Eliciting and Aggregating Information: An Information Theoretic Approach

by

Yuqing Kong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2018

Doctoral Committee:

Assistant Professor Grant Schoenebeck, Chair
Assistant Professor Walter Lasecki
Professor Mingyan Liu
Professor David Parkes

Yuqing Kong

yuqkong@umich.edu

ORCID iD: 0000-0002-5901-3004

Dedicated to my mom and dad

ACKNOWLEDGMENTS

To my advisor, Grant Schoenebeck: thank you for the time, support and encouragement you have given to me especially when I was upset. I am so lucky to have you as my advisor! I have learned everything (including English ☺) from you and I could not have imagined having a better advisor and mentor for my Ph.D study. My goal is to become an advisor that is as good as you!

I would also like to thank my committee members: Professor David Parkes, Professor Mingyan Liu, Professor Walter Lasecki, for generously offering their time, support, guidance and insightful comments.

I am also grateful to my friends/roommates who have supported me along the way, to the faculty in the theory group and the university staff who have provided me great help during my Ph.D study. I want to thank the University of Michigan as well which is very pretty and provided great facilities (e.g. Duderstadt Center, Chez Betty) to me.

I have a very special gratitude to Professor David Parkes and Professor Yiling Chen: you two have made a great difference to my Ph.D life and also my future, thank you!

Finally, I want to thank my parents and my family. To my parents: I am so proud and lucky to have you two as my parents. Both of you are loved not only by me but also by many students. My dream is to become a teacher that is as good as you two and can provide help and support to many students like you two. I have the best parents in the world who are the best teachers in the world.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

Crowdsourcing—outsourcing tasks to a crowd of workers (e.g. Amazon Mechanical Turk, peer grading for massive open online courseware (MOOCs), scholarly peer review, and Yahoo answers)—is a fast, cheap, and effective method for performing simple tasks even at large scales. Two central problems in this area are:

(1) Information Elicitation how to design reward systems that incentivize high quality feedback from agents; and

(2) Information Aggregation how to aggregate the collected feedback to obtain a high quality forecast.

This thesis shows that the combination of game theory, information theory, and learning theory can bring a unified framework to both of the central problems in crowdsourcing area. This thesis builds a natural connection between information elicitation and information aggregation, distills the essence of eliciting and aggregating information to the design of proper information measurements and applies the information measurements to both the central problems:

In the setting where information cannot be verified, this thesis proposes a simple yet powerful information theoretical framework, the *Mutual Information Paradigm (MIP)*, for information elicitation mechanisms. The framework pays every agent a measure of mutual information between her signal and a peer's signal. The mutual information measurement is required to have the key property that any "data processing" on the two random variables will decrease the mutual information between

them. We identify such information measures that generalize Shannon mutual information. MIP overcomes the two main challenges in information elicitation without verification: (1) how to incentivize effort and avoid agents colluding to report random or identical responses (2) how to motivate agents who believe they are in the minority to report truthfully.

To elicit expertise without verification, this thesis also defines a natural model for this setting based on the assumption that *more sophisticated agents know the beliefs of less sophisticated agents* and extends MIP to a mechanism design framework, the *Hierarchical Mutual Information Paradigm (HMIP)*, for this setting.

Aided by the information measures and the frameworks, this thesis (1) designs several novel information elicitation mechanisms (e.g. the disagreement mechanism, the f-mutual information mechanism, the multi-hierarchical mutual information mechanism, the common ground mechanism) in various of settings such that honesty and efforts are incentivized and expertise is identified; (2) addresses an important unsupervised learning problem—co-training by reducing it to an information elicitation problem—forecast elicitation without verification.

CHAPTER I

Introduction

Crowdsourcing, outsourcing tasks to a crowd of workers (e.g. Amazon Mechanical Turk, peer grading for massive open online courses, scholarly peer review, and Yahoo answers), is a fast, cheap, and effective method for performing simple tasks even at large scales. To attract a large number of workers, crowdsourcing is usually open to the public rather than just professional experts. Two central problems in this area are:

(1) Information Elicitation how to design reward systems that incentivize high quality feedback from agents, even when the information is unverifiable; and

(2) Information Aggregation how to aggregate the collected feedback to obtain a high quality forecast.

The elicitation and aggregation of information play a central role in many decision-making contexts (e.g. reputation systems, purchasing, product development, pricing), and also deal with a key challenge in big data—the lack of labeled data: crowdsourcing can be used to provide a massive amount of noisy labels; and the effective use of the mass of noisy labels is an information aggregation problem. The two central problems are hard especially when the information is unverifiable. However, in the context of many applications, the information is expensive or even impossible to

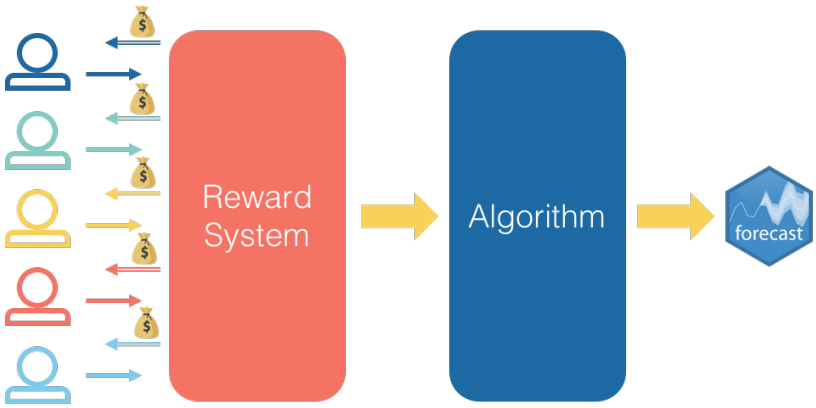

Figure 1.1: Eliciting and aggregating information

verify (e.g. do you like this restaurant? Y/N). In the situation where the lack of labeled data is a key challenge, it is expensive to obtain the ground truth or to verify the information, otherwise the lack of labeled data would not be a problem. Prediction markets/stock markets and spot-checking (randomly picking some questions and checking the answers provided by the participant) are two examples that elicit information with (possibly future) verification. Supervised learning is an example of information aggregation with verification. This thesis focuses on the elicitation and aggregation of information *without verification*.

Information elicitation is a mechanism design problem which is closely related to game theory and information aggregation is an algorithmic design problem which is closely related to learning theory. The subject we are dealing with is information which is related to information theory. Note that recently, the combination of game theory and learning theory has made innovative progress (e.g. Generative Adversarial Networks). Thus, a central contention of this thesis is that the combination of game theory, information theory, and learning theory can bring a unified framework to both of the central problems in crowdsourcing. Few previous works have deeply connected the three fields in crowdsourcing or any other area while this thesis builds several simple yet powerful connections among game theory, learning theory and

information theory to solve several main challenges in the elicitation and aggregation of information.

Thesis statement This thesis shows that the combination of game theory, information theory, and learning theory can bring a unified framework to both of the central problems in crowdsourcing area. This thesis builds a natural connection between information elicitation and information aggregation, distills the essence of eliciting and aggregating information to the design of proper information measurements and applies the information measurements to both the central problems.

Main contribution In the setting where information cannot be verified, this thesis proposes a simple yet powerful information theoretical frameworks, the *Mutual Information Paradigm (MIP)*, for information elicitation mechanisms. The framework pays every agent a measure of mutual information between her signal and a peer's signal. The mutual information measurement is required to have the key property that any "data processing" on the two random variables will decrease the mutual information between them. We identify such information measures that generalize Shannon mutual information. MIP overcomes the two main challenges in information elicitation without verification: (1) how to incentivize effort and avoid agents colluding to report random or identical responses; (2) how to motivate agents who believe they are in the minority to report truthfully.

To elicit expertise without verification, this thesis also defines a natural model for this setting based on the assumption that *more sophisticated agents know the beliefs of less sophisticated agents* and extends MIP to a mechanism design framework, the *Hierarchical Mutual Information Paradigm (HMIP)*, for this setting.

Aided by the information measures and the frameworks, this thesis (1) designs several novel information elicitation mechanisms (*disagreement mechanism* (Section 4.3), *f-mutual information mechanism* (Section 3.3), *multi-HMIM* (Section 5.2.2), *Learn-*

ing based multi-HMIM (Section 5.2.3), *single-HMIM* (Section 5.3.2), *common ground mechanism* (Section 6.5.2) and *multi-task common ground mechanism* (Section 6.5.1)) in various of settings such that honesty and efforts are incentivized and expertise is identified; (2) addresses an important unsupervised learning problem—co-training by reducing it to an information elicitation problem—forecast elicitation without verification.

1.1 Eliciting information without verification

User feedback requests (e.g Ebay's reputation system) are increasingly prominent and important. However, the overwhelming number of requests can lead to low participation rates, which in turn may yield unrepresentative samples. To encourage participation, a system can reward people for answering requests. But this may cause perverse incentives: some people may answer a large of number of questions simply for the reward and without making any attempt to answer accurately. Moreover, people may be motivated to lie when they face a potential loss of privacy or can benefit in the future by lying now. It is thus important to develop reward systems that motivate honesty. If we can verify the information people provide in the future (e.g prediction markets), we can motivate honesty via this future verification. However, sometimes we need to elicit information without verification since the objective truth is hard to access (e.g. a self-report survey for unethical activities) or even does not exist (e.g. subjective ratings). This thesis focuses on the situation where the objective truth is not observable—peer prediction [45]. A key problem in peer prediction literature is:

(+) **how to motivate honest reporting without verification?**

Two main challenges in solving problem (+) are: without verification,

1. (avoiding collusion) *how to avoid colluding agents who report random or identical responses; and*

2. (motivating the minority) *how to motivate agents who believe they are in the minority to report truthfully.*

Traditional reward systems (e.g. flat payment, majority vote, spot-checking) fail to solve problem (+) since they either distort users' incentives (e.g. flat payment, majority vote) or require partial verification and expensive gold-standard questions (e.g. spot checking). In previous peer prediction literature, problem (+) is also not fully solved in many important settings, even when we assume people are homogeneous (have the same expertise, ability ...). Avoiding collusion is more difficult compared with motivating the minority. Few previous works[1] deal with collusion and their results are typically proved by clever algebraic computations, sometimes lack a deeper intuition, and fail to extend to important settings.

After answering problem (+), an advanced central problem—task (++)—remains to be solved:

(++) **how to incentivize effort and identify expertise without verification?**

Previous peer prediction literature does not consider settings where

1. *Agents have different levels of expertise* or

2. *A lack of effort can systemically bias agents' reports.*

The following two tasks exemplify settings 1) and 2) respectively:

Example 1. Which state (from a list all 50 states) in the United States of America is closest to Africa? (Single-task)

Example 2. Peer grading several essays by providing a grade from the set $\{1, 2, 3, 4, 5\}$. (Multi-tasks)

[1]Prelec [51], Dasgupta and Ghosh [18], and Kamble et al. [31] deal with collusion and an independent work with this thesis, Shnayder et al. [63], also propose a mechanism that avoids collusion.

In the first example, an agent can guess randomly (no effort), look up the correct answer (full effort), or guess at the correct answer (partial effort). Most people will guess Florida, even though experts will know the correct answer is Maine. Thus differing levels of expertise yield different answers. In the second example, a student can, instead of carefully grading (full effort) or assigning a random grade (zero effort), quickly check the name of the top of the paper and spot check the grammar (partial effort). Thus partial effort can systematically bias agents: consider an essay from a top student in impeccable pose, but which contains large conceptual errors. Here partial effort can give some information about the correct answer, but also enable agents to "coordinate" on an incorrect answer.

Gao, Wright, and Leyton-Brown [22] show that the effects of the settings 1) and 2) are devastating to previous peer-prediction mechanisms, which generally fail in motivating the agents to invest effort for "expensive signals" when "cheap signals" (that ensure agreement and may even be correlated with the sought signal) exist. The main (very high-level) idea behind previous peer-prediction mechanisms can be understood as a "clever majority vote"—every agent is paid according to a specific similarity between her and her peer. Thus, they point out that in the peer-grading example, coordinating on just checking the grammar can guarantee good agreement with other agents, but with substantially reduced effort.

In fact, Gao et al point out that things are likely even worse than this. If the cheap signals correlate more than the expensive signals, then the peer-prediction techniques incentivize agents to not report the true answer, but instead focus on cheap signals! For example, in the essay grading above, it is likely that assessments of grammatical correctness will agree more than assessments of overall essay quality. Because of this, peer-prediction mechanisms will pay agents more overall for lower-quality information. In Example 1, even if agents know the answer is Maine, they may report Florida, expecting that most others will do likewise.

Such behavior undermines the goal of applying crowd-sourcing to increasingly complex tasks, and, in fact, undercuts any application of crowd-sourcing to perform any task where the answers are not "common knowledge." The field must overcome this key challenge of rewarding rather than suppressing expertise in order to begin the project of expanding crowd-sourcing beyond simple labeling tasks.

This thesis proposes a simple yet powerful idea to solve problem (+) and problem (++) that can be applied to various settings—rewarding every agent based on the *amount* and *value* of the information she provides. It remains to design proper information measurements to *quantify* information and *evaluate* information without verification. Thus, this thesis distills the essence of eliciting information to the design of proper information measurements.

Information theory is not typically used in the information elicitation literature, a key *novelty* of this thesis is showing how the insights of information theory illuminate the work and challenges in the information elicitation field.

1.1.1 Quantifying information

In a line of work [34] on designing mechanisms to elicit truthful, but unverifiable information, this thesis noticed that many current mechanisms can be understood in terms of information theory, specifically mutual information. This observation provides a unified framework—the *Mutual Information Paradigm*—for the field, simplifies several existing and foundational results, and provides several novel mechanisms in a variety of settings [38, 35, 34]. These mechanisms overcome a serious flaw of many previous information elicitation mechanisms: agents can obtain high reward by reporting meaningless information (e.g. everyone reports the a priori most likely answer).

In the mutual information paradigm, each agent i is paid the mutual information between her information and her peers' information—

$$MI(\text{her information}; \text{her peers' information}).$$

If we pick "correct" mutual information measures, no agent can obtain strict benefit by lying since intuitively the amount of information each agent has will not increase no matter what kind of strategy she applies to her information. That is, the mutual information measurement should be "information-monotone". We found two families of "(weakly) information-monotone" mutual information measures—f-mutual information and Bregman mutual information—both of which generalize the Shannon mutual information[2].

By assuming that agents are expected utility maximizers, it is sufficient to construct an unbiased estimator of the information measures. Unlike calculating the information measure, obtaining an unbiased estimator of the information measure only requires a small number of samples in many situations.

Section 1.1.3 will give an overview of the applications of the mutual information paradigm in a variety of important settings and the techniques used to construct an unbiased estimator of the information measures.

1.1.2 Evaluating information

Previous peer prediction mechanisms treat all information (cheap/expensive) equally, and thus agents lack an incentive to invest effort to obtain expensive signals. Moreover, even when an expert can easily obtain the expensive signal, previous mechanisms discourage her from providing it when she believes the non-experts will disagree.

A successful mechanism must break the symmetry between weak and expensive signals and between expert and non-expert signals. We propose the following natural assumption which will allow a mechanism to break this symmetry.

[2]Bregman mutual information is strictly weaker than f-mutual information since it only satisfies information-monotonicity in one of its two coordinatess and is asymmetric.

Assumption 3. *Agents with high effort or expertise, know the beliefs of agents with less effort or less expertise.*

We can see that this assumption is very natural in Example 1 and Example 2. Agents who look up or know the answer in Example 1 also know most people will answer "Florida." Agents that carefully grade an essay can also approximate the score of an agent who spends very little effort. We will define a hierarchical information structure to naturally capture Assumption 3.

Our mechanisms solicit not only agents' own opinions but also their predictions for the opinions of the other agents who have less information. This differs with the previous peer prediction mechanisms which ask agents to provide their predictions for *all* other agents' opinions. For example, in the peer-grading example, we might ask agents to report their own evaluation, and to *optionally* report one or more low-effort / low expertise evaluations (e.g. scores based on grammar, thesis statement, student name, naive reading, etc).

The information theoretic techniques and Assumption 3 lead to a mechanism design framework from which we construct several mechanisms for a variety of settings, such that expensive signals are incentivized and identified [37].

1.1.3 Overview of results and techniques

We present our results in designing *multi-choice* peer prediction mechanisms here. In these mechanisms, the elicited information is a discrete signal from a finite set. We will introduce our results in the context where the elicited information is a forecast in the information aggregation section since we will show essentially the forecast elicitation can be seen as an information aggregation problem. Important solution desiderata in peer prediction literature are:

(Strictly) truthful: truth-telling is a (strict) Bayesian Nash equilibrium. A strictly truthful mechanism motivates minority since for each agent, if she believes

everyone else tells the truth, she should tell the truth even she is minority.

(Symmetric) focal: the truth-telling equilibrium is paid more than other (Symmetric) equilibria in expectation. A symmetric focal mechanism avoids "all agree collusion".

Dominantly truthful: truth-telling maximizes the expected payment regardless of the other agents' strategies.

Before presenting the results, let's introduce a series of important settings.

Homogeneous/Heterogeneous In the homogeneous setting, we assume the prior over the signals agents will receive is symmetric in the sense agents have the same expertise. We do not have this assumption in heterogeneous setting.

Single/Multi-task In the *single-task* setting, each agent is assigned a single task (e.g. have you ever texted while driving before?). Miller, Resnick, and Zeckhauser [45] and Prelec [51] are two seminal works in this setting. Another is the *multi-task* setting in which each agent is assigned a batch of *apriori similar* tasks (e.g., peer grading, or is there a cat in this picture?). Dasgupta and Ghosh [18] is the foundational work in this setting.

Known prior/Detail free *Detail free* mechanisms require no knowledge of the prior over the signals agents will receive (e.g. with probability 0.6, 70% agents will receive "yes", with probability 0.4, 70% agents will receive "no") while *known prior* mechanisms are the opposite.

Minimal/Non-minimal *Minimal* mechanisms only require agents to report their information rather than forecasts for other agents' reports (e.g. is there a cat in this picture?) while *non-minimal* mechanisms requires the agents to report both (e.g. have you texted while driving before and what percentage of your peers have texted while driving before?).

Small/Medium/Large group *Large group* mechanisms require the number of participants to be large or even infinite. *Small group* mechanisms can be applied to the situation where the number of participants is greater than a small constant (e.g 3,6). In the *medium group* mechanisms, the number of participants are required to be greater than an integer that depends on the agents' prior $N(\text{Prior})$.

In addition to the novel information theoretic mechanism design frameworks, we also propose *disagreement mechanism* (Section 4.3) in the single-task, homogeneous setting, *f-mutual information mechanism* (Section 3.3) in the multi-task, homogeneous setting. The above mechanisms are all *detail free*. Figure 1.3, 1.4 show the comparison between these mechanisms and previous literature in homogeneous setting. In the heterogeneous setting, by assuming Assumption 3, we apply HMIP to create the following mechanisms:

Multi-HMIM: (Section 5.2.2) which works in the multiple task setting even for a small number of tasks but requires the mechanism to know the hierarchical information structure.

Learning based Multi-HMIM: (Section 5.2.3) which works in the multiple task setting even when the mechanism does not know the hierarchical information structure; however requires a large number of tasks.

Single-HMIM: (Section 5.3.2) which works in the single-task setting.

All of the above mechanisms work for small populations. Prelec, Seung, and McCoy [52] and Agarwal et al. [2] design mechanisms for settings with heterogeneous participants. Prelec, Seung, and McCoy [52] only consider the single-task setting and make a different assumption on the expertise. The mechanism in Prelec, Seung, and McCoy [52] requires an infinite number of participants. The mechanism in Agarwal

et al. [2] does not assume the hierarchy of the information and cannot be applied to identify and elicit expertise.

