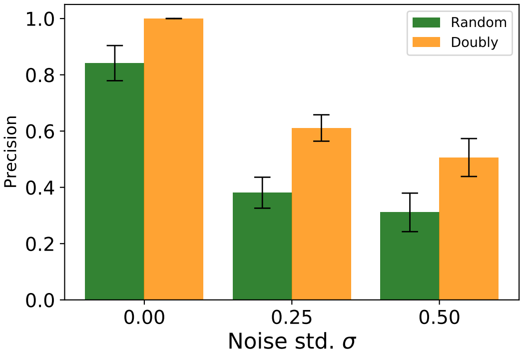
Double Connectivity against Perception Noise

Figure 2: Performance in review graphs of different connectivity and noise level



Double Connectivity against Perception Noise

Figure 2: Performance in review graphs of different connectivity and noise level



Review assignments with double-connectivity can help LSC calibrate.

Conclusion

Conclusion

- LSC is a simple yet powerful unsupervised learning framework for calibration in peer review system.

Conclusion

- LSC is a simple yet powerful unsupervised learning framework for calibration in peer review system.
- It exploits both the robustness of linear regression methods and the topological structure of review graphs.

Conclusion

- LSC is a simple yet powerful unsupervised learning framework for calibration in peer review system.
- It exploits both the robustness of linear regression methods and the topological structure of review graphs.
- We provide a general guideline on the assignment rules in peer review for more effective calibration.

Conclusion

- LSC is a simple yet powerful unsupervised learning framework for calibration in peer review system.
- It exploits both the robustness of linear regression methods and the topological structure of review graphs.
- We provide a general guideline on the assignment rules in peer review for more effective calibration.
- We wish to apply our LSC framework in real conferences!

References i

- Yukino Baba and Hisashi Kashima, *Statistical quality estimation for general crowdsourcing tasks*, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 554–562.
- Hong Ge, Max Welling, and Zoubin Ghahramani, *A bayesian model for calibrating reviewer scores*.
- Robert S MacKay, Ralph Kenna, Robert J Low, and Sarah Parker, *Calibration with confidence: a principled method for panel assessment*, Royal Society open science **4** (2017), no. 2, 160760.

References ii

- Magnus Roos, Jörg Rothe, and Björn Scheuermann, *How to calibrate the scores of biased reviewers by quadratic programming*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 25, 2011.
- Mehdi SM Sajjadi, Morteza Alamgir, and Ulrike von Luxburg, *Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines*, Proceedings of the third (2016) ACM conference on Learning@ Scale, 2016, pp. 369–378.
- Jingyan Wang and Nihar B. Shah, *Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings*, 2018.

References iii

Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar B Shah, *Debiasing evaluations that are biased by evaluations*, arXiv preprint arXiv:2012.00714 (2020).