To apply MIP and HMIP in designing these mechanisms, we need to construct an unbiased estimator of the information measure using agents' reports. In the homogeneous setting, we pay each agent the unbiased estimator such that in expectation, each agent is paid based on the amount of the information. In the heterogeneous setting, we first evaluate the value of the information based on Assumption 3 and then use the same method as in the homogeneous setting to quantify the information using agents' reports. In the end, we pay each agent based on both the value and the amount of the information.

To construct an unbiased estimator of the information measure using agents' reports, different settings have different techniques. In the single-task setting, we use a non-minimal mechanism to ask agents their posterior (e.g. what percentage of your peers have texted while driving before?) and construct the estimator using both the first order information (e.g Y/N) and the second order information (e.g. 80% Yes). In the multi-task setting, either we ask a large number of questions to estimate the prior and use the prior to calculate the information measure, or we ask a small number questions but require the knowledge of information structure.

To give a flavor of the techniques used in constructing the estimator, we use a special case of the f-mutual information mechanism as an example. This special case is the TVD-mutual information mechanism which is also independently proposed by Shnayder et al. [63] (Section 3.5).

We assume that there are two agents: Alice and Bob. They are both asked to grade the same three essays. Their payment is

Average agreements for the same essay − Average agreements for different tasks

Figure 1.2: TVD-mutual information mechanism

When both Alice and Bob's answers are $(0, 1, 1)$, their average agreements for the same task are 1 and their average agreements for different tasks are $2/6 = 1/3$. Thus, they will be paid $2/3$.

By assuming Alice's answer is positively correlated with Bob's answer, the above payment is an unbiased estimator of the TVD-mutual information between Alice's answer and Bob's answer (Section 3.5). The major difference between this mutual information style payment and the naive "pay for agreements" is that this payment also punishes the agreements for different tasks. When both Alice and Bob's answers are $(1, 1, 1)$, although they have the maximal average agreements for the same essay, they also have maximal average agreements for the different essays. In this case, both Alice and Bob are paid nothing. Later we will show an extension of this idea in forecast elicitation where agents' reports are forecasts (Figure 1.7).

	Truthful	Focal	Small Population
Bayesian Truth Serum [Prelec 2004]	☺	☺	
Multi-Signal SM [PW 2014]	☺		☺
Multi-Valued RBTS+ [RF 2014]	☺		☺
Disagreement Mechanism	☺	☺	☺

Figure 1.3: Multi-choice single-task mechanisms comparison in homogeneous setting

	Truthful	Focal	Dominant truthful	Non-binary choice	Small number of tasks
Correlation mechanism [DG 2013]	☺	☺			☺
CA mechanism [SAFP 2016 *]	☺	☺		☺	☺
f-mutual information mechanism	☺	☺	☺	☺	

*Independent work

Figure 1.4: Multi-choice multi-task mechanisms comparison in homogeneous setting

1.2 Aggregating information without verification

Co-training/multiview learning is a problem that asks to aggregate two views of data into a prediction for the latent label, and was first proposed by Blum and Mitchell [9]. Although co-training is an important learning problem, it lacks a unified and rigorous approach to the general setting. The current thesis will make an innovative connection between the co-training problem and a peer prediction style mechanism design problem: forecast elicitation without verification, and develop a unified theory for both of them via the same information theoretic approach.

1.2.1 Aggregating information=Eliciting information

We use "forecasting whether a startup company will succeed" as our running example. We have two possible sources of information for each startup: the features X_A (e.g. products, business idea, target customer) of the startup; and the survey feedback X_B, collected from the crowd (e.g. a survey of amateur investors). Sometimes we have access to both the sources, and sometimes we have access to only one of the sources. We want to learn how to forecast the result Y (succeed/fail) of a startup company, using both or one of the sources.

We are given a set of candidates predictors $\{P_A\}$ (e.g. a set of hypotheses) such that each candidate predictor P_A maps the features X_A to a forecast for the result Y of the startup (e.g. succeed with 73% probability, fail with 27% probability). We are also given a set of candidates predictors $\{P_B\}$ (e.g. a set of aggregation algorithms like majority vote/weighted average) such that each candidate predictor P_B maps the survey feedback X_B to a forecast for the result Y. Our goal is to evaluate the performance of a specific pair P_A, P_B. The learning problem, learning how to forecast, can be reduced to this goal since if we know how to evaluate the two candidates P_A, P_B's performance, we can select the two candidates P_A^*, P_B^* which have the highest performance and use them to forecast.

Given a batch of past startup data each with the features X_A, the crowdsourced feedback X_B, and the result Y, we can evaluate the performance of the predictors through many existing measurements (e.g. proper scoring rules, loss functions). This evaluation method is related to the supervised learning setting. However, there may be only very few data points about the startups with results Y.[3] When we only use a few labeled data points to train the predictor, the predictor will likely over-fit. Thus, we can boldly ask:

(*Learning) *Can we evaluate the performance of the candidate predictors, as well as learn how to forecast the ground truth Y, without access to any data labeled with Y?* (see Figure 1.5) It is impossible to solve this problem without making an additional assumption on the relationship between X_A, X_B and Y. However, it turns out we can solve this problem with a natural assumption; conditioning on Y, X_A and X_B are independent. With this assumption, a naive approach is to learn the joint distribution of X_A and X_B using the past data, and then solve the relationship between Y and X_A, X_B by some calculations, using the lemma that X_A and X_B are independent conditioning on Y. However, this naive approach will not work if either X_A or X_B

[3]For example, if we focus on cryptographic or self-driving currencies, there are very few startups labeled with results.

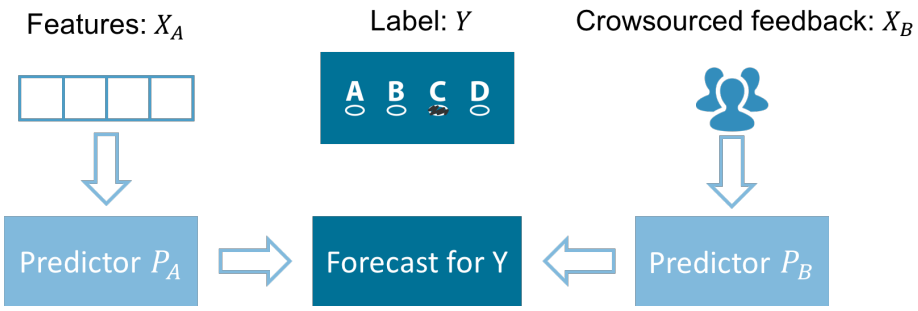

Figure 1.5: Problem (*): Finding the common ground truth

has very high dimension. We will solve this problem using learning methods. Before we go further on the learning problem, let's consider a corresponding mechanism design problem. In the scenario where the forecasts are provided by human beings, we want to ask a mechanism design problem:

(**Mechanism design) *Can we design proper* instant *reward schemes to incentivize high quality forecast for* Y *without instant access to* Y*?* (see Figure 1.6)

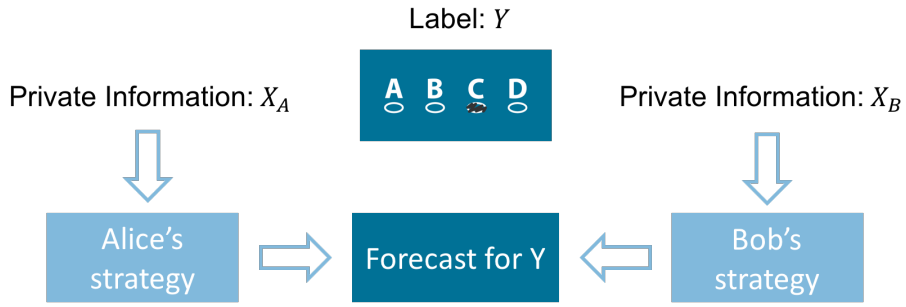

Figure 1.6: Problem (**): Forecast elicitation

People will obtain instant payments from *instant* reward schemes. If we do not require the reward schemes to be instant, proper scoring rules will work by rewarding people in the future after Y is revealed. It turns out the above learning problem (*) and mechanism design problem (**) are essentially the same, since there is a natural correspondence between an evaluation of their performance and their rewards. A first try would be rewarding the predictors according to their "agreement", since high quality predictors should have a lot of agreement with each other. However, if we

train the predictors based on this criterion, then the output of the training process will be two meaningless constant predictors which perfectly agree with each other (e.g. always forecast 100% success). We call this problem the "naive agreement" issue.

Note that the mechanism design problem (**) is closely related to the peer prediction literature, incentivizing high quality information reports without verification. It is natural to leverage the techniques and insights from peer prediction to address problems (*) and (**). In fact, the peer prediction literature provides an information theoretic idea to address the "naive agreement" issue, that is, replacing "agreement" by mutual information. In the thesis, we will show that with a natural assumption, conditioning on Y, X_A, and X_B are independent, we can address problem (*) and (**) simultaneously via rewarding the predictors the mutual information between them and using the predictors' reward as the evaluation of their performance.

1.2.2 Overview of results and techniques

Our contribution We build a natural connection between mechanism design and machine learning by simultaneously addressing a learning problem and a mechanism design problem in the context where ground truth is unknown, via the same information theoretic approach.

Learning We focus on the co-training problem [9]: learning how to forecast Y using two sources of information X_A and X_B, without access to any data labeled with ground truth Y (Section 6.3). By making a typical assumption in the co-training literature, conditioning on Y, X_A and X_B are independent, we reduce the learning problem to an optimization problem $\max_{P_A,P_B} MIG^f(P_A, P_B)$ such that solving the learning problem is equivalent to picking the P_A^*, P_B^* that maximize $MIG^f(P_A, P_B)$, i.e., the f-mutual information gain between P_A and P_B (Section 6.4). Formally, we define *the Bayesian posterior predictor* as the

predictor that maps any input information $X = x$ to its Bayesian posterior forecast for $Y = y$, i.e., $Pr(Y = y|X = x)$. Then when both P_A, P_B are Bayesian posterior predictors, $MIG^f(P_A, P_B)$ is maximized and the maximal value is the f-mutual information between X_A and X_B. With an additional mild restriction on the prior, $MIG^f(P_A, P_B)$ is maximized if and only if both P_A, P_B are permuted versions of the Bayesian posterior predictor.

We also design another family of optimization goals, *PS-gain*[4], based on the family of proper scoring rules (Section 6.6). We can also reduce the learning problem to the PS-gain optimization problem. We will show a special case of the PS-gain, picking PS as the logarithmic scoring rule LSR, corresponds to the maximum likelihood estimator method. The range of applications of PS-gain is more limited when compared with the range of applications of f-mutual information gain, since the application of PS-gain requires either one of the information sources to be low dimensional or that we have a simple generative model for the distribution over one of the information sources and ground truth labels, while f-mutual information gain does not have these restrictions.

As is typical in related literature, we do not investigate the computation complexity or data requirement of the learning problem.

To the best of our knowledge, this is the first optimization goal in the co-training literature that guarantees that the maximizer corresponds to the Bayesian posterior predictor, without any additional assumption. Thus, our method optimally aggregates the two sources of information.

Mechanism design Consider the scenario where we elicit forecasts for ground truth Y from agents and pay agents immediately. Without access to Y, given the prior on the distribution of Y, i.e., $Pr[Y]$, [5] by assuming agents' private information

[4] PS is a proper scoring rule.

[5] This is not a very strong assumption since we do not need the knowledge of the joint distribution

are independent conditioning on Y, in the single-task setting (there is only a single forecasting task), we design a *strictly truthful* mechanism, the *common ground mechanism*, where truth-telling is a strict equilibrium (Section 6.5.2); in the multi-task (there are at least two a priori similar forecasting tasks) setting, we design a family of *focal* mechanisms, the *multi-task common ground mechanism* $MCG(f)s$, where the truth-telling equilibrium pays better than any other strategy profile and *strictly* higher than any non-permutation strategy profile (Section 6.5.1).

Technical contribution Our main technical ingredient is a novel performance measurement, the *f-mutual information gain*, which is an unbiased estimator of the f-mutual information. To give a flavor of this measurement, we give an informal presentation here: both P_A and P_B are assigned a batch of forecasting tasks, the f-mutual information gain between P_A and P_B is

The agreements between P_A's forecast and P_B's forecast for the same task

$-$ $f^\star$(The agreements between P_A's forecast and P_B's forecast for different tasks)

where $f^\star$ is the conjugate of the convex function f. With this measurement, two agreeing constant predictors have small gain since their outputs have large agreements for both the same task and different tasks. The formal definition will be introduced in Section 6.4.1 and the agreement measure is introduced in Definition 115.

The f-mutual information gain is conceptually similar to the correlation payment scheme (Figure 1.2) proposed by Dasgupta and Ghosh [18] (in the binary choice setting), and Shnayder et al. [63] (in to multiple choice setting), which pays agents "the agreements for the same task *minus* the agreements for the distinct task". In Dasgupta and Ghosh [18] and Shnayder et al. [63], the payment scheme is designed

over the event and agents' private information.

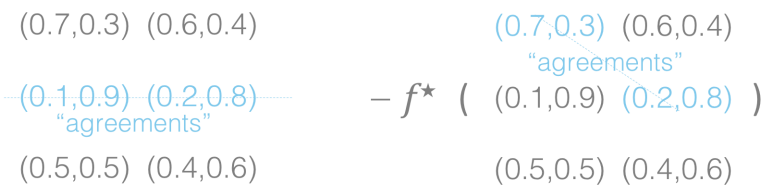

Figure 1.7: f-mutual information gain

P_A and P_B are assigned three forecasting tasks. P_A's outputs are $(0.7, 0.3), (0.1, 0.9), (0.5, 0.5)$ and P_B's outputs are $(0.6, 0.4), (0.2, 0.8), (0.4, 0.6)$. To calculate the f-mutual information gain between them, we pick a task (e.g. Task no.2) uniformly at random and calculate the agreements a_s between P_A and P_B's forecasts for this task; we also pick a pair of distinct task (i, j) uniformly at random (e.g. (Task no.1, Task no.2)) and calculate the agreements a_d between P_A's forecast for task i and P_B's forecast for this task j. The f-mutual information gain is set as $a_s - f^\star(a_d)$. We can also calculate a more concentrated version of the f-mutual information gain by replacing a_s and $f^\star(a_d)$ by their empirical expectations. The formal definition (Section 6.4.1) uses the concentrated version.

for discrete signals and the measure of agreements is a simple indicator function. This thesis also show that this correlation payment is related to a special f-mutual information, TVD-mutual information (Section 3.5). Thus, the f-mutual information gain can be seen as an extension of the correlation payment scheme and work for forecasts report.

1.3 Roadmap

This thesis will start by introducing the general information theoretic mechanism design frameworks in Chapter II that quantify (Mutual information paradigm) and evaluate information (Hierarchical mutual information paradigm) without verification and price information such that agents will be incentivized to invest effort and provide high quality information.

Both Chapter III and Chapter IV assume agents and information are homogeneous. Chapter III applies the mutual information paradigm into multi-task setting and propose two families of novel mechanisms: the f-mutual information mechanism and the Bregman mutual information mechanism, and map the seminal work [18] in

multi-task peer prediction literature into the mutual information paradigm. Chapter IV considers the single-task setting and proposes a novel mechanism—the Disagreement mechanism. The Disagreement mechanism is the first detail free, strictly truthful, focal mechanism in the single-task setting that works for a small number of participants. Although the design of the Disagreement mechanism does not directly use the mutual information mechanism, it also employs the information theory tools and uses the information monotonicity property. Chapter IV also maps the first detail free and truthful mechanism—Bayesian truth serum [51]—into the mutual information paradigm such that the results can be constructed easily.

Chapter V considers the setting where agents have different expertise and applies the hierarchical mutual information paradigm in both single-task and multi-task setting to propose several novel mechanisms that can identify expertise and incentivize low cost agents to invest high level effort and provide an honest report.

Chapter VI considers an important unsupervised learning problem—co-training [9] which can be also seen as an information aggregation problem. Chapter VI reduces this learning problem to a peer-prediction-style mechanism design problem—forecast elicitation without verification and addresses them simultaneously using the same information theoretic approach—the mutual information paradigm.

Chapter VII concludes this thesis and proposes several potential future works.

CHAPTER II

An Information Theoretic Framework

2.1 Preliminaries

2.1.1 Transition probability

We define a $m \times m'$ transition matrix $M \in \mathbb{R}^{m \times m'}$ as a matrix such that for any $i, j \in [m] \times [m']$, $M_{i,j} \geq 0$ and $\sum_j M_{i,j} = 1$. We define a *permutation transition matrix* π as a $m \times m$ permutation matrix.

Given a random variable X with m possible outcomes, by abusing notation a little bit, a $m \times m'$ transition matrix M defines a **transition probability** M that transforms X to $M(X)$ such that $X' := M(X)$ is a new random variable that has m' possible outcomes where $\Pr[X' = j | X = i] = M_{i,j}$.

If the distribution of X is represented by an $m \times 1$ column vector $\mathbf{p}$, then the distribution over $M(X)$ is $M^T \mathbf{p}$ where M^T is the transpose of M.

2.1.2 f-divergence

f-divergence $D_f : \Delta_\Sigma \times \Delta_\Sigma \to \mathbb{R}$ is a non-symmetric measure of the difference between distribution $\mathbf{p} \in \Delta_\Sigma$ and distribution $\mathbf{q} \in \Delta_\Sigma$ and is defined to be

$$D_f(\mathbf{p}, \mathbf{q}) = \sum_{\sigma \in \Sigma} \mathbf{p}(\sigma) f\left(\frac{\mathbf{q}(\sigma)}{\mathbf{p}(\sigma)}\right)$$

where $f(\cdot)$ is a convex function and $f(1) = 0$. Now we introduce the properties of f-divergence:

Fact 4 (Non-negativity [16])**.** For any $\mathbf{p}, \mathbf{q}$, $D_f(\mathbf{p}, \mathbf{q}) \geq 0$ and $D_f(\mathbf{p}, \mathbf{q}) = 0$ if and only if $\mathbf{p} = \mathbf{q}$.

Fact 5 (Joint Convexity [16])**.** For any $0 \leq \lambda \leq 1$, for any $\mathbf{p_1}, \mathbf{p_2}, \mathbf{q}_1, \mathbf{q}_2 \in \Delta_\Sigma$,

$$D_f(\lambda \mathbf{p_1} + (1-\lambda)\mathbf{p_2}, \lambda \mathbf{q_1} + (1-\lambda)\mathbf{q_2}) \leq \lambda D_f(\mathbf{p_1}, \mathbf{q_1}) + (1-\lambda) D_f(\mathbf{p_2}, \mathbf{q_2}).$$

Fact 6 (Information Monotonicity ([3, 40, 5]))**.** For any strictly convex function f, f-divergence $D_f(\mathbf{p}, \mathbf{q})$ satisfies information monotonicity so that for any transition matrix $\theta \in \mathbb{R}^{|\Sigma| \times |\Sigma|}$, $D_f(\mathbf{p}, \mathbf{q}) \geq D_f(\theta^T \mathbf{p}, \theta^T \mathbf{q})$.

Moreover, the inequality is strict if and only if there exists $\sigma, \sigma', \sigma''$ such that $\frac{\mathbf{p}(\sigma'')}{\mathbf{p}(\sigma')} \neq \frac{\mathbf{q}(\sigma'')}{\mathbf{q}(\sigma')}$ and $\theta_{\sigma', \sigma} \mathbf{p}(\sigma') > 0$, $\theta_{\sigma'', \sigma} \mathbf{p}(\sigma'') > 0$.

If the strictness condition does not satisfied, we can see $\theta^T \mathbf{p}$ and $\theta^T \mathbf{q}$ are $\mathbf{p}$ and $\mathbf{q}$'s sufficient statistic which means the transition θ does not lose any information, thus, the equality holds.

Proof. The proof follows from algebraic manipulation and one application of convexity.

$$D_f(\theta^T\mathbf{p}, \theta^T\mathbf{q}) = \sum_\sigma (\theta^T\mathbf{p})(\sigma) f\left(\frac{(\theta^T\mathbf{q})(\sigma)}{(\theta^T\mathbf{p})(\sigma)}\right) \tag{2.1}$$

$$= \sum_\sigma \theta^T_{\sigma,\cdot}\mathbf{p} f\left(\frac{\theta^T_{\sigma,\cdot}\mathbf{q}}{\theta^T_{\sigma,\cdot}\mathbf{p}}\right) \tag{2.2}$$

$$= \sum_\sigma \theta^T_{\sigma,\cdot}\mathbf{p} f\left(\frac{1}{\theta^T_{\sigma,\cdot}\mathbf{p}} \sum_{\sigma'} \theta^T_{\sigma,\sigma'}\mathbf{p}(\sigma')\frac{\mathbf{q}(\sigma')}{\mathbf{p}(\sigma')}\right) \tag{2.3}$$

$$\leq \sum_\sigma \theta^T_{\sigma,\cdot}\mathbf{p}\frac{1}{\theta^T_{\sigma,\cdot}\mathbf{p}} \sum_{\sigma'} \theta^T_{\sigma,\sigma'}\mathbf{p}(\sigma') f\left(\frac{\mathbf{q}(\sigma')}{\mathbf{p}(\sigma')}\right) \tag{2.4}$$

$$= \sum_\sigma \mathbf{p}(\sigma) f\left(\frac{\mathbf{q}(\sigma)}{\mathbf{p}(\sigma)}\right) = D_f(\mathbf{p}, \mathbf{q}) \tag{2.5}$$

The second equality holds since $(\theta^T\mathbf{p})(\sigma)$ is dot product of the σ^{th} row of θ^T and $\mathbf{p}$.

The third equality holds since $\sum_{\sigma'} \theta^T_{\sigma,\sigma'}\mathbf{p}(\sigma')\frac{\mathbf{q}(\sigma')}{\mathbf{p}(\sigma')} = \theta^T_{\sigma,\cdot}\mathbf{q}$.

The fourth inequality follows from the convexity of $f(\cdot)$.

The last equality holds since $\sum_\sigma \theta^T_{\sigma,\sigma'} = 1$.

We now examine under what conditions the inequality in Equation 2.4 is strict. Note that for any strictly convex function g, if $\forall u, \lambda_u > 0$, $g(\sum_u \lambda_u x_u) = \sum_u \lambda_u g(x_u)$ if and only if there exists x such that $\forall u, x_u = x$. By this property, the inequality is strict if and only if there exists $\sigma, \sigma', \sigma''$ such that $\frac{\mathbf{q}(\sigma')}{\mathbf{p}(\sigma')} \neq \frac{\mathbf{q}(\sigma'')}{\mathbf{p}(\sigma'')}$ and $\theta^T_{\sigma,\sigma'}\mathbf{p}(\sigma') > 0$, $\theta^T_{\sigma,\sigma''}\mathbf{p}(\sigma'') > 0$. □

Definition 7. Given two signals $\sigma', \sigma'' \in \Sigma$, we say two probability measures $\mathbf{p}, \mathbf{q}$ over Σ can **distinguish** $\sigma', \sigma'' \in \Sigma$ if $\mathbf{p}(\sigma') > 0$, $\mathbf{p}(\sigma'') > 0$ and $\frac{\mathbf{q}(\sigma')}{\mathbf{p}(\sigma')} \neq \frac{\mathbf{q}(\sigma'')}{\mathbf{p}(\sigma'')}$

Fact 6 directly implies

Corollary 8. *Given a transition matrix θ and two probability measures $\mathbf{p}, \mathbf{q}$ that can distinguish $\sigma', \sigma'' \in \Sigma$, if there exists $\sigma \in \Sigma$ such that $\theta(\sigma', \sigma), \theta(\sigma'', \sigma) > 0$, we have $D_f(\mathbf{p}, \mathbf{q}) > D_f(\theta^T\mathbf{p}, \theta^T\mathbf{q})$ when f is a strictly convex function.*

Now we introduce two f-divergences in common use: KL divergence, and Total variation Distance.

Example 9 (KL divergence)**.** Choosing $-\log(x)$ as the convex function $f(x)$, f-divergence becomes KL divergence $D_{KL}(\mathbf{p}, \mathbf{q}) = \sum_\sigma \mathbf{p}(\sigma) \log \frac{\mathbf{p}(\sigma)}{\mathbf{q}(\sigma)}$

Example 10 (Total Variation Distance)**.** Choosing $|x - 1|$ as the convex function $f(x)$, f-divergence becomes Total Variation Distance $D_{tvd}(\mathbf{p}, \mathbf{q}) = \sum_\sigma |\mathbf{p}(\sigma) - \mathbf{q}(\sigma)|$

2.1.3 Proper scoring rules

Informally, a scoring rule measures the accuracy of the forecasts. Formally, a scoring rule [66, 24] $PS : \Sigma \times \Delta_\Sigma \to \mathbb{R}$ takes in a signal $x \in \Sigma$ and a distribution over signals $\delta_\Sigma \in \Delta_\Sigma$ and outputs a real number. A scoring rule is *proper* if, whenever the first input is drawn from a distribution δ_Σ, then δ_Σ will maximize the expectation of PS over all possible inputs in Δ_Σ to the second coordinate. A scoring rule is called *strictly proper* if this maximum is unique. We will assume throughout that the scoring rules we use are strictly proper. Slightly abusing notation, we can extend a scoring rule to be $PS : \Delta_\Sigma \times \Delta_\Sigma \to \mathbb{R}$ by simply taking $PS(\delta_\Sigma, \delta'_\Sigma) = \mathbb{E}_{x \leftarrow \delta_\Sigma}(x, \delta'_\Sigma)$. We note that this means that any proper scoring rule is linear in the first term.

Example 11 (Log Scoring Rule [66, 24])**.** Fix an outcome space Σ for a signal x. Let $\mathbf{q} \in \Delta_\Sigma$ be a reported distribution. The Logarithmic Scoring Rule maps a signal and reported distribution to a payoff as follows:

$$L(x, \mathbf{q}) = \log(\mathbf{q}(x)).$$

Let the signal x be drawn from some random process with distribution $\mathbf{p} \in \Delta_\Sigma$. Then the expected payoff of the Logarithmic Scoring Rule

$$\mathbb{E}_{x \leftarrow \mathbf{p}}[L(x, \mathbf{q})] = \sum_x \mathbf{p}(x) \log \mathbf{q}(x) = L(\mathbf{p}, \mathbf{q})$$

This value will be maximized if and only if $\mathbf{q} = \mathbf{p}$.

Intuitively, more information should imply a more accurate prediction. This intuition is valid when the accuracy is measured by a proper scoring rule. When predicting a random variable Y, assuming that all agents have a common prior, the agent who has more information will have higher prediction score when the prediction score is measured by a proper scoring rule. We denote the prediction of Y conditioning on X as $\Pr[\boldsymbol{Y}|X] := (\Pr[Y = 1|X], \Pr[Y = 2|X], ..., \Pr[Y = |\Sigma||X]) \in \Delta_\Sigma$.

Fact 12 (Information monotonicity of proper scoring rules)**.** Given any strictly proper scoring rule PS,

$$\mathbb{E}_{X,Y,Z} PS(Y, \Pr[\boldsymbol{Y}|X, Z]) \geq \mathbb{E}_{X,Y} PS(Y, \Pr[\boldsymbol{Y}|X]).$$

The equality holds if and only if $\Pr[\boldsymbol{Y}|X = x, Z = z] = \Pr[\boldsymbol{Y}|X = x]$ for all (x, z) where $\Pr[X = x, Z = z] > 0$.

We defer the proof to the appendix.

2.2 (Weakly) Information-monotone information measures

2.2.1 f-mutual information

Given two random variables X, Y, let $\mathbf{U}_{X,Y}$ and $\mathbf{V}_{X,Y}$ be two probability measures where $\mathbf{U}_{X,Y}$ is the joint distribution of (X, Y) and $\mathbf{V}$ is the product of the marginal distributions of X and Y. Formally, for every pair of (x, y),

$$\mathbf{U}_{X,Y}(X = x, Y = y) = \Pr[X = x, Y = y] \qquad \mathbf{V}_{X,Y}(X = x, Y = y) = \Pr[X = x] \Pr[Y = y].$$

If $\mathbf{U}_{X,Y}$ is very different with $\mathbf{V}_{X,Y}$, the mutual information between X and Y should be high since knowing X changes the belief for Y a lot. If $\mathbf{U}_{X,Y}$ equals to

$\mathbf{V}_{X,Y}$, the mutual information between X and Y should be zero since X is independent with Y. Intuitively, the "distance" between $\mathbf{U}_{X,Y}$ and $\mathbf{V}_{X,Y}$ represents the mutual information between them.

Definition 13 (f-mutual information)**.** The f-mutual information between X and Y is defined as

$$MI^f(X;Y) = D_f(\mathbf{U}_{X,Y}, \mathbf{V}_{X,Y})$$

where D_f is f-divergence.

Example 14 (KL divergence and $I(\cdot;\cdot)$)**.** Choosing f-divergence as KL divergence, f-mutual information becomes the Shannon (conditional) mutual information [15]

$$I(X;Y) := MI^{KL}(X;Y) = \sum_{x,y} \Pr[X=x, Y=y] \log \frac{\Pr[X=x, Y=y]}{\Pr[X=x]\Pr[Y=y]}$$

$$\begin{aligned} I(X;Y|Z) :=& MI^{KL}(X;Y|Z) \\ =& \sum_{x,y} \Pr[X=x, Y=y, Z=z] \log \frac{\Pr[X=x, Y=y|Z=z]}{\Pr[X=x|Z=z]\Pr[Y=y|Z=z]}. \end{aligned}$$

Example 15 (Total Variation Distance and $MI^{tvd}(\cdot;\cdot)$)**.** Choosing f-divergence as Total Variation Distance, f-mutual information becomes

$$MI^{tvd}(X;Y) := \sum_{x,y} |\Pr[X=x, Y=y] - \Pr[X=x]\Pr[Y=y]|.$$

For the strictness guarantee, we introduce the following definition:

Definition 16 (Fine-grained distribution)**.** $P \in \Delta_{\Sigma_X \times \Sigma_Y}$ is a fine-grained joint distribution over X and Y if for every two distinct pairs $(x,y), (x',y')$, $U_{X,Y}(X,Y) := P(X,Y)$ and $V_{X,Y}(X,Y) := P(X)P(Y)$ can **distinguish** (see Definition 7) (x,y) and (x',y').

Fact 17 (General data processing inequality)**.** When f is strictly convex, f-mutual information MI^f is information-monotone and strictly information-monotone with respect to all fine-grained joint distributions over X and Y.

Definition 18 (Fine-grained prior)**.** Given general setting (n, Σ), Q is fine-grained prior if for every pair i, j, $Q(\Psi_i, \Psi_j)$ is a fine-grained joint distribution over Ψ_i and Ψ_j.

Proof of Theorem 17. We will apply the information monotonicity of f-divergence to show the data processing inequality of f-mutual information. We first introduce several matrix operations to ease the presentation of the proof.

Definition 19 (vec operator [27])**.** The vec operator creates a column vector $\text{vec}(A)$ from a matrix A by stacking the column vectors of A.

Definition 20 (Kronecker Product [27])**.** The Kronecker product of two matrices $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times q}$ is defined as the $mp \times nq$ matrix $A \otimes B = \{A_{i,j}B\} = \begin{bmatrix} A_{11}B & \dots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \dots & A_{mn}B \end{bmatrix}$.

Fact 21 (vec operator and Kronecker Product [61])**.** For any matrices $A \in \mathbb{R}^{n_1\times n_2}$, $X \in \mathbb{R}^{n_2\times n_3}$, $B \in \mathbb{R}^{n_3\times n_4}$, $\text{vec}(AXB) = B^T \otimes A\text{vec}(X)$.

Let $X : \Omega \mapsto \Sigma_X$, $Y : \Omega \mapsto \Sigma_Y$ be two random variables. $U_{X,Y}$ and $V_{X,Y}$ can be seen as two $\Sigma_X \times \Sigma_Y$ matrices. Let M be a $|\Sigma_X| \times |\Sigma_X|$ transition matrix.

We define $\Sigma_{X,Y}$ as $\Sigma_X \times \Sigma_Y$.

Note that the vectorization of the matrix that represents the probability measure over X and Y will not change the probability measure. Thus,

$$D_f(U_{M(X),Y}, V_{M(X),Y}) = D_f(\text{vec}(U_{M(X),Y}), \text{vec}(V_{M(X),Y})).$$

We define I as a $|\Sigma_Y| \times |\Sigma_Y|$ identity matrix. For any transition matrix M, by simple calculations, we can see the Kronecker product between M and the identity matrix I is a transition matrix as well.

When Y is independent with $M(X)$ conditioning on X, for any probability measure $P \in \Delta_{\Sigma_X \times \Sigma_Y}$ on X and Y,

$$\begin{aligned} P(M(X) = x', Y = y) &= \sum_x P(M(X) = x' | X = x, Y = y) P(X = x, Y = y) \quad (2.6) \\ &= \sum_x P(M(X) = x' | X = x) P(X = x, Y = y) \\ &\qquad (Y \text{ is independent with } M(X) \text{ conditioning on } X) \end{aligned}$$

$$\begin{aligned} MI^f(M(X); Y) =& D_f(U_{M(X),Y}, V_{M(X),Y}) \quad (2.7) \\ =& D_f(\mathrm{vec}(U_{M(X),Y}), \mathrm{vec}(V_{M(X),Y})) \\ =& D_f(\mathrm{vec}(M^T U_{X,Y} I), \mathrm{vec}(M^T V_{(X),Y} I)) \\ & \qquad (\text{equation (2.6), replacing } P \text{ by } U_{X,Y} \text{ and } V_{X,Y}) \\ =& D_f(I^T \otimes M^T \mathrm{vec}(U_{X,Y}), I^T \otimes M^T \mathrm{vec}(U_{X,Y})) \quad (\text{Fact 21}) \\ \leq& D_f(\mathrm{vec}(U_{X,Y}), \mathrm{vec}(V_{X,Y})) \\ & \qquad (\text{information monotonicity of } f\text{-divergence}) \\ =& D_f(U_{X,Y}, V_{X,Y}) \\ =& MI^f(X; Y) \end{aligned}$$

Now we show the strictness guarantee. When M is a non-permutation matrix, $\Theta := (I^T \otimes M^T)^T = M \otimes I$ is a non-permutation matrix as well. Thus there must exist $(x, y), (x', y'), (x'', y'')$ such that both $\Theta((x, y), (x', y'))$ and $\Theta((x, y), (x'', y''))$ are strictly positive where $(x', y') \neq (x'', y'')$. According to the definition of fine-grained prior (see Definition 18), $U_{X,Y}$ and $V_{X,Y}$ can distinguish (x', y') and (x', y'). Then

Corollary 8 implies that the inequality in (2.7) is strict.

□

Fact 22 (Convexity of f-mutual information)**.** For any $0 \leq \lambda \leq 1$, for any random variables X_1, X_2, Y, let B_λ be an independent Bernoulli variable such that $B_\lambda = 1$ with probability λ and 0 with probability $1 - \lambda$. Let X be a random variable such that if $B_\lambda = 1$, $X = X_1$, otherwise, $X = X_2$,

$$MI^f(X;Y) \leq \lambda MI^f(X_1;Y) + (1-\lambda)MI^f(X_2;Y).$$

Proof. Based on the definition of X,

$$U_{X,Y} = \lambda U_{X_1,Y} + (1-\lambda)U_{X_2,Y} \qquad V_{X,Y} = \lambda V_{X_1,Y} + (1-\lambda)V_{X_2,Y}.$$

Combining the joint convexity of D_f (Fact 5) and the fact that $MI^f(X;Y) = D_f(U_{X,Y}, V_{X,Y})$,

$$MI^f(X;Y) \leq \lambda MI^f(X_1;Y) + (1-\lambda)MI^f(X_2;Y).$$

□

2.2.2 Bregman mutual information

It is naturally to ask whether in addition to f-divergence, can we use another commonly used divergence—Bregman divergence D_{PS}—to define an information-monotone information measure. Since the general Bregman divergence may not satisfy information monotonicity, the answer is likely to be negative. However, surprisingly, by properly using the Bregman divergence, we can obtain a new family of information measures BMI^{PS} that satisfies almost all information-monotone properties of f-mutual information except the symmetry and one half of the data processing

inequality. Therefore, by plugging BMI^{PS} into the Mutual Information Paradigm, we may lose the focal property but can preserve the dominantly truthful property.

Bregman Divergence [10, 24] Bregman divergence $D_{PS} : \Delta_\Sigma \times \Delta_\Sigma \to \mathbb{R}$ is a non-symmetric measure of the difference between distribution $\mathbf{p} \in \Delta_\Sigma$ and distribution $\mathbf{q} \in \Delta_\Sigma$ and is defined to be

$$D_{PS}(\mathbf{p}, \mathbf{q}) = PS(\mathbf{p}, \mathbf{p}) - PS(\mathbf{p}, \mathbf{q})$$

where PS is a proper scoring rule (see the definition of PS in Section 2.1.3).

Inspired by the f-mutual information, we can first try $D_{PS}(\mathbf{U}_{X,Y}, \mathbf{V}_{X,Y})$ to define the Bregman mutual information. However, since the Bregman divergence may not satisfy the information monotonicity, this idea does not work. Intuitively, more information implies a more accurate prediction. Inspired by this intuition, we define Bregman mutual information between X and Y as an accuracy gain—*the accuracy of the posterior* $\Pr[\mathbf{Y}|X]$ *minus the accuracy of the prior* $\Pr[\mathbf{Y}]$. With this definition, if X changes the belief for Y a lot, then the Bregman mutual information between them is high; if X is independent with Y, $\Pr[\mathbf{Y}|X] = \Pr[\mathbf{Y}]$, then the Bregman mutual information between them is zero.

We define $\mathbf{U}_{Y|X=x}$ and $\mathbf{U}_Y$ as two probability distribution over Y such that

$$\mathbf{U}_{Y|X=x}(Y = y) = \Pr[Y = y|X = x] \qquad \mathbf{U}_Y(Y = y) = \Pr[Y = y].$$

Definition 23 (Bregman mutual information)**.** The Bregman mutual information between X and Y is defined as

$$BMI^{PS}(X;Y) = \mathbb{E}_X D_{PS}(\mathbf{U}_{Y|X}, \mathbf{U}_Y) = \mathbb{E}_X PS(\Pr[\mathbf{Y}|X], \Pr[\mathbf{Y}|X]) - PS(\Pr[\mathbf{Y}|X], \Pr[\mathbf{Y}]).$$

Bridging log scoring rule and Shannon mutual information Inspired by the definition of Bregman mutual information, we will show a novel connection between log scoring rule and Shannon information theory concepts—*the log scoring rule can be used to construct an unbiased estimator of (conditional) Shannon mutual information.* A powerful application of this connection is the information theoretic reconstruction of Prelec [51] (Section 4.4.0.2).

The definition of Bregman mutual information says that the accuracy gain measured by a proper scoring rule PS equals the information gain measured by the (conditional) Bregman mutual information BMI^{PS}. The following theorem (Theorem 24) shows that we can bridge the log scoring rule and Shannon mutual information by showing the accuracy gain measured by log scoring rule equals the information gain measured by (conditional) Shannon mutual information. Therefore, like f-mutual information, Bregman mutual information also generalizes Shannon mutual information (Corollary 25).

Theorem 24 (expected accuracy gain = information gain)**.** *For random variables X, Y, Z, when predicting Y, the logarithm score of prediction $\Pr[\boldsymbol{Y}|Z, X]$ minus the logarithm score of prediction $\Pr[\boldsymbol{Y}|Z]$*

$$\mathbb{E}_{X,Y,Z} L(Y, \Pr[\boldsymbol{Y}|Z, X]) - L(Y, \Pr[\boldsymbol{Y}|Z]) = I(X; Y|Z)$$

where $L : \Sigma \times \Delta_\Sigma \mapsto \mathbb{R}$ is the log scoring rule and $I(X; Y|Z)$ is the Shannon mutual information between X and Y conditioning on Z.

Proof.

$$
\begin{aligned}
&\mathbb{E}_{X,Y,Z} L(Y, \Pr[\boldsymbol{Y}|Z, X]) - L(Y, \Pr[\boldsymbol{Y}|Z]) \\
&= \sum_{x,y,z} \Pr[X = x, Y = y, Z = z] \log(\frac{\Pr[Y = y|Z = z, X = x]}{\Pr[Y = y|Z = z]}) \\
&= \sum_{x,y,z} \Pr[X = x, Y = y, Z = z] \log(\frac{\Pr[Y = y, X = x|Z = z]}{\Pr[Y = y|Z = z] \Pr[X = x|Z = z]}) \\
&= I(X; Y|Z)
\end{aligned}
$$

□

Recall that the conditional mutual information (Definition 29) is defined as

$$\sum_{z} \Pr[Z = z] MI(X; Y|Z = z).$$

Thus,

$$BMI^{PS}(X; Y|Z) = \mathbb{E}_{X,Z} PS(\Pr[\mathbf{Y}|X, Z], \Pr[\mathbf{Y}|X, Z]) - PS(\Pr[\mathbf{Y}|X, Z], \Pr[\mathbf{Y}|Z])$$

which is the accuracy of posterior $\Pr[\mathbf{Y}|X, Z]$ minus the accuracy of prior $\Pr[\mathbf{Y}|Z]$. Therefore, Fact 24 directly implies Corollary 25.

Corollary 25. *$BMI^{L(\cdot,\cdot)}(X; Y|Z) = I(X; Y|Z)$ where $BMI^{L(\cdot,\cdot)}$ is a Bregman mutual information that chooses Log scoring rule $L(\cdot, \cdot)$ as the proper scoring rule.*

Definition 26 (Quasi Information-monotone mutual information)**.** We say MI is quasi information-monotone if and only if it is always non-negative and satisfies the data processing inequality for the first entry.

A quasi information-monotone mutual information may not be symmetric. Thus, even if it satisfies the data processing inequality for the first entry, it may not satisfy

the data processing inequality for the second entry which means data processing methods operating on Y may increase $MI(X;Y)$.

Theorem 27. *The Bregman mutual information is quasi information-monotone.*

Intuitively, more information about X provides a more accurate prediction for random variable Y. That is, $\Pr[\mathbf{Y}|M(X)]$ is less accurate than $\Pr[\mathbf{Y}|X]$. We will show the property of the proper scoring rules directly implies the above intuition and then the quasi information-monotonicity of BMI^{PS} follows.

Proof. The definition of proper scoring rules implies the non-negativity of Bregman divergence as well as that of Bregman mutual information.

For any transition probability M that operates on X,

$$
\begin{aligned}
BMI^{PS}(M(X);Y) &= \mathbb{E}_{M(X)} PS(\Pr[\mathbf{Y}|M(X)], \Pr[\mathbf{Y}|M(X)]) - PS(\Pr[\mathbf{Y}|M(X)], \Pr[\mathbf{Y}]) \\
&= \mathbb{E}_{X,M(X)} PS(\Pr[\mathbf{Y}|X, M(X)], \Pr[\mathbf{Y}|M(X)]) - PS(\Pr[\mathbf{Y}], \Pr[\mathbf{Y}]) \\
&\qquad\qquad (PS \text{ is linear for the first entry}) \\
&= \mathbb{E}_{X,M(X)} PS(\Pr[\mathbf{Y}|X], \Pr[\mathbf{Y}|M(X)]) - PS(\Pr[\mathbf{Y}], \Pr[\mathbf{Y}]) \\
&\qquad\qquad (\text{conditioning on } X,\ M(X) \text{ is independent with } Y) \\
&\leq \mathbb{E}_X PS(\Pr[\mathbf{Y}|X], \Pr[\mathbf{Y}|X]) - PS(\Pr[\mathbf{Y}], \Pr[\mathbf{Y}]) \\
&\qquad\qquad (PS \text{ is proper}) \\
&= \mathbb{E}_X PS(\Pr[\mathbf{Y}|X], \Pr[\mathbf{Y}|X]) - PS(\Pr[\mathbf{Y}|X], \Pr[\mathbf{Y}]) \\
&= BMI^{PS}(X;Y)
\end{aligned}
$$

□

2.3 Mutual information paradigm (MIP)

General Setting We introduce the general setting (n, Σ) of the mechanism design framework where n is the number of agents and Σ is the set of possible private information. Each agent i receives a random private information / signal $\Psi_i : \Omega \mapsto \Sigma$ where Ω is the underlying sample space. She also has a prior for other agents' private information.

Formally, each agent i believes the agents' private information is chosen from a joint distribution Q_i before she receives her private information. Thus, from agent i's perspective, before she receives any private information, the probability that agent 1 receives $\Psi_1 = \sigma_1$, agent 2 receives $\Psi_2 = \sigma_2$, ..., agent n receives $\Psi_n = \sigma_n$ is $Q_i(\Psi_1 = \sigma_1, \Psi_2 = \sigma_2, ..., \Psi_n = \sigma_n)$. After she receives her private information based on her prior, agent i will also update her knowledge to a posterior distribution which is the prior conditioned on her private information. Without assuming a common prior, agents may have different priors, that is, Q_i may not equal Q_j. We define Δ_Σ as the set of all possible probability distributions over Σ.

2.3.1 Mechanism design framework: MIP

The original idea of peer prediction [45] is based on a clever insight: every agent's information is related to her peers' information and therefore can be checked using her peers' information. Inspired by this, we propose a natural yet powerful information theoretic mechanism design idea—paying every agent the "mutual information" between her reported information and her peer's reported information where the "mutual information" should be *information-monotone*—any "data processing" on the two random variables will decrease the "mutual information" between them.

Definition 28 (Information-monotone mutual information)**.** We say MI is information-monotone if and only if for any random variables $X : \Omega \mapsto \Sigma_X$ and $Y : \Omega \mapsto \Sigma_Y$:

Symmetry $MI(X;Y) = MI(Y;X)$;

Non-negativity $MI(X;Y)$ is always non-negative and is 0 if X is independent with Y;

Data processing inequality for any transition probability $M \in \mathbb{R}^{|\Sigma_X| \times |\Sigma_X|}$, when Y is independent with $M(X)$ conditioning on X, $MI(M(X);Y) \leq MI(X;Y)$.

We say MI is *strictly* information-monotone with respect to a probability measure $P \in \Delta_{\Sigma_X \times \Sigma_Y}$ if when the joint distribution over X and Y is P, for any non-permutation M, when Y is independent with $M(X)$ conditioning on X, $MI(M(X);Y) < MI(X;Y)$.

Definition 29 (Conditional mutual information)**.** Given three random variables X, Y, Z, we define $MI(X;Y|Z)$ as

$$\sum_z Pr[Z = z] MI(X;Y|Z = z)$$

where $MI(X;Y|Z = z) := MI(X';Y')$ where $Pr[X' = x, Y' = y] = Pr[X = x, Y = y|Z = z]$.

We now provide a paradigm for designing information elicitation mechanisms—the Mutual Information Paradigm. We warn the reader that *this paradigm represents some "wishful thinking" in that is it clear the paradigm cannot compute the payments given the reports.*

Mutual Information Paradigm (MIP(MI)) Given a general setting (n, Σ),

Report For each agent i, she is asked to provide her private information Ψ_i. We denote the actual information she reports as $\hat{\Psi}_i$.

Payment/Information Score We uniformly randomly pick a reference agent $j \neq i$ and denote his report as $\hat{\Psi}_j$. Agent i is paid by her information score

$$MI(\hat{\Psi}_i; \hat{\Psi}_j)$$

where MI is information-monotone.

Given a general setting (n, Σ), we say MI is *strictly information-monotone with respect to prior* Q if for every pair i, j, MI is strictly information-monotone with respect to $Q(\Psi_i, \Psi_j)$.

Resolving "wishful thinking" MIP pays agents according to the information measure. The calculation of the information measure requires the knowledge of the prior, i.e., the joint distribution which is unrealistic in practical. To removing this "wishful thinking", a key observation is that paying agents an unbiased estimator of the information measure is sufficient when we assume agents are expected payment maximizers. To construct an unbiased estimator of the information measure using agents' reports, different settings have different techniques. In the multi-task setting, either we ask a large number of questions to estimate the prior and use the prior to calculate the information measure (f-mutual information mechanism), or we ask a small number questions but require the knowledge of information structure and use a special f-mutual information, MI^{tvd} (TVD-mutual information mechanism). In the single-task setting (disagreement mechanism, BTSPrelec [51]), we ask agents their posterior (e.g. what percentage of your peers have texted while driving before?) and construct the estimator using both the first order information (e.g Y/N) and the second order information (e.g. 80% Yes). Thus, although the proposed mechanisms are based on the MIP, they are all detail free in the sense that they do not need any priori knowledge of the distributions (nor wishful thinking).

2.3.2 Analysis of MIP

Definition 30 (Mechanism)**.** We define a mechanism $\mathcal{M}$ for a setting (n, Σ) as a tuple $\mathcal{M} := (\mathcal{R}, M)$ where $\mathcal{R}$ is a set of all possible reports the mechanism allows, and $M : \mathcal{R}^n \mapsto \mathbb{R}^n$ is a mapping from all agents' reports to each agent's reward.

The mechanism requires agents to submit a report r. For example, r can simply be an agent's private information. In this case, $\mathcal{R} = \Sigma$. We call this kind of mechanism a *minimal* mechanism. We define $\mathbf{r}$ to be a report profile $(r_1, r_2, ..., r_n)$ where r_i is agent i's report.

Typically, the strategy of each agent should be a mapping from her received knowledge including her prior and her private signal, to a probability distribution over her report space $\mathcal{R}$. But since all agents' priors are fixed during the time when they play the mechanism, without loss of generality, we omit the prior in the definition of strategy.

Definition 31 (Strategy)**.** Given a mechanism $\mathcal{M}$, we define the strategy of each agent in the mechanism $\mathcal{M}$ for setting (n, Σ) as a mapping s from σ (private signal) to a probability distribution over $\mathcal{R}$.

We define a strategy profile $\mathbf{s}$ as a profile of all agents' strategies $(s_1, s_2, ..., s_n)$ and we say agents play $\mathbf{s}$ if for all i, agent i plays strategy s_i.

Note that actually the definition of a strategy profile only depends on the setting and the definition of all possible reports $\mathcal{R}$. We will need the definition of a mechanism when we define an equilibrium.

A *Bayesian Nash equilibrium* consists of a strategy profile $\mathbf{s} = (s_1, \ldots, s_n)$ such that no agent wishes to change her strategy since other strategy will decrease her expected payment, given the strategies of the other agents and the information contained in her prior and her signal.

Definition 32 (Agent Welfare)**.** Given a mechanism $\mathcal{M}$, for a strategy profile $\mathbf{s}$, we define the agent welfare of $\mathbf{s}$ as the sum of expected payments to agents when they play $\mathbf{s}$ under $\mathcal{M}$.

We can use transition matrices to represent agents' strategies of reporting their private information. Given the general setting (n, Σ), for the minimal mechanisms, fixing the priors of the agents, each agent i's strategy s_i can be seen as a transition matrix that transforms her private information Ψ_i to her reported information $\hat{\Psi}_i = s_i(\Psi_i)$. We define **truth-telling $\mathbf{T}$** as the strategy where an agent truthfully reports her private signal. $\mathbf{T}$ corresponds to an identity transition matrix.

We say agent i plays a permutation strategy if s_i corresponds to a permutation transition matrix. An example is that an agent relabels / permutes the signals and reports the permuted version (e.g. she reports "good" when her private signal is "bad" and reports "bad" when her private signal is "good"). Note that $\mathbf{T}$ [1] is a permutation strategy as well. We call the strategy profile where all agents play a permutation strategy a *permutation strategy profile*. Note that in a permutation strategy profile, agents may play different permutation strategies. When a permutation strategy profile is a Bayesian Nash equilibrium, we call such equilibrium a *permutation equilibrium*.

We hope our mechanisms can be strictly truthful, focal, and even dominantly truthful (see informal definitions in Section 1.1.3 and formal definitions will be introduced later). Here we propose two additional, stronger equilibrium goals. A mechanism $\mathcal{M}$ is strongly focal if the truth-telling strategy profile maximizes *every* agent's expected payment among all strategy profiles, while in the focal mechanism, truth-telling maximizes the agent welfare—the sum of agents' expected payment. A mechanism $\mathcal{M}$ is truth-monotone if when any truthful agent changes to play a non-truthful strategy s, no matter what strategies other agents play, it decreases every

[1] The above definitions of $\mathbf{T}$ and the permutation strategy are sufficient to analyze the framework. When considering more general settings, we will provide generalized definitions of $\mathbf{T}$ and the permutation strategy.

agent's expected payment. Note that the truth-monotone property is stronger than the strongly focal or focal property and it says any non-truthful behavior of any agent will hurt everyone. In addition to the above equilibrium goals, we also hope the mechanism can be minimal and detail free (see definitions in Section 1.1.3).

For the strictness guarantee, it turns out no truthful detail free mechanism can make truth-telling strategy profile be strictly better than any permutation strategy profile. Therefore, the best we can hope is making the truth-telling strategy profile be strictly better than any other non-permutation strategy profile. We give the formal definitions for the equilibrium goals with the strictness guarantee in the following paragraph.

Mechanism Design Goals

(Strictly) Truthful A mechanism $\mathcal{M}$ is (strictly) truthful if for every agent, **T** (uniquely) maximizes her expected payment given that everyone else plays **T**.

(Strictly) Dominantly truthful A mechanism $\mathcal{M}$ is dominantly truthful if for every agent, **T** maximizes her expected payment no matter what strategies other agents play. A mechanism $\mathcal{M}$ is strictly dominantly truthful if for every agent, if she believes at least one other agent will tell the truth, playing **T** pays her strictly higher than playing a non-permutation strategy.

(Strictly) Focal A mechanism $\mathcal{M}$ is (strictly) focal if the truth-telling equilibrium maximizes the agent welfare among all equilibria (and any other non-permutation equilibrium has strictly less agent welfare).

(Strictly) Strongly focal A mechanism $\mathcal{M}$ is (strictly) strongly focal if the truth-telling strategy profile maximizes *every* agent's expected payment among all strategy profiles (and in any other non-permutation strategy profile, every agent's expected payment is strictly less).

(Strictly) Truth-monotone A mechanism $\mathcal{M}$ is (strictly) truth-monotone if when any truthful agent changes to play a non-truthful strategy s, no matter what strategies other agents play, it decreases every agent's expected payment (and strictly decreases every other truthful agent's expected payment if s is a non-permutation strategy).

Section 2.5 will show that it is impossible to ask the truth-telling strategy profile to be strictly better than other permutation strategy profiles when the mechanism is detail free. Thus, the strictly truth-monotone is the optimal property for equilibrium selection when the mechanism is detail free.

Theorem 33. *Given a general setting (n, Σ), when MI is (strictly) information-monotone (with respect to every agent's prior), the Mutual Information Paradigm $MIP(MI)$ is (strictly) dominantly truthful, (strictly) truth-monotone.*

Theorem 33 almost immediately follows from the data processing inequality of the mutual information. The key observation in the proof is that *applying any strategy to the information is essentially data processing and thus erodes information.*

Note that the Mutual Information Paradigm is not a mechanism since it requires the mechanism to know the full joint distribution over all agents' random private information while agents only report (or even have access to) a realization / sample of the random private information. Rather, if we design mechanisms such that the payment in the mechanism is an unbiased estimator of the payment in Mutual Information Paradigm, the designed mechanisms will obtain the desirable properties immediately according to Theorem 33. In the future sections, we will see how to design such mechanisms in both the multi-question and single-question settings.

Proof. For each agent i, for any strategy s_i she plays, comparing with the case she

honestly reports Ψ_i, her expected information score is

$$\sum_{j\neq i}\frac{1}{n-1}MI(\hat{\Psi}_i;\hat{\Psi}_j)=\sum_{j\neq i}\frac{1}{n-1}MI(s_i(\Psi_i);\hat{\Psi}_j)\leq\sum_{j\neq i}\frac{1}{n-1}MI(\Psi_i;\hat{\Psi}_j)$$

since MI is information-monotone. Thus, MIP(MI) is dominantly truthful when MI is information-monotone.

For the strictness guarantee, we need to show when agent i believes at least one agent tells the truth, for agent i, any non-permutation strategy will strictly decrease her expected payment. Let's assume that agent i believes agent $j_0 \neq i$ plays $\mathbf{T}$. When MI is strictly information-monotone with respect to every agent's prior, MI is strictly information-monotone with respect to $Q_i(\Psi_i, \Psi_{j_0})$ as well. Then the inequality of the above formula is strict if agent i plays a non-permutation strategy s_i since $MI(s_i(\Psi_i);\hat{\Psi}_{j_0}) = MI(s_i(\Psi_i);\Psi_{j_0}) < MI(\Psi_i,\Psi_{j_0})$.

Thus, when MI is strictly information-monotone with respect to every agent's prior, MIP(MI) is strictly dominantly truthful.

Fixing other agents' strategies except agent k, for $i \neq k$, agent i's expected payment is

$$\begin{aligned}\sum_{j\neq i}\frac{1}{n-1}MI(\hat{\Psi}_i;\hat{\Psi}_j)&=\sum_{j\neq i,k}\frac{1}{n-1}MI(\hat{\Psi}_i;\hat{\Psi}_j)+\frac{1}{n-1}MI(\hat{\Psi}_i;\hat{\Psi}_k)\\&\leq\sum_{j\neq i,k}\frac{1}{n-1}MI(\hat{\Psi}_i;\hat{\Psi}_j)+\frac{1}{n-1}MI(\hat{\Psi}_i;\Psi_k).\end{aligned}$$

Thus, agent i's expected payment decreases when truthful agent k changes to play a non-truthful strategy. For $i = k$, the dominantly truthful property already shows agent $i = k$'s expected payment will decrease when truthful agent k changes to play a non-truthful strategy. Therefore when MI is information-monotone, MIP(MI) is truth-monotone.

For the strictness guarantee, when MI is strictly information-monotone with respect to every agent's prior, if truthful agent k changes to play a non-permutation strategy s_k, then a truthful agent i's expected payment will strictly decrease since $MI(\Psi_i; s_k(\Psi_k)) < MI(\Psi_i; \Psi_k)$ if s_k is a non-permutation strategy and MI is strictly information-monotone.

Therefore, when MI is (strictly) information-monotone (with respect to every agent's prior), MIP(MI) is (strictly) truth-monotone. □

Theorem 17 and Theorem 33 imply the following corollary.

Corollary 34. *Given a general setting (n, Σ), when f is (strictly) convex (and every agent's prior is fine-grained), the Mutual Information Paradigm MIP(MI^f) is (strictly) dominantly truthful, (strictly) truth-monotone.*

If we use Bregman mutual information instead of f-mutual information, the dominantly truthful property will still be preserved.

Theorem 35. *Given a general setting (n, Σ), when MI is quasi information-monotone, the Mutual Information Paradigm MIP(MI) is dominantly truthful.*

Proof. For each agent i, for any strategy s_i she plays, comparing with the case she honestly reports Ψ_i, her expected information score is

$$\sum_{j \neq i} MI(\hat{\Psi}_i; \hat{\Psi}_j) = \sum_{j \neq i} MI(s_i(\Psi_i); \hat{\Psi}_j) \leq \sum_{j \neq i} MI(\Psi_i; \hat{\Psi}_j)$$

which is less than if she had reported truthfully since quasi information-monotone MI has data processing inequality for the first entry.

□

Corollary 36. *Given a general setting (n, Σ), the Mutual Information Paradigm MIP(BMI^{PS}) is dominantly truthful.*

2.4 Hierarchical mutual information paradigm (HMIP)

In this section, we will define the hierarchical information structure and provide a mechanism design framework that helps design mechanisms which elicit the hierarchical information. Section 6.3 defines the information model; Section 2.4.2 defines our mechanism framework; and Section 2.4.3 analyzes the framework. We will use the peer grading process (Figure 2.1) as a running example to throughout this section.

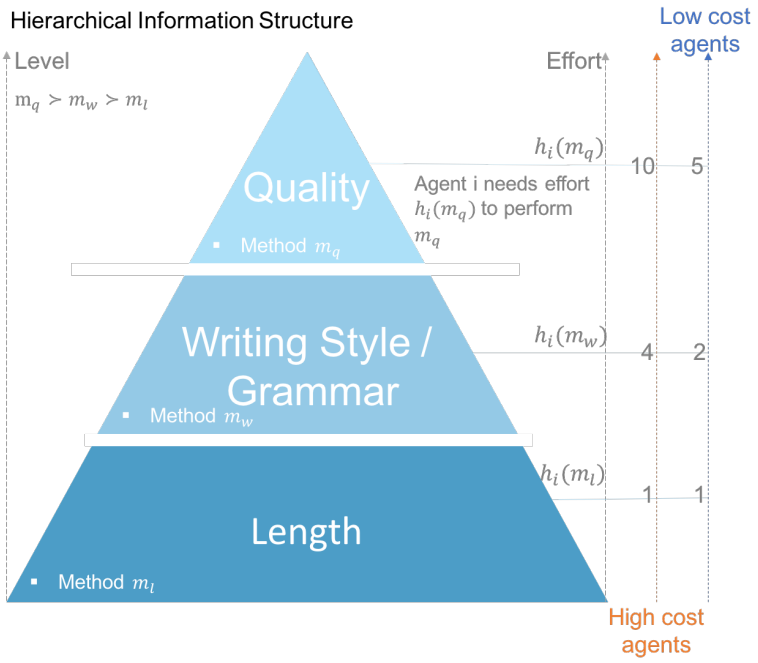

Figure 2.1: An illustration of the hierarchical information structure in the peer grading process.

2.4.1 Hierarchical information structure

There are n agents and one task. The agents have a finite set M of methods to perform on the task based on the task's attributes $\mathbf{a} \in A$ where a is a random (possibly high dimensional) vector drawn from a distribution $Q_A \in \Delta_A$. Each method $m : A \mapsto \Sigma_m$ maps the attributes $\mathbf{a} \in A$ to a signal $m(\mathbf{a})$ from a finite set Σ_m. We now introduce our peer grading example.

10 evaluators are asked to judge one essay. The essay has eight possible attributes: $\mathbf{a} = (q_i, w_j, l_k), i, j, k \in \{0, 1\}$. (q_1, w_0, l_1) means the essay has (good quality, bad writing, long length); (q_0, w_1, l_0) means the essay has (bad quality, good writing, short length). The distribution over the attributes space Q_A is defined as:

$Q_A((q_0, w_0, *))$	$Q_A((q_0, w_1, *))$	$Q_A((q_1, w_0, *))$	$Q_A((q_1, w_1, *))$
0.4	0.1	0.1	0.4

With this distribution, an essay of good quality usually has good writing as well. Moreover, we assume the essay's length is independent with the essay's quality and writing and an essay has long length with probability 0.5. That is: $Q_A((q_i, w_j, l_1)) = Q_A((q_i, w_j, *)) * 0.5, Q_A((q_i, w_j, l_0)) = Q_A((q_i, w_j, *)) * 0.5$.
Each evaluator can perform three methods: $m_l(\mathbf{a})$, $m_w(\mathbf{a})$, and $m_q(\mathbf{a})$ which are, respectively, (possibly noisy) signals about the essay's length; writing style and grammar; and quality. $\Sigma_l = \Sigma_w = \Sigma_q = \{\text{☺}, \text{☹}\}$.

We define $\psi_i^m(\mathbf{a})$ as agent i's received output by performing m on attributes $\mathbf{a}$. Different agents may receive different signals by performing the same method on the same attributes. But we assume the distribution is symmetric/homogeneous in the sense that for any permutation $\pi : [n] \mapsto [n]$, the probability that $\psi_1^m(\mathbf{a}) = \sigma_1$, $\psi_2^m(\mathbf{a}) = \sigma_2$, ... $\psi_n^m(\mathbf{a}) = \sigma_n$ equals the probability that $\psi_{\pi(1)}^m(\mathbf{a}) = \sigma_1$, $\psi_{\pi(2)}^m(\mathbf{a}) = \sigma_2$, ... $\psi_{\pi(n)}^m(\mathbf{a}) = \sigma_n$. We also assume that each agent performs methods independently (see (2.8)). When the attributes $\mathbf{a}$ is drawn from a distribution Q_A, we can define define Ψ_i^m as agent i's received output by performing m on a random attributes $\mathbf{a}$ that is drawn from a distribution Q_A. Analogously, we define a random variable Ψ_{-i}^m as an arbitrary agent $j \neq$'s received output by performing m on a random attributes $\mathbf{a}$ that is drawn from a distribution Q_A. This definition is well-defined since we have assumed the distribution is symmetric. We define prior Q as a joint distribution over all $\{\Psi_i^m\}_{i \in [n], m \in M}$.

Conditioning on the attributes of the essay $\mathbf{a} = (q_i, w_j, l_k), i, j, k \in \{0, 1\}$, for each method m, each agent will receive $\psi_i^m(\mathbf{a}) = ☺$ with probability $p_{m,\mathbf{a}}$ independently by performing m. That is, agents' received signals by performing m is a Binomial distribution $B(n = 10, p_{m,\mathbf{a}})$. [a]

	good quality essay[b]	bad quality essay
$\Pr[m_q(\mathbf{a}) = ☺]$	70%	30%

This means conditioning on the essay having good quality, the distribution over agents' received quality signals by performing m_q is $Q_{m_q,(q_1,*,*)} = B(10, 0.7)$; while conditioning on the essay having good quality, the distribution is $Q_{m_q,(q_0,*,*)} = B(10, 0.3)$. Similarly, we have

	good writing essay	bad writing essay
$\Pr[m_w(\mathbf{a}) = ☺]$	90%	10%

	long essay	short essay
$\Pr[m_l(\mathbf{a}) = ☺]$	100%	0%

Note that the cheap length signal is noiseless. We also assume that fixing the attributes, every agent performs the different methods independently. That is, when $\mathbf{a} = (q_1, w_1, l_1)$

$$\Pr\left(\Psi_i^{m_l}(\mathbf{a}) = ☺, \Psi_i^{m_w}(\mathbf{a}) = ☺, \Psi_i^{m_q}(\mathbf{a}) = ☺\right) = 0.7 * 0.9 * 1. \tag{2.8}$$

With the above set up, the probability that agent i receives a ☺ writing signal and agent j receives a ☺ quality signal will be
$\Pr[\Psi_i^{m_w} = ☺, \Psi_j^{m_q} = ☺] = 0.4*0.1*0.3+0.1*0.3*0.9+0.1*0.7*0.1+0.4*0.7*0.9 = 0.298$.

[a] To give a concrete example, we use the Binomial distribution here. In fact, we only need

the distribution to be symmetric.

[b]This means $\mathbf{a} = (q_1, *, *)$

We define a partial order on the methods. We say $m_1 \succeq m_2$—the level of m_1 is higher than that of m_2—if method m_1 cannot be performed without performing m_2. By $m_1 \succ m_2$ we mean $m_1 \succeq m_2$ but $m_2 \not\succeq m_1$. Note that the partial order $\succ$ is transitive—$m_1 \succ m_2, m_2 \succ m_3 \Rightarrow m_1 \succ m_3$. Each agent i needs effort $h_i(m) > 0$ to perform method m and when she spends effort $h_i(m)$ to perform m, the methods that have lower levels than m are performed as well without additional effort. We assume, as is natural, that $h_i(m)$ is an increasing function, that is, $h_i(m_1) \geq h_i(m_2)$ when $m_1 \succeq m_2$. The higher the level of the method an agent performs, the more effort she must invest. However, it may be the case that some agents (low cost agents) can perform methods more economically than others (high cost agents). The partial order definition is essentially our key assumption (Assumption 3).

$m_q \succ m_w \succ m_a$. Among the 10 evaluators, there are 2 low cost evaluators who need $1, 2, 5$ effort to perform m_l, m_w, m_q respectively. There are 8 high cost evaluators who need $1, 4, 10$ effort to perform m_l, m_w, m_q respectively (Figure 2.1). Based on the partial order definition, when an evaluator spends sufficient effort to perform m_q and obtains the quality signal, she also obtains the length and writing signals without additional effort, which is natural in real life.

We assume agents share a hierarchical information structure and allow agents to have different priors Q[2]. We will design mechanisms that incentivize agents to invest efforts based on their costs and report honestly.

2.4.2 Mechanism design framework: HMIP

We start by introducing the formal definition of a mechanism.

[2]To ease the presentation of the example, in our peer grading example, we assume agents share the same prior Q.

Definition 37 (Mechanism)**.** We define a *mechanism* $\mathcal{M}$ for n agents as a tuple $\mathcal{M} := (R, S)$ where R is a set of all possible reports the mechanism allows, and $S : R^n \mapsto \mathbb{R}^n$ is a mapping from all agents' reports to each agent's payment.

We will extend the Mutual Information Paradigm to the Hierarchical Mutual Information Paradigm that handles the hierarchical information structure. We can naturally extend the Mutual Information Paradigm for Peer Prediction Mechanisms to the hierarchical model by paying agent i

$$MI^f(\text{her information}; \{\Psi_{-i}^m\}_{m \in M})^3.$$

This idea has a severe drawback: sometimes low level information has very large correlation with the high level information. In this case, $MI^f(\text{her information}; \{\Psi_{-i}^m\}_{m \in M})$ will pay low level information nearly as much as high level information; and so agents will lack incentive to perform high level methods.

To solve the above problem, we pay agents method by method. For each m, we only value the "information gain" in the sense that we pay each agent i the mutual information between her information and the method m's information *conditioning on* the information output by the methods are lower than m.

Formally, we chose a payment scale α_m for each m and pay each agent i

$$\sum_m \alpha_m MI^f(\text{her information}; \Psi_{-i}^m | \{\Psi_{-i}^{m'}\}_{m' \prec m}).$$

In our actual paradigm, we hope to pay each agent i using the above payment when the mechanism has access to all levels of honest information provided by other agents.

[3]Recall that $(\Psi_1^m, \Psi_2^m, ..., \Psi_n^m)$ are the random signals agents receive by performing method m on the same random attributes.

In the peer grading example, the information about the writing style / grammar may already have a very high correlation with the quality of the essay. With the above concrete set up, we are ready to calculate the (conditional) Shannon mutual information (Euler number base) between agent i's received signals and agent j's received signals. For example, the 2×2 entry is the mutual information between agent i's received length signal, writing signal by performing method m_w and agent j's writing signal, conditioning on agent j's length signal, which is 0.2259. We calculate the values by first calculating the joint distribution over 6 random variables—agent i's length, writing, quality signals and agent j's length, writing, quality signals.

We show the values in the following table and defer the calculation to Appendix. According to the information monotonicity, for each column, the values increase from bottom to top.

$MI(\cdot;\cdot)$ agent j's / agent i's	length	writing|length [a]	quality | writing, length	length, writing, quality[b]
length, writing, quality	0.6931	0.2259	**0.0115**	**0.9305**
length, writing	0.6931	0.2218	**0.0041**	**0.9190**
length	0.6931	0	0	0.6931

Even though performing the quality method provides the information that has the highest mutual information 0.9305 with other agents' information, performing writing method already outputs information that has $0.9190 \approx 0.9305 * 0.98$ mutual information with other agents' information.

In this case, what we really value is the additional quality of information after conditioning on the information of cheap signals like writing style / grammar. In other words, we value the information about an essay which has a high quality but is written carelessly (or low quality but impeccable prose).

Each agent, performing the writing method only has 0.0041 mutual information with other agents' quality signal conditioning on other agents' writing and length

signals while performing the quality method has $0.0115 \approx 0.0041 * 2.80$ conditional mutual information.

Looking ahead, we seek to pay each evaluator i by:

$$\begin{aligned} &\alpha_l MI^f(\text{her information}; \text{agent } j\text{'s length signal}) \qquad (2.9)\\ &+\alpha_w MI^f(\text{her information}; \text{agent } j\text{'s writing signal}|\text{agent } j\text{'s length signal})\\ &+\alpha_q MI^f(\text{her information}; \text{agent } j\text{'s quality signal}|\text{agent } j\text{'s length \& writing signal}). \end{aligned}$$

where α_q is set to be rather larger than α_l and α_w.

[a] x|y means x conditioning on y.

[b] Since we use Shannon mutual information which satisfies chain rule, the last column is the sum of the previous columns.

Hierarchical Mutual Information Paradigm (HMIP(MI^f,$\{\alpha_m\}_m$)) We now present our hierarchical Mutual Information Paradigm. We emphasize that this is not a mechanism that can be run. Instead we engage in the wishful thinking that the reports of the agents are distributions rather than draws from the distribution. Of course, this will never happen. Nonetheless, we will show that using the HMIP paradigm we can design actual mechanisms in both the multiple-task setting (Section 6.5.1) and the single-task setting (Section 6.5.2).

The paradigm requires as parameters a payment scale $\alpha_m \in \mathbb{R}_{\geq 0}$ for each method m.

Report For each agent i, for each $m \in M$, she is asked to optionally provide the random signal Ψ_i^m. We denote the set of methods whose outputs are reported by agent i as M_i and the actual random signal she reports for each $\ell \in M_i$ as $\hat{\Psi}_i^\ell$.

Payment/Information Score We define M_{-i} as $\bigcup_{j \neq i} M_j$. For each $m \in M_{-i}$, we

arbitrarily pick an agent $j \neq i$ who provides method m's output and denote his report for method m's output as $\hat{\Psi}^m_{-i}$.

Agent i is paid by her information score

$$\sum_{m \in M_{-i}} \alpha_m MI^f(\{\hat{\Psi}^\ell_i\}_{\ell \in M_i}; \hat{\Psi}^m_{-i} | \{\hat{\Psi}^{m'}_{-i}\}_{m' \prec m, m' \in M_{-i}})$$

We use the same techniques introduced in Section 2.3.1 to resolve "wishful thinking".

2.4.3 Analysis of HMIP

For each agent i, we define her utility as her payment minus her effort.

Definition 38 (Strategy)**.** We define the *effort strategy* of each agent i as a mapping e_i from her priors to a probability distribution over the methods she will perform. We define the *report strategy* of each agent i as a mapping s_i from her received information to a probability distribution over R.

Definition 39 (Amount of information in HMIP)**.** In HMIP, for agent i, the amount of information acquired with method m_i is defined as

$$AOI(m_i, \text{HMIP}(MI^f, \{\alpha_m\}_m)) := \sum_{m \in M} \alpha_m MI^f(\{\Psi^\ell_i\}_{\ell \preceq m_i}; \Psi^m_{-i} | \{\Psi^{m'}_{-i}\}_{m' \prec m}).$$

We have already give the example of the amount of information in (2.9). Later in the proof of Theorem 43, we will see the amount of information is also the optimal payment of agent i who performs method m_i when HMIP has access to all levels of honest signals reported by other agents.

An especially desirable strategy in HMIP is a *prudent strategy*. Informally, agents play a prudent strategy if they (a) choose the method they perform to maximize their

utility—trading off the amount of information acquired with the effort it costs; (b) report all received information honestly.

Definition 40 (Prudent strategy in HMIP)**.** For each agent i, we say she plays a prudent strategy in HMIP(MI^f,$\{\alpha_m\}_m$) if she chooses to (a) perform method m_i^* such that

$$m_i^* \in \arg\max_{m_i} \left(AOI(m_i, \text{HMIP}(MI^f, \{\alpha_m\}_m)) - h_i(m_i)\right); and$$

(b) reports all received information honestly.

Definition 41 (Truthful strategy in HMIP)**.** We say an agent plays truthful strategy if she always reports her received information honestly.

A truthful strategy is a special report strategy. An agent can play any effort strategy and truthful strategy simultaneously. If an agent invests no effort and reports nothing or meaningless information, she is still considered as playing truthful strategy.

Mechanism design goals A mechanism $\mathcal{M}$ is (strictly) *potent* if for each agent, when she believes everyone else plays their prudent strategy, she can (strictly) maximize her expected utility by playing a prudent strategy as well. A mechanism $\mathcal{M}$ is *dominant truthful* if for each agent, regardless of other agents' strategies, she can maximize her expected utility by playing a pure effort strategy and truthful strategy.

The dominant truthful property is incomparable with the potent property. A flat payment scheme is dominant truthful but not potent since investing no effort and reporting nothing is also considered as a pure effort and truthful strategy. The potent property is desirable since it encourages low cost agents to invest high level effort and high cost agents to invest low level effort, and incentivizes them to report honestly as well.

In order to design potent mechanism, the coefficients $\{\alpha_m\}_m$ should be chosen appropriately.

We say a method m is *maximal* if there does not exist $m' \neq m \in M$ such that $m' \succ m$.

Definition 42 (potent coefficients for HMIP)**.** Given the priors $\{Q_m\}_m$, we say the coefficients $\{\alpha_m\}_m$ are potent for HMIP(MI^f,$\{\alpha_m\}_m$) if given the coefficients $\{\alpha_m\}_m$, for every maximal m, there exists at least **two** agents whose prudent strategy in HMIP(MI^f,$\{\alpha_m\}_m$) is performing method m.

This is a weak requirement since we only need to set sufficiently high coefficients to incentivize **two** low cost agents such that for each agent (including one of the low cost agent), she will believe there exists a low cost agent who will be incentivized to report all levels of information. Potent coefficients exist since we can always set the coefficient of the highest level information sufficiently high and the coefficients of other levels arbitrarily close to zero such that agents will be incentivized to invest the highest level effort. We use our peer grading example to show how to set potent coefficients. With our example, we will see we can always set the optimal potent coefficients that minimize the mechanism's cost by solving a linear programming.

> In our example, the 2 low cost agents need efforts 1,2,5 to perform m_l, m_w, m_q respectively and 8 high cost agents need efforts 1,4,10. With the above set up, we need $\alpha_q * 0.0115 + \alpha_w * 0.2259 + \alpha_l * 0.6931 - 5 > \max\{\alpha_q * 0.0041 + \alpha_w * 0.2218 + \alpha_l * 0.6931 - 2, \alpha_l * 0.6931 - 1, 0\}$ to make the coefficients potent and we also want to minimize the mechanism's cost which is

$$2 * (\alpha_q * 0.0115 + \alpha_w * 0.2259 + \alpha_l * 0.6931)$$

$$+ 8 * \begin{cases} v_q := \alpha_q * 0.0115 + \alpha_w * 0.2259 + \alpha_l * 0.6931 & \text{if } v_q - 10 \geq v_w - 4, v_l - 1, 0 \\ v_w := \alpha_q * 0.0041 + \alpha_w * 0.2218 + \alpha_l * 0.6931 & \text{if } v_w - 4 \geq v_q - 10, v_l - 1, 0 \\ v_l := \alpha_l * 0.6931 & \text{if } v_l - 1 \geq v_w - 4, v_q - 10, 0 \\ 0 & \text{otherwise} \end{cases}$$

After solving this linear programming, the optimal solution is around $\alpha_l, \alpha_w, \alpha_q = \epsilon^a, 0.5562, 423.8571$ and the amount of information for performing m_l,m_w,and m_q are $O(\epsilon)$, $1.86+O(\epsilon)$, and $5+O(\epsilon)$ respectively. The minimal cost is $2*(5+O(\epsilon)) = 10 + O(\epsilon)$.

[a] ϵ is an arbitrarily small positive real number, we need ϵ since we want agents will be incentivized to report the length signal as well.

Theorem 43. *Given a convex function f, HMIP(MI^f,$\{\alpha_m\}_m$) is dominant truthful; moreover, when $\{\alpha_m\}_m$ are potent for HMIP(MI^f,$\{\alpha_m\}_m$), HMIP(MI^f,$\{\alpha_m\}_m$) is potent and dominant truthful.*

The proof of the theorem uses the information monotonicity of MI^f. The key observation in the proof is that *applying any strategy to the information is essentially data processing and thus erodes information.*

Proof for Theorem 43. In order to show the dominant truthful property, we will show for each agent, fixing any other agents' strategies, she can maximize her payment as well as her utility by reporting her received information honestly. The information monotonicity property of f-mutual information MI^f (Fact 17) says any data processing decreases the (conditional) mutual information. For each $m \in M_{-i}$, fixing the strategies other agents use, the distribution of $\hat{\Psi}^m_{-i}$, whose randomness comes from random variable Ψ^m_{-i} and the agents' strategies, is also fixed. Any strategy (data

processing) agent i applies to her received signals decreases

$$MI^f(\text{her received signals}; \hat{\Psi}^m_{-i} | \{\hat{\Psi}^{m'}_{-i}\}_{m' \prec m, m' \in M_{-i}}).$$

Thus, for agent i, honestly reporting her received signals maximizes her payment no matter what strategies other agents use.

We start to show HMIP is potent when the coefficients are potent . When the coefficients are potent , for every agent i, when she believes everyone else plays a prudent strategy, she will believe for each m, there exists an agent $j(m) \neq i$ who reports $\{\Psi^\ell_j\}_{\ell \preceq m}$

□

HMIP provides a framework to design information elicitation mechanisms for our hierarchical information model. To apply the HMIP framework in different settings, it remains to design the report requirement for agents and to use agents' reports to calculate the (conditional) mutual information without underlying distributions. We apply HMIP in both the multi-task setting (Section 6.5.1) and the single-task setting (Section 6.5.2).

2.5 Impossibility (Tightness) results

In this section, we will show an impossibility result that implies the optimality of the information theoretical framework. We will see when the mechanism knows no information about the prior profile, no non-trivial mechanism has truth-telling as the *unique* "best" equilibrium. Thus, it is too much to ask for a mechanism where truth-telling is paid strictly higher than any other non-truthful equilibrium. The best we can hope is to construct a mechanism where truth-telling is paid strictly higher than all non-truthful equilibria / strategy profiles excluding all *permutation strategy profiles* (Definition 48) when the prior is symmetric; or all non-truthful equilibria /

strategy profiles excluding all *generalized permutation strategy profiles* (Definition 49) when the prior may be asymmetric. Because permutation strategies seem unnatural, risky, and require the same amount of effort as truth-telling these are still strong guarantees.

Actually we will show a much more general result in this section that is sufficiently strong to imply the optimality of the framework. Recall that a mechanism is strictly focal if truth-telling is strictly better than any other strategy profiles excluding generalized permutations strategy profiles. The results of this section imply that no truthful detail free mechanism can pay truth-telling $\mathbf{T}$ strictly better than all generalized permutations strategy profiles (Definition 49) *no matter what definition is the truth-telling strategy* $\mathbf{T}$.

We omit the prior in the definition of strategy before since we always fix the prior. However, when proving the impossibility results, the prior is not fixed. Therefore, we use the original definition of strategy in this section.

Definition 44 (Strategy)**.** Given a mechanism $\mathcal{M}$, we define the strategy of $\mathcal{M}$ for setting (n, Σ) as a mapping s from (σ, Q) (private signal and prior) to a probability distribution over $\mathcal{R}$.

(Generalized) Permutation Strategy Profiles A permutation $\pi : \Sigma \mapsto \Sigma$ can be seen as a relabelling of private information. Given two lists of permutations $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_n)$, $\boldsymbol{\pi}' = (\pi'_1, \pi'_2, ..., \pi'_n)$, we define the product of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ as

$$\boldsymbol{\pi} \cdot \boldsymbol{\pi}' := (\pi_1 \cdot \pi'_1, \pi_2 \cdot \pi'_2, ..., \pi_n \cdot \pi'_n)$$

where for every i, $\pi_i \cdot \pi'_i$ is the group product of π_i and π'_i such that $\pi_i \cdot \pi'_i$ is a new permutation with $\pi_i \cdot \pi'_i(\sigma) = \pi_i(\pi'_i(\sigma))$ for any σ.

We also define $\boldsymbol{\pi}^{-1}$ as $(\pi_1^{-1}, \pi_2^{-1}, ..., \pi_n^{-1})$.

By abusing notation a little, we define $\boldsymbol{\pi} : \mathcal{Q} \mapsto \mathcal{Q}$ as a mapping from a prior Q

to a *generalized permuted prior* $\boldsymbol{\pi}(Q)$ where for any $\sigma_1, \sigma_2, ..., \sigma_n \in \Sigma$,

$$\boldsymbol{\pi}(Q)(\sigma_1, \sigma_2, ..., \sigma_n) = Q(\pi_1^{-1}(\sigma_1), \pi_2^{-1}(\sigma_2), ..., \pi_2^{-1}(\sigma_n))$$

where σ_i is the private signal of agent i. Notice that it follows that:

$$\boldsymbol{\pi}(Q)(\pi_1(\sigma_1), \pi_2(\sigma_2), ..., \pi_2(\sigma_n)) = Q(\sigma_1, \sigma_2, ..., \sigma_n).$$

Intuitively, $\boldsymbol{\pi}(Q)$ is the same as Q after the signals are relabelled according to $\boldsymbol{\pi}$.

Definition 45 (Permutation List Operator on Strategy)**.** For every agent i, given her strategy is s_i and a permutation list $\boldsymbol{\pi}$, we define $\boldsymbol{\pi}(s_i)$ as the strategy such that $\boldsymbol{\pi}(s_i)(\sigma, Q) = s_i(\pi_i(\sigma), \boldsymbol{\pi}(Q))$ for every private information σ and prior Q.

Definition 46 (Permutation List Operator on Strategy Profile)**.** Given a permutation list $\boldsymbol{\pi}$, for any strategy profile $\mathbf{s}$, we define $\boldsymbol{\pi}(\mathbf{s})$ as a strategy profile with $\boldsymbol{\pi}(\mathbf{s}) = (\boldsymbol{\pi}(s_1), \boldsymbol{\pi}(s_2), ..., \boldsymbol{\pi}(s_n))$.

Note that $\boldsymbol{\pi}^{-1}\boldsymbol{\pi}Q = Q$ which implies $\boldsymbol{\pi}^{-1}\boldsymbol{\pi}(\mathbf{s}) = \mathbf{s}$.

We say $(\pi, \pi, ..., \pi)$ is a *symmetric* permutation list for any permutation π. For convenience, we write $(\pi, \pi, ..., \pi)(Q)$ as $\pi(Q)$, $(\pi, \pi, ..., \pi)(s)$ as $\pi(s)$ and $(\pi, \pi, ..., \pi)(\boldsymbol{s})$ as $\pi(\boldsymbol{s})$.

We define a permutation strategy (profile) and then give a generalized version of this definition.

Definition 47 (Permutation Strategy)**.** We define a strategy s as a permutation strategy if there exists a permutation π such that $s = \pi(\mathbf{T})$.

Definition 48 (Permutation Strategy Profile)**.** We define a strategy profile $\mathbf{s}$ as a permutation strategy profile if there exists a permutation π such that $\mathbf{s} = \pi(\mathbf{T})$.

Definition 49 (Generalized Permutation Strategy Profile)**.** We define a strategy profile $\mathbf{s}$ as a generalized permutation strategy profile if there exists a permutation list $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_n)$ such that $\mathbf{s} = \boldsymbol{\pi}(\mathbf{T}) = (\pi_1, \pi_2, ..., \pi_n)(\mathbf{T})$.

2.5.1 Tightness proof

Definition 50. Given a prior profile $\mathbf{Q} = (Q_1, Q_2, ..., Q_n)$ and a strategy profile $\mathbf{s} = (s_1, s_2, ..., s_n)$, and a mechanism $\mathcal{M}$, for every agent i, we define

$$\nu_i^{\mathcal{M}}(n, \Sigma, \mathbf{Q}, \mathbf{s})$$

as agent i's ex ante expected payment when agents play $\mathbf{s}$ and all agents' private information is drawn from Q_i that is, from agent i's viewpoint.

The impossibility result is stated as following:

Proposition 51. Let $\mathcal{M}$ be a mechanism that does not know the prior profile, then for any strategy profile s, and any permutation list $\boldsymbol{\pi}$:
(1) $\mathbf{s}$ is a strict Bayesian Nash equilibrium of $\mathcal{M}$ for any prior profile iff $\boldsymbol{\pi}(\mathbf{s})$ is a strict Bayesian Nash equilibrium of $\mathcal{M}$ for any prior profile.
(2) For every agent i, there exists a prior profile $\mathbf{Q}$ such that $\nu_i^{\mathcal{M}}(n, \Sigma, \mathbf{Q}, \mathbf{s}) \leq \nu_i^{\mathcal{M}}(n, \Sigma, \mathbf{Q}, \boldsymbol{\pi}(\mathbf{s}))$.

Additionally, if the mechanism knows the prior is symmetric, the above results only hold for any symmetric permutation list $(\pi, \pi, ..., \pi)$.

Proposition 51 implies

Corollary 52. *Let $\mathcal{M}$ be a truthful mechanism, given truth-telling strategy $\mathbf{T}$, when $\mathcal{M}$ knows no information about the prior profile of agents, if there exists a permutation list $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}(\mathbf{T}) \neq \mathbf{T}$, $\mathbf{T}$ cannot be always paid strictly higher than all generalized permutation strategy profiles.*

Additionally, if the mechanism knows the prior is symmetric, the above results only hold for any symmetric permutation list $(\pi, \pi, ..., \pi)$ *and all permutation strategy profiles.*

We note that the requirement that there exists $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}(\mathbf{T}) \neq \mathbf{T}$ only fails for very trivial mechanisms where the truthfully reported strategy does not depend on the signal an agent receives.

The key idea to prove this theorem is what we refer to as **Indistinguishable Scenarios**:

Definition 53 (Scenario)**.** We define a scenario for the setting (n, Σ) as a tuple $(\mathbf{Q}, \mathbf{s})$ where $\mathbf{Q}$ is a prior profile, and $\mathbf{s}$ is a strategy profile.

Given mechanism $\mathcal{M}$, for any scenario $A = (\mathbf{Q}_A, \mathbf{s}_A)$, we write $\nu_{i_A}^{\mathcal{M}}(n, \Sigma, A)$ as agent i_A's ex ante expected payment when agents play $\mathbf{s}_A$ and all agents' private signals are drawn from Q_{i_A}.

For two scenarios $A = (\mathbf{Q}_A, \mathbf{s}_A)$, $B = (\mathbf{Q}_B, \mathbf{s}_B)$ for setting (n, Σ), let $\sigma_A := (\sigma_{1_A}, \sigma_{2_A}, ..., \sigma_{n_A})$ be agents $(1_A, 2_A, ..., n_A)$' private signals respectively in scenario A, $\sigma_B := (\sigma_{1_B}, \sigma_{2_B}, ..., \sigma_{n_B})$ be agents $(1_B, 2_B, ..., n_B)$' private signals respectively in scenario B.

Definition 54 (Indistinguishable Scenarios)**.** We say two scenarios A, B are indistinguishable $A \approx B$ if there is a coupling of the random variables σ_A and σ_B such that $\forall i$, $s_{i_A}(\sigma_{i_A}, Q_{i_A}) = s_{i_B}(\sigma_{i_B}, Q_{i_B})$ and agent i_A has the same belief about the world as agent i_B, in other words, for every j, $Pr(\hat{r}_{j_A} = \hat{r} | \sigma_{i_A}, \mathbf{Q}_A, \mathbf{s}_A) = Pr(\hat{r}_{j_B} = \hat{r} | \sigma_{i_B}, \mathbf{Q}_B, \mathbf{s}_B)$ $\forall \hat{r} \in \mathcal{R}$.

Now we will prove two properties of indistinguishable scenarios which are the main tools in the proof for our impossibility result.

Observation 55. If $(\mathbf{Q}_A, \mathbf{s}_A) \approx (\mathbf{Q}_B, \mathbf{s}_B)$, then (i) for any mechanism $\mathcal{M}$, $\mathbf{s}_A$ is a

(strict) equilibrium for the prior profile $\mathbf{Q}_A$ iff $\mathbf{s}_B$ is a (strict) equilibrium for the prior profile $\mathbf{Q}_B$. (ii) $\forall i$, $\nu^{\mathcal{M}}_{i_A}(n, \Sigma, A) = \nu^{\mathcal{M}}_{i_B}(n, \Sigma, B)$

At a high level, (1) is true since any reported profile distribution that agent i_A can deviate to, agent i_B can deviate to the same reported profile distribution as well and obtain the same expected payment as agent i_A.

Formally, we will prove the $\Rightarrow$ direction in (1) by contradiction. The proof of the other direction will be similar. Consider the coupling for σ_A, σ_B mentioned in the definition of indistinguishable scenarios. For the sake of contradiction, assume there exists i and σ_{i_B} such that $\hat{r}' \neq s_{i_B}(\sigma_{i_B}, Q_{i_A})$ is a best response for agent i_B. Since agent i_A has the same belief about the world as agent i_B and $s_{i_A}(\sigma_{i_A}, Q_{i_A}) = s_{i_B}(\sigma_{i_B}, Q_{i_B})$, $\hat{r}' \neq s_{i_A}(\sigma_{i_A}, Q_{i_A})$ is a best response to agent i_A as well, which is a contradiction to the fact that $\mathbf{s}_A$ is a strictly equilibrium for prior Q_{i_A}.

To gain intuition about (2), consider the coupling again. For any i, agent i_A reports the same thing and has the same belief for the world as agent i_B, which implies the expected payoff of agent i_A is the same as agent i_B. (2) follows.

Now we are ready to prove our impossibility result:

of Proposition 51. We prove part (1) and part (2) separately.

Proof of Part (1) Let $A := (\mathbf{Q}, \mathbf{s}), B := (\boldsymbol{\pi}^{-1}(\mathbf{Q}), \boldsymbol{\pi}(\mathbf{s}))$. We will show that for any strategy profile $\mathbf{s}$ and any prior Q, $A \approx B$. Based on our above observations, part (1) immediately follows from that fact.

To prove $(Q, \mathbf{s}) \approx (\boldsymbol{\pi}^{-1}Q, \boldsymbol{\pi}(\mathbf{s}))$, for every i, we can couple $(\sigma_1, \sigma_2, ..., \sigma_n)$ with $(\pi_1^{-1}(\sigma_1), \pi_2^{-1}(\sigma_2), .., \pi_n^{-1}(\sigma_n))$ where $(\sigma_1, \sigma_2, ..., \sigma_n)$ is drawn from Q_i. It is a legal coupling since

$$\boldsymbol{\pi}^{-1}(Q_i)(\pi_1^{-1}(\sigma_1), \pi_2^{-1}(\sigma_2), .., \pi_n^{-1}(\sigma_n)) = Q_i(\sigma_1, \sigma_2, ..., \sigma_n)$$

according to the definition of $\boldsymbol{\pi}^{-1}(Q)$.

Now we show this coupling satisfies the condition in Definition 54. First note that $\boldsymbol{\pi}(s_i)(\boldsymbol{\pi}^{-1}(\sigma_i), \boldsymbol{\pi}^{-1}(Q)) = s_i(\sigma_i, Q)$. Now we begin to calculate $Pr(\hat{r}_{j_B} = \hat{r}|\sigma_{i_B}, \mathbf{Q}_B, \mathbf{s}_B)$

$$
\begin{aligned}
Pr(\hat{r}_{j_B} = \hat{r}|\sigma_{i_B}, \mathbf{Q}_B, \mathbf{s}_B) =& Pr(\hat{r}_{j_B} = \hat{r}|\pi_i^{-1}(\sigma_{i_A}), \boldsymbol{\pi}^{-1}(Q_{j_A}), \boldsymbol{\pi}(\mathbf{s}_A)) && (2.10)\\
=& \sum_{\sigma'} \boldsymbol{\pi}^{-1}(Q_{i_A})(\sigma'|\pi_i^{-1}(\sigma_{i_A}))Pr(\boldsymbol{\pi}(s_{j_A})(\sigma', \boldsymbol{\pi}^{-1}(Q_{j_A})) = \hat{r}) && (2.11)\\
=& \sum_{\sigma'} \boldsymbol{\pi}^{-1}(Q_{i_A})(\sigma'|\pi_i^{-1}(\sigma_{i_A}))Pr(s_{j_A}(\boldsymbol{\pi}(\sigma'), \boldsymbol{\pi}\boldsymbol{\pi}^{-1}(Q_{j_A})) = \hat{r}) && (2.12)\\
=& \sum_{\sigma'} Q_{i_A}(\pi_j(\sigma')|\sigma_{i_A})Pr(s_{j_A}(\boldsymbol{\pi}(\sigma'), Q_{j_A}) = \hat{r}) && (2.13)\\
=& \sum_{\sigma''} Q_{i_A}(\sigma''|\sigma_{i_A})Pr(s_{j_A}(\sigma'', Q_{i_A}) = \hat{r}) && (2.14)\\
=& Pr(\hat{r}_{j_A} = \hat{r}|\sigma_{i_A}, \mathbf{Q}_A, s_A) && (2.15)
\end{aligned}
$$

From (2.10) to (2.11): To calculate the probability that agent j_B has reported $\hat{r}$, we should sum over all possible private signals agent j_B has received and calculate the probability agent j_B reported $\hat{r}$ conditioning on he received private signal σ', which is determined by agent j_B's strategy $\boldsymbol{\pi}(s_{j_A})$.

By abusing notation a little bit, we can write $\boldsymbol{\pi}(s_{j_A})(\sigma', \boldsymbol{\pi}^{-1}Q_{j_A})$ as a random variable (it is actually a distribution) with $Pr(\boldsymbol{\pi}(s_{j_A})(\sigma', \boldsymbol{\pi}^{-1}Q) = \hat{r}) = \boldsymbol{\pi}(s_{j_A})(\sigma', \boldsymbol{\pi}^{-1}Q)(\hat{r})$. According to above explanation, (2.11) follows.

(2.12) follows from the definition of permuted strategy.

(2.13) follows from the definition of permuted prior.

By replacing $\pi_j(\sigma')$ by σ'', (2.14) follows.

We finished the proof $A \approx B$, as previously argued, result (1) follows.

Proof for Part (2) We will prove the second part by contradiction:

Fix permutation strategy profile $\boldsymbol{\pi}$. First notice that there exists a positive integer O_d such that $\boldsymbol{\pi}^{O_d} = I$ where I is the identity and agents play I means they tell the truth.

Given any strategy profile s, for the sake of contradiction, we assume that there exists a mechanism $\mathcal{M}$ with unknown prior profile such that $\nu_{i_A}^{\mathcal{M}}(n, \Sigma, \mathbf{Q}, \mathbf{s}) > \nu_{i_A}^{\mathcal{M}}(n, \Sigma, \mathbf{Q}, \boldsymbol{\pi}(\mathbf{s}))$ for any prior Q. For positive integer $k \in \{0, 1, ..., O_d\}$, we construct three scenarios:

$$A_k := (\boldsymbol{\pi}^k(\mathbf{Q}), \mathbf{s}),\ A_{k+1} := (\boldsymbol{\pi}^{k+1}(\mathbf{Q}), \mathbf{s}),\ B_k := (\boldsymbol{\pi}^k(\mathbf{Q}),\ \boldsymbol{\pi}(\mathbf{s}))$$

and show for any k,

(I)$\nu_{i_A}^{\mathcal{M}}(n, \Sigma, A_k) > \nu_{i_A}^{\mathcal{M}}(n, \Sigma, B_k)$,

(II) $\nu_{i_A}^{\mathcal{M}}(n, \Sigma, A_{k+1}) = \nu_{i_A}^{\mathcal{M}}(n, \Sigma, B_k)$.

Combining (I), (II) and the fact $A_0 = A_{O_d}$, we have

$$\nu_{i_A}^{\mathcal{M}}(n, \Sigma, A_0) > \nu_{i_A}^{\mathcal{M}}(n, \Sigma, A_1) > ...\nu_{i_A}^{\mathcal{M}}(n, \Sigma, A_{O_d}) = \nu_{i_A}^{\mathcal{M}}(n, \Sigma, A_0)$$

which is a contradiction.

Now it is only left to show (I) and (II). Based on our assumption

$$\nu_{i_A}^{\mathcal{M}}(n, \Sigma, \mathbf{Q}, \mathbf{s}) > \nu_{i_A}^{\mathcal{M}}(n, \Sigma, \mathbf{Q}, \boldsymbol{\pi}(\mathbf{s}))$$

for any prior $\mathbf{Q}$, we have (I). By the same proof we have in part (1), we have $A_{k+1} \approx B_k$, which implies (II) according to our above observations.

When the mechanism knows the prior is symmetric, the above proof is still valid if we replace the permutation list $\boldsymbol{\pi}$ by symmetric permutation list $(\pi, \pi, ..., \pi)$. $\square$

CHAPTER III

Multi-task Signal Elicitation

3.1 Related work

Since Miller, Resnick, and Zeckhauser [45] introduced peer prediction, several works follow the peer prediction framework and design information elicitation mechanisms without verification in different settings. In this section, we introduce these works in multi-task, detail free, minimal setting.

Dasgupta and Ghosh [18] consider a setting where agents are asked to answer multiple a priori similar binary choice questions. They propose a mechanism M_d that pays each agent the correlation between her answer and her peer's answer, and show each agent obtains the highest payment if everyone tells the truth. In retrospect, one can see that our techniques are a recasting and generalization of those of Dasgupta and Ghosh [18]. Kamble et al. [31] considers both homogeneous and heterogeneous populations and design a mechanism such that truth-telling pays higher than non-informative equilibria in the presence of a large number of a priori similar questions. However, they leave the analysis of other non-truthful equilibria as a open question. Agarwal et al. [2] consider a peer prediction mechanism for heterogeneous users.

3.1.1 Independent work

Like this thesis, Shnayder et al. [63] also extends Dasgupta and Ghosh [18]'s binary signals mechanism to multiple signals setting. However, the two works differ both in the specific mechanism and the technical tools employed.

Shnayder et al. [63] analyze how many questions are needed (whereas we simply assume infinitely many questions). Like this thesis, they also analyze to what extent truth-telling can pay strictly more than other equilibria. Additionally, they show their mechanism does not need a large number of questions when "the signal correlation structure" is known (that is the pair-wise correlation between the answers of two questions). While the this thesis does not state such results, we note that the techniques employed are sufficiently powerful to immediately extend to this interesting special case (Section 3.5)—when the signal structure is known, it is possible to construct an unbiased estimator for f-mutual information of the distribution, when the total variation distance is used to define the f-mutual information. Both Shnayder et al. [63] and this thesis also show their results generalize Dasgupta and Ghosh [18]'s.

Moreover, when the number of questions is large, f-mutual information mechanism has truth-telling as a dominant strategy while Shnayder et al. [63] do not.

3.2 Background and assumptions

In this section, we introduce the multi-task setting which was previously studied in Dasgupta and Ghosh [18] and Radanovic and Faltings [55]: n agents are assigned the same T questions (multi-tasks). For each question k, each agent i receives a **private signal** $\sigma_i^k \in \Sigma$ about question k and is asked to report this signal. We call this setting (n, T, Σ).

We see mechanisms in which agents are not required to report their forecasts for

other agents' answer (minimal), and were the mechanism does not know the agents' priors (detail free). Agent i may lie and report $\hat{\sigma}_i^k \neq \sigma_i^k$. Dasgupta and Ghosh [18] give the following example for this setting: n workers are asked to check the quality of m goods, they may receive signal "high quality" or "low quality".

Agents have priors for questions. Each agent i believes agents' private signals for question k are chosen from a joint distribution Q_i^k over Σ^n. Note that different agents may have different priors for the same question.

In the multi-task setting, people usually make the following assumption:

Assumption 56 (A Priori Similar and Random Order)**.** *For any i, any $k \neq k'$, $Q_i^k = Q_i^{k'}$. Moreover, all tasks/questions appear in a random order, independently drawn for each agent.*

This means agents cannot distinguish each question without the private signal they receive.

We define $(\Psi_1, \Psi_2, ..., \Psi_n)$ as the joint random variables such that

$$\Pr(\Psi_1 = \sigma_1, \Psi_2 = \sigma_2, ..., \Psi_n = \sigma_n)$$

equals the probability that agents $1, 2, .., n$ receive private signals $(\sigma_1, \sigma_2, ..., \sigma_n)$ correspondingly for a question which is picked uniformly at random.

We define $(\hat{\Psi}_1, \hat{\Psi}_2, ..., \hat{\Psi}_n)$ as the joint random variables such that

$$\Pr(\hat{\Psi}_1 = \hat{\sigma}_1, \hat{\Psi}_2 = \hat{\sigma}_2, ..., \hat{\Psi}_n = \hat{\sigma}_n)$$

equals the probability that agents $1, 2, .., n$ reports signals $(\hat{\sigma}_1, \hat{\sigma}_2, ..., \hat{\sigma}_n)$ correspondingly a question which is picked uniformly at random. Note that the joint distribution over $(\hat{\Psi}_1, \hat{\Psi}_2, ..., \hat{\Psi}_n)$ depends on the strategies agents play.

For each question k, each agent i's effort strategy is λ_i^k and conditioning on that she

invests full effort e_i, her strategy is s_i^k. We say agent i plays a **consistent strategy** if for any k, k', $\lambda_i^k = \lambda_i^{k'}$ and $s_i^k = s_i^{k'}$.

Recall that in the minimal mechanism, the strategy corresponds to a transition matrix. We define **truth-telling T** as the strategy where an agent truthfully reports her private signal for every question. **T** corresponds to the identity matrix. We say agent i plays a **permutation** strategy if there exists a permutation transition matrix π such that $s_i^k = \pi, \forall k$. Note that a permutation strategy is a consistent strategy. We define a **consistent strategy profile** as the strategy profile where all agents play a consistent strategy.

With the a priori similar and random order assumption, Dasgupta and Ghosh [18] make the following observation:

Observation 57. [18] When questions are a priori similar and agents receive questions in random order (Assumption 56), for every agent, using different strategies for different questions is the same as a mixed consistent strategy.

With the above observation, it is sufficient to only consider the consistent strategy profiles.

3.3 The f-mutual information mechanism and Bregman mutual information mechanism

In this section, we give direct applications of the Mutual Information Paradigm in multi-task setting—the f-mutual information mechanism and the Bregman mutual information mechanism. Both of them are a family of mechanisms that can be applied to the non-binary setting / multiple-choices questions which generalize the mechanism in Dasgupta and Ghosh [18] that can only be applied to the binary setting / binary choices questions. Moreover, both the f-mutual information mechanism and the Bregman mutual information mechanism are dominantly truthful without considering

efforts. Later we will map the mechanism in Dasgupta and Ghosh [18] to a special case of the f-mutual information mechanism[1].

f-mutual Information Mechanism $\mathcal{M}_{MI^f}$ Given a multi-task setting (n, T, Σ),

Report For each agent i, for each question k, she is asked to provide her private signal σ_i^k. We denote the actual answer she reports as $\hat{\sigma}_i^k$.

Payment/Information Score We arbitrarily pick a reference agent $j \neq i$. We define a probability measure P over $\Sigma \times \Sigma$ such that $T * P(\hat{\Psi}_i = \sigma_i; \hat{\Psi}_j = \sigma_j)$ equals the number of questions that agent i answers σ_i and agent j answers σ_j.

Agent i is paid by her information score

$$MI^f(\hat{\Psi}_i; \hat{\Psi}_j)$$

where $(\hat{\Psi}_i; \hat{\Psi}_j)$ draws from the probability measure P.

Theorem 58. *Given a multi-task setting (n, T, Σ) with the a priori similar and random order assumption (56), when the number of questions is infinite, f is (strictly) convex (and every agent's prior is fine-grained), $\mathcal{M}_{MI^f}$ is detail free, minimal, (strictly) dominantly truthful, (strictly) truth-monotone.*

Proof. We would like to show that the f-mutual information mechanism is the same as MIP(MI^f). Then Corollary 34 directly implies the theorem.

Based on observation 57, it is sufficient to only consider the consistent strategy profiles. When the number of questions is infinite and $\forall i$, agent i play the consistent strategy,

$$P(\hat{\Psi}_i = \sigma_i; \hat{\Psi}_j = \sigma_j) = \Pr(\hat{\Psi}_i = \sigma_i; \hat{\Psi}_j = \sigma_j).$$

[1]Although f-mutual information mechanism requires infinite number of question for clean analysis, with an extra positively correlated assumption for the information structure, we can construct an unbiased estimator for f-mutual information of the distribution via only 3 questions (See Section 3.1.1, Appendix 3.5).

Therefore, with Assumption 56, when the number of questions is infinite, the f-mutual information mechanism is the same as MIP(MI^f) in the multi-task setting. Theorem 58 follows immediately from Corollary 34.

□

Bregman mutual Information Mechanism $\mathcal{M}_{BMI^{PS}}$ We can define *Bregman mutual information mechanism* via the same definition of f-mutual information except replacing MI^f by BMI^{PS}.

Corollary 36 directly imply the following theorem.

Theorem 59. *Given a multi-task setting (n, T, Σ) with the a priori similar and random order assumption (56), when the number of questions is infinite, without considering efforts, the Bregman mutual information mechanism $\mathcal{M}_{BMI^{PS}}$ is detail free, minimal, dominantly truthful.*

3.4 Mapping Dasgupta and Ghosh [2013] into our information theoretic framework

This section maps Dasgupta and Ghosh [18] to a special case of f-mutual information mechanism—TVD-mutual information mechanism $\mathcal{M}_{MI^{tvd}}$ (restricted to the binary choice setting)—using the specific f-divergence, total variation distance. With the mapping, we can simplify the proof in Dasgupta and Ghosh [18] to a direct application of our framework.

3.4.0.1 Prior Work

We first state the mechanism M_d and the main theorem in Dasgupta and Ghosh [18].

Mechanism M_d Agents are asked to report binary signals 0 or 1 for each question. Uniformly randomly pick a reference agent j for agent i. We denote C_i as the set of questions agent i answered. We denote C_j as the set of questions agent j answered. We denote $C_{i,j}$ as the set of questions both agent i and agent j answered. For each question $k \in C_{i,j}$ that both agent i and agent j answered, pick subsets $A \subseteq C_i \backslash k, B \subseteq C_j \backslash (k \cup A)$ with $|A| = |B| = d$. If such A, B do not exist, agent i's reward is 0. Otherwise, we define $\bar{\hat{\sigma}}_i^A = \frac{\sum_{l \in A} \hat{\sigma}_i^l}{|A|}$ to be agent i's average answer for subset A, $\bar{\hat{\sigma}}_j^B = \frac{\sum_{l \in B} \hat{\sigma}_j^l}{|B|}$ is agent j's average answer for subset B.

Agent i's reward for each question $k \in C_{i,j}$ is

$$R_{i,j}^k := [\hat{\sigma}_i^k * \hat{\sigma}_j^k + (1 - \hat{\sigma}_i^k) * (1 - \hat{\sigma}_j^k)] - [\bar{\hat{\sigma}}_i^A * \bar{\hat{\sigma}}_j^B + (1 - \bar{\hat{\sigma}}_i^A) * (1 - \bar{\hat{\sigma}}_j^B)]$$

By simple calculations, essentially agent i's reward for each question $k \in C_{i,j}$ is the correlation between her answer and agent j's answer—$\mathbb{E}[\hat{\Psi}_i \hat{\Psi}_j] - \mathbb{E}[\hat{\Psi}_i]\mathbb{E}[\hat{\Psi}_j]$.

Dasgupta and Ghosh [18] also make an additional assumption:

Assumption 60 (Positively Correlated)**.** *Each question k has a unknown ground truth a^k and for every agent i, with probability greater or equal to $\frac{1}{2}$, agent i receives private signal a^k.*

We succinctly interpret the main results of Dasgupta and Ghosh [18] as well as the results implied by the main results into the following theorem.

Theorem 61. *[18] Given an multi-question setting (n, T, Σ) with the a priori similar and random order assumption (56), the positively correlated assumption (89), when $T \geq d + 1$, M_d is truthful, and strongly focal.*

The parameter d can be any positive integer. Larger d will make the mechanism more robust. We will see M_d equals a special case of the f-mutual information mechanism only if agent i, j's reported answers are positively correlated. Thus, without considering efforts, M_d is not dominantly truthful while the f-mutual information

mechanism is. Although M_d only requires a small number of questions, it only applies to binary choice questions, makes an extra assumption, and obtains weaker properties than the f-mutual information mechanism.

3.4.0.2 Using our information theoretic framework to analyze Dasgupta and Ghosh [18]

Proof Outline We will first connect the expected payment in M_d with a specific f-mutual information—MI^{tvd}. Then the result follows from the information monotone property of f-mutual information. Formally, we use the following claim to show the connection between mechanism M_d and f-mutual information mechanism.

Claim 62. $[M_d \approx \mathcal{M}_{MI^{tvd}}]$ With a priori similar and random order assumption, in M_d, for every pairs i, j, for every reward question k,

$$\mathbb{E}[R^k_{i,j}] = \frac{1}{2} MI^{tvd}(\Psi_i; \Psi_j)$$

if both of them play $\mathbf{T}$;

$$\mathbb{E}[R^k_{i,j}] \leq \frac{1}{2} MI^{tvd}(\hat{\Psi}_i; \hat{\Psi}_j)$$

if one of them does not play $\mathbf{T}$.

Claim 141 shows the connection between M_d and $\mathcal{M}_{MI^{tvd}}$. The only difference between M_d and $\mathcal{M}_{MI^{tvd}}$ is that for agents i, j, when one of the agent does not play $\mathbf{T}$, the correlation between their reports is upper-bounded by rather than equal to the *tvd*-mutual information. Therefore, in M_d, truth-telling is not a dominant strategy. But the information-monotone property of MI^{tvd} still guarantees the informative truthful and strongly focal property of M_d.

Proof of Theorem 61. We start to show the truthful property of M_d.

For every agent i, given that everyone else plays $\mathbf{T}$, agent i's expected payment for each reward question is

$$\mathbb{E}[R^k_{i,j}] \leq \frac{1}{2} MI^{tvd}(\hat{\Psi}_i; \Psi_j) \leq \frac{1}{2} MI^{tvd}(\Psi_i; \Psi_j)$$

since MI^{tvd} is information-monotone. Thus, M_d is truthful. Moreover,

$$\mathbb{E}[R^k_{i,j}] \leq \frac{1}{2} MI^{tvd}(\hat{\Psi}_i; \hat{\Psi}_j) \leq \frac{1}{2} MI^{tvd}(\Psi_i; \Psi_j)$$

Thus, the truth-telling strategy profile maximizes *every* agent's expected payment among all strategy profiles which implies M_d is strongly focal.

□

Proof for Claim 141 We first show that

$$\mathbb{E}[R^k_{i,j}] = \frac{1}{2} MI^{tvd}(\Psi_i; \Psi_j)$$

if both of agents i, j play $\mathbf{T}$.

Note that by simple calculations, Assumption 89 implies that for any $\sigma \in \{0, 1\}$,

$$\Pr[\Psi_j = \sigma | \Psi_i = \sigma] \geq \Pr[\Psi_j = \sigma],$$

$$\Pr[\Psi_j = \sigma | \Psi_i = \sigma'] \leq \Pr[\Psi_j = \sigma], \forall \sigma' \neq \sigma.$$

When both of agents i, j play $\mathbf{T}$,

$$\frac{1}{2}MI^{tvd}(\Psi_i;\Psi_j) = \frac{1}{2}\sum_{\sigma,\sigma'} |\Pr[\Psi_i=\sigma,\Psi_j=\sigma'] - \Pr[\Psi_i=\sigma]\Pr[\Psi_j=\sigma']| \quad \text{(Definition of } MI^{tvd}\text{)}$$

$$= \frac{1}{2}\sum_{\sigma,\sigma'} \mathbb{1}(\sigma=\sigma')\left(\Pr[\Psi_i=\sigma,\Psi_j=\sigma'] - \Pr[\Psi_i=\sigma]\Pr[\Psi_j=\sigma']\right)$$

$$+ \mathbb{1}(\sigma\neq\sigma')\left(\Pr[\Psi_i=\sigma]\Pr[\Psi_j=\sigma'] - \Pr[\Psi_i=\sigma,\Psi_j=\sigma']\right) \quad \text{(Assumption 89)}$$

$$= \sum_{\sigma}\left(\Pr[\Psi_i=\sigma,\Psi_j=\sigma] - \Pr[\Psi_i=\sigma]\Pr[\Psi_j=\sigma]\right)$$

$$\text{(Combining like terms, } \Pr[E]-\Pr[\neg E] = 2\Pr[E]-1\text{)}$$

$$= \mathbb{E}[R^k_{i,j}] \quad \text{(Definition of } R^k_{i,j} \text{ in } M_d\text{)}$$

The proof of

$$\mathbb{E}[R^k_{i,j}] \leq \frac{1}{2}MI^{tvd}(\hat{\Psi}_i;\hat{\Psi}_j)$$

is similar to above proof. We only need to replace Ψ_i by $\hat{\Psi}_i$ and change the second equation to greater than, that is,

$$\frac{1}{2}\sum_{\sigma,\sigma'} |\Pr[\hat{\Psi}_i=\sigma,\hat{\Psi}_j=\sigma'] - \Pr[\hat{\Psi}_i=\sigma]\Pr[\hat{\Psi}_j=\sigma']|$$

$$\geq \frac{1}{2}\sum_{\sigma,\sigma'} \mathbb{1}(\sigma=\sigma')\Big(\Pr[\hat{\Psi}_i=\sigma,\hat{\Psi}_j=\sigma'] - \Pr[\hat{\Psi}_i=\sigma]\Pr[\hat{\Psi}_j=\sigma']\Big)$$

$$+ \mathbb{1}(\sigma\neq\sigma')\Big(\Pr[\hat{\Psi}_i=\sigma]\Pr[\hat{\Psi}_j=\sigma'] - \Pr[\hat{\Psi}_i=\sigma,\hat{\Psi}_j=\sigma']\Big). \quad \left(\sum |x| \geq \sum x\right)$$

We have finished the proof of Claim 141.

3.5 Independent work analysis

The analysis in Section 3.4 is not restricted to Dasgupta and Ghosh [18]. Replacing the $R^k_{i,j}$ defined in M_d by the $R^k_{i,j}$ defined in the non-binary extension of M_d in the independent work of Shnayder et al. [63] will not change the analysis. Thus, Shnayder et al. [63] is also a special case of f-mutual information mechanism—TVD-mutual information mechanism $\mathcal{M}_{MI^{tvd}}$ in the non-binary settings.

CA applies to a more general setting than the positively correlated setting (Assumption 89), in the sense that CA assumes the knowledge of signal structure but the signal structure does not need to be positively correlated. Here we give the analysis for CA in the special setting where the signal structure is positively correlated. The analysis for other settings is similar.

The Correlated Agreement (CA) Mechanism [63] In the special setting where the signal structure is positively correlated, the non-binary extension of M_d—the CA mechanism—can be reinterpreted as M_d by defining

$$R^k_{i,j} := \mathbb{1}(\hat{\sigma}^k_i = \hat{\sigma}^k_j) - \mathbb{1}(\hat{\sigma}^{\ell_A}_i = \hat{\sigma}^{\ell_B}_j)$$

where ℓ_A is picked from subset A uniformly at random and ℓ_B is picked from subset B uniformly at random.

With this new definition of $R^k_{i,j}$, Claim 141 is still valid since the proof of Claim 141 that uses the definition of $R^k_{i,j}$—

$$\begin{aligned}&\sum_{\sigma} \left(\Pr[\Psi_i = \sigma, \Psi_j = \sigma] - \Pr[\Psi_i = \sigma]\Pr[\Psi_j = \sigma]\right)\\ &= \mathbb{E}[R^k_{i,j}] \qquad \text{(Definition of } R^k_{i,j} \text{ in } M_d)\end{aligned}$$

—is still valid for this new definition of $R^k_{i,j}$. Therefore, Theorem 61 is still valid when

we replace M_d by the CA mechanism which means we can also use our information theoretic framework to analyze Shnayder et al. [63].

CHAPTER IV

Single-task Signal Elicitation

4.1 Related work

After Miller, Resnick, and Zeckhauser [45] introducing peer prediction, a host of results (see, e.g., [51, 67, 53, 55, 74, 59, 21, 71, 70, 69, 28, 29, 36]) have followed. In this section, we will introduce them and classifies them into several categories according to the properties they (do not) have.

(1) Single-task, detail free, focal (not small group): Bayesian Truth Serum (BTS) [51] first successfully weakened the known common prior assumption (detail free) and addresses the equilibrium multiplicity issue (focal). Prelec [51] also provides an important framework for mechanisms without known common prior. BTS requires the agents report—in addition to their reported signal—a forecast (prediction) of the other agents' reported signals, and uses this predictions in lieu of the common prior. BTS incentives agents to report accurate forecasts by rewarding forecasts that have the ability to predict the other agents' reported signal. However, BTS has two weakness: (1) BTS requires that the number of agents goes to infinity (or is large enough in a modified version) since the mechanism needs agents to believe it has access to the true distribution of from which agents' signals are drawn. (2) The analysis of non-truthful equilibria provided in [51] requires that the number of agents goes to infinity and only proves that truth-telling has total expected payment at least as high

as other equilibrium. Specifically, it does not rule out the existence of many other equilibrium which are all paid the same as the truth-telling equilibrium. *Logarithmic Peer Truth Serum (PTS)* [54] extends BTS to a slightly different setting involving sensors, but still requires a large number of agents.

(2) Single-task, small group, detail free (not focal): Several mechanisms [67, 53, 55, 59, 71, 70, 69] are based on the BTS framework and address the first weakness of BTS. *Robust Bayesian Truth Serum (RBTS)* [67] is a mechanism which can only be applied to binary signals. *Multi-Valued RBTS* [53] and *Multi-Signal Shadowing Method (Multi-Signal SM)* [68] can be applied to non-binary signals while they require an *additional assumption* that an agent will think the probability that other agents receive signal σ higher if he himself also receives σ. *Divergence-based BTS* [55] can be applied to non-binary signals and does not require additional assumptions on the prior. All of those works do not address the equilibrium multiplicity issue, but do work for a small number of agents. *Minimal Truth Serum (MTS)* [59] is a mechanism where agents have the option to report or not report their predictions, and also lacks analysis of non-truthful equilibria. MTS uses a typical zero-sum technique such that all equilibria are paid equally. In contrast, we show that in our *Disagreement Mechanism* any equilibrium that is even close to paying more than the truth-telling equilibrium must be close to a small set of permutation equilibrium. The *Divergence based BTS* only requires the common prior assumption to be truthful. Because of its generality, we use it as a building block in our Disagreement Mechanism. However, the *Divergence based BTS* contains effortless equilibrium that pay significantly more than truth-telling. Moreover, analysing the set of equilibria in *Divergence-based BTS* is very complicated and becomes a main technical obstacle in this chapter. Thus, while the above work addresses the first weakness of BTS, it exacerbates the second.

(3) Single-task, small group, focal (not detail free): Jurca and Faltings [28, 29] use algorithmic mechanism design to build their own peer prediction style mechanism

where truth-telling is paid strictly better than non-truthful *pure* strategies but leaves the analysis of mixed strategies as a open question. Kong, Schoenebeck, and Ligett [36] modify the peer prediction mechanism such that truth-telling is paid strictly better than any other non-truthful equilibrium. Additionally, they optimize the cost their mechanism needs over a natural space. The assumption that the mechanism knows the prior, allows these mechanisms to only require that agent's report a signal (there is no prediction report). However, unlike the current work, the mechanism still needs to know the prior and the analysis only works for the case of binary signals.

(4) Different Settings: We have introduced multi-task setting in the previous chapter. In addition to the multi-task setting, there are many other works in the settings that are different from our results. For example, Cai, Daskalakis, and Papadimitriou [11] and Liu and Chen [41] consider the machine learning setting. Kamble et al. [31], and Agarwal et al. [2] consider the heterogeneous participants setting in the multi-task setting. Mandal et al. [42] consider the heterogeneous tasks setting. Zhang and Chen [74] consider a sequential game. Faltings et al. [21] consider a setting where they have an estimation of the public distribution of previous answers on other a priori similar questions.

4.2 Preliminary and background

We will defer the proofs for most claims to Appendix A.1.4.

4.2.1 Prior definitions and assumptions

We consider a setting with n agents and a set of signals Σ, and define a *setting* as a tuple (n, Σ). Each agent i has a private signal $\sigma_i \in \Sigma$ chosen from a joint distribution Q over Σ^n called the prior. Given a prior Q, for $\sigma \in \Sigma$, let $q_i(\sigma) = \Pr_Q[\sigma_i = \sigma]$ be the *a priori* probability that agent i receives signal σ. Let $q_{j,i}(\sigma'|\sigma) = \Pr_Q[\sigma_j = \sigma'|\sigma_i = \sigma]$ be the probability that agent j receives signal σ given that agent i received signal σ'.

We say that a prior Q over Σ is *symmetric* if for all σ, $\sigma' \in \Sigma$ and for all pairs of agents $i \neq j$ and $i' \neq j'$ we have $q_i(\sigma) = q_{i'}(\sigma)$ and $q_{i,j}(\sigma|\sigma') = q_{i',j'}(\sigma|\sigma')$. That is, the first two moments of the prior do not depend on the agent identities.

Assumption 63 (Symmetric Prior)**.** *We assume throughout that the agents' signals* $\boldsymbol{\sigma}$ *are drawn from some joint* **symmetric prior** Q*.*

Because we will assume that the prior is symmetric, we denote $q_i(\sigma)$ by $q(\sigma)$ and $q_{i,j}(\sigma|\sigma')$ (where $i \neq j$) by $q(\sigma|\sigma')$. We also define $\mathbf{q}_\sigma = q(\cdot|\sigma)$.

Assumption 64 (Non-zero Prior)**.** *We assume that for any* $\sigma, \sigma' \in \Sigma$, $q(\sigma) > 0, q(\sigma|\sigma') > 0$.

Assumption 65 (Informative Prior)**.** *We assume if agents have different private signals, they will have different expectations for the fraction of at least one signal. That is for any* $\sigma \neq \sigma'$*, there exists* σ'' *such that* $q(\sigma''|\sigma) \neq q(\sigma''|\sigma')$*.*

The following assumption conceptually states that one state is not just a more likely version of another state, and can be thought of as a weaker version of assuming $q(\sigma|\cdot)$ are linearly independent.

Assumption 66 (Fine-grained Prior)**.** *We assume that for any* $\sigma \neq \sigma' \in \Sigma$*, there exists* σ'', σ''' *such that*

$$\frac{q(\sigma|\sigma'')}{q(\sigma'|\sigma'')} \neq \frac{q(\sigma|\sigma''')}{q(\sigma'|\sigma''')}$$

If this assumption does not hold, then in some since σ and σ' are the same signal. We can create a new prior by replacing σ and σ' with a new signal $\sigma_0 := \sigma$ or σ', and not lose any information, in the sense that we can still recover the original prior. To see this, we first define $p = \frac{q(\sigma)}{q(\sigma')}$, and note that for all σ'', $p = \frac{q(\sigma|\sigma'')}{q(\sigma'|\sigma'')}$. Whenever σ_0 is drawn in the new prior, we simply replace it by σ with probability p and σ' with probability $1 - p$. This produces the same prior for agents that have no information or other their signal's information.

We illustrate this in the following example:

Example 67. $Q = \begin{pmatrix} q(s_1|s_1) & q(s_1|s_2) & q(s_1|s_3) \\ q(s_2|s_1) & q(s_2|s_2) & q(s_2|s_3) \\ q(s_3|s_1) & q(s_3|s_2) & q(s_3|s_3) \end{pmatrix} = \begin{pmatrix} 0.1 & 0.2 & 0.3 \\ 0.2 & 0.4 & 0.6 \\ 0.7 & 0.4 & 0.1 \end{pmatrix}$ is not a fine-grained prior since

$$\frac{q(s_1|s_1)}{q(s_2|s_1)} = \frac{q(s_1|s_2)}{q(s_2|s_2)} = \frac{q(s_1|s_3)}{q(s_2|s_3)}$$

Note that in this example, even we combine s_1 and s_2 to be a single signal s_0 which is defined as $s_0 := s_1$ or s_2, we do not lose any information: if an agent knows that the fraction of agents who report s_0 is x, we know his belief for the expectation of the fraction of s_1 must be $\frac{x}{3}$ no matter what private signal he receives.

We only require the fine-grained prior assumption to show that truth-telling is *strictly* "better" than any other symmetric equilibrium (excluding permutation equilibrium). In the above example where the prior is not fine-grained, if agents always report s_1 when they receive s_1 or s_2, this does not lose information (is not "worse") comparing with the case agents always tell the truth. So we cannot say truth-telling is strictly "better" than any other equilibrium when the prior is not fine-grained. However, this assumption is not necessary to show that truth-telling is a strict Bayesian equilibrium of our mechanism, nor to show that the agent welfare of truth-telling is at least as high as other symmetric equilibrium.

Assumption 68 (Ensemble Prior)**.** *Although we talk of a single prior, in fact we have an ensemble $Q = \{Q_n\}_{n\in N, n\geq 3}$ of priors; one for each possible number of agents greater than 3. We assume that all Q_n are over the same signal set Σ have have identical $q(\sigma)$ and $q(\sigma'|\sigma)$.*

When the number of agents n changes, the joint prior actually changes as well, but the first two moments of the prior are fixed. This allows us to make meaningful

statements about n going to infinity.

We sometimes will denote the class of priors that satisfy all five of these assumptions as SNIFE priors.

4.2.2 Game setting and equilibrium concepts

Given a setting (n, Σ) with prior Q, we consider a game in which each agent i is asked to report his private signal $\sigma_i \in \Sigma$ and his prediction $\mathbf{p}_i \in \Delta_\Sigma$, a distribution over Σ, where $\mathbf{p}_i = \mathbf{q}_{\sigma_i}$. For any $\sigma \in \Sigma$, $\mathbf{p_i}(\sigma)$ is agent i's (reported) expectation for the fraction of other agents who has received σ given he has received σ_i. However, agents may not tell the truth. We denote $\Sigma \times \Delta_\Sigma$ by $\mathcal{R}$. We define a report profile of agent i as $r_i = (\hat{\sigma}_i, \hat{\mathbf{p}}_i) \in \mathcal{R}$ where $\hat{\sigma}_i$ is agent i's reported signal and $\hat{\mathbf{p}}_i$ is agent i's reported prediction.

We would like to encourage truth-telling, namely that agent i reports $\hat{\sigma}_i = \sigma_i, \hat{\mathbf{p}}_i = \mathbf{q}_{\sigma_i}$. To this end, agent i will receive some payment $\nu_i(\hat{\sigma}_i, \hat{\mathbf{p}}_i, \hat{\sigma}_{-i}, \hat{\mathbf{p}}_{-i})$ from our mechanism.

Now we consider the strategy an agent plays in the game.

Definition 69 (Strategy)**.** Given a mechanism $\mathcal{M}$, we define the strategy of $\mathcal{M}$ for setting (n, Σ) as a mapping s from (σ, Q) (the signal and common prior received) to a probability distribution over $\mathcal{R}$ (the reported signal, prediction pair).

That is, for each possible signal σ and prior Q an receives, he will choose a signal, prediction pair to report from some distribution $s(\sigma, Q)$. We define a strategy profile $\mathbf{s}$ as a profile of all agents' strategies $\{s_1, s_2, ...s_n\}$ and we say agents play $\mathbf{s}$ if for any i, agent i plays strategy s_i. We say a strategy profile is **symmetric** if each agent plays the same strategy.

We define the **agent welfare** of a strategy profile $\mathbf{s}$ and a mechanism $\mathcal{M}$ for setting (n, Σ) with prior Q to be the expectation of the sum of payments to each agent and we write it as $AW_{\mathcal{M}}(n, \Sigma, Q, \mathbf{s})$. Note that for symmetric strategy profile,

the **agent welfare** is proportional to each agent's expected payment since everyone plays the same strategy.

A *Bayesian Nash equilibrium* consists of a strategy profile $s = (s_1, \ldots, s_n)$ such that no player wihes to change her strategy, given the strategies of the other players and the information contained in the prior and her signal. Formally,

Definition 70 (Bayesian Nash equilibrium)**.** Given a family of priors $\mathcal{Q}$, a strategy profile $s = (s_1, \ldots, s_n)$ is a Bayesian Nash equilibrium if and only if for any prior $Q \in \mathcal{Q}$, for any i, and for any s_i'

$$\begin{aligned}&\mathbb{E}_{(\hat{\sigma}_i', \hat{\mathbf{p}}_i') \leftarrow s_i'(\sigma_i, Q), (\hat{\sigma}_{-i}, \hat{\mathbf{p}}_{-i}) \leftarrow s_{-i}(\sigma_{-i}, Q)}\left[\nu_i(\hat{\sigma}_i', \hat{\mathbf{p}}_i', \hat{\sigma}_{-i}, \hat{\mathbf{p}}_{-i})\right] \\ \leq &\mathbb{E}_{(\hat{\sigma}_i, \hat{\mathbf{p}}_i) \leftarrow s_i(\sigma_i, Q), (\hat{\sigma}_{-i}, \hat{\mathbf{p}}_{-i}) \leftarrow s_{-i}(\sigma_{-i}, Q)}\left[\nu_i(\hat{\sigma}_i, \hat{\mathbf{p}}_i, \hat{\sigma}_{-i}, \hat{\mathbf{p}}_{-i})\right]\end{aligned}$$

In the case where, for some i, the equality holds if and only if $s_i' = s_i$, we say this strategy profile is a *strict Bayesian Nash equilibrium* for prior family $\mathcal{Q}$.

Remark 71 (Equilibrium for a Given Prior). Note that we assume agents have a common prior Q, so often for convenience, we will implicitly assume Q is fixed, at which point a strategy is a mapping from Σ to a probability distribution over $\mathcal{R}$. We will call such a strategy profile $\mathbf{s}$ an equilibrium for prior Q if it satisfies the condition of Bayesian Nash equilibrium when Q is fixed.

Assuming a fixed prior Q, for any strategy profile $s = (s_1, s_2, ..., s_n)$, we will represent the marginal distribution of an agent i's strategy for her signal report as a matrix θ_i where $\theta_i(\hat{\sigma}, \sigma)$ is the probability that agent will report signal $\hat{\sigma}$ when his private signal is σ. Note that θ_i is a **transition matrix**, that is the sum of every column is 1. We call θ_i the signal strategy of agent i. We also call $(\theta_1, \theta_2, ..., \theta_n)$ the signal strategy of s. We define the *average signal strategy* of s as $\bar{\theta}_n = \frac{\sum_i \theta_i}{n}$. The following claim relates this average signal strategy to the distribution of all reported signals:

Claim 72. Assume that the distribution over all agents' private signals is $\omega \in \Delta_\Sigma$, the distribution over all agents' reported signals will be $\bar{\theta}_n \omega$.

Note that the mechanism actually collects agents' reported signals, so in order to estimate the distribution over their private signals, we hope $\bar{\theta}_n$ is (close to) the identity matrix I.

4.2.3 Special strategy profiles

In this section, we will introduce three special types of strategy profiles that we call *truth-telling*, *best prediction strategy profiles*, and *permutation strategy profiles*.

Definition 73 (Truth-telling)**.** We define a strategy profile as truth-telling if for all i, and for all Q, $s(\sigma_i, Q) = (\sigma_i, \mathbf{q}_{\sigma_i})$ with probability 1. We write the truth-telling strategy profile as $\mathbf{T}$.

For every agent i, let $\hat{\sigma}$ be a randomly chosen agent's reported signal, when other agents tell the truth, the distribution of $\hat{\sigma}$ is $\mathbf{q}_{\sigma_i}$. However, if agents play strategy $\mathbf{s}$, for agent i, the distribution of $\hat{\sigma}$ depends on not only his prior Q but also the strategy $\mathbf{s}$. We define the distribution of $\hat{\sigma}$ for agent i as $\mathbf{q}^{\mathbf{s}}_{\sigma_i}$.

Claim 74.

$$\mathbf{q}^{\mathbf{s}}_{\sigma_i} = \theta_{-i} \mathbf{q}_{\sigma_i}$$

where $(\theta_1, \theta_2, ..., \theta_n)$ is $\mathbf{s}$'s signal strategy and $\theta_{-i} = \frac{\sum_{j \neq i} \theta_j}{n-1}$.

When agents play strategy $\mathbf{s}$, to best predict other agents' reported signal, agent i should be report $\mathbf{q}^{\mathbf{s}}_{\sigma_i}$ rather than $\mathbf{q}_{\sigma_i}$. This motivates our definition for *best prediction strategy profile* which is a strategy profile where every agent i gives his "best prediction" $\mathbf{q}^{\mathbf{s}}_{\sigma_i}$.

Definition 75 (Best Prediction Strategy Profile)**.** We say a strategy profile $\mathbf{s}$ is a best prediction strategy profile if for every agent i, he reports $\mathbf{q}^{\mathbf{s}}_{\sigma_i}$. We call a best strategy

prediction strategy profile $\mathbf{s}$ a *symmetric best strategy prediction strategy profile* if $\theta_i = \theta$ for every i.

Now we begin to introduce the definition of a permutation strategy profile. Intuitively, if agents "collude" to relabel the signals and then tell the truth with relabeled signals, they actually play what we will call permutation strategy profile.

Given a permutation $\pi : \Sigma \mapsto \Sigma$ (which is actually a relabeling of signals), by abusing notation a little bit, we define $\pi : \mathcal{Q} \mapsto \mathcal{Q}$ as a mapping from a prior Q to a *permuted prior* $\pi(Q)$ where for any $\sigma_1, \sigma_2, ..., \sigma_n \in \Sigma$,

$$Pr_{\pi(Q)}(\sigma_1, \sigma_2, ..., \sigma_n) = Pr_Q(\pi^{-1}(\sigma_1), \pi^{-1}(\sigma_2), ..., \pi^{-1}(\sigma_n))$$

where σ_i is the private signal of agent i. Notice that it follows that:

$$Pr_{\pi(Q)}(\pi(\sigma_1), \pi(\sigma_2), ..., \pi(\sigma_n)) = Pr_Q(\sigma_1, \sigma_2, ..., \sigma_n).$$

Intuitively, $\pi(Q)$ is the same with Q when the signals are relabeled according to π.

For any strategy s, we define $\pi(s)$ as the strategy such that $\pi(s)(\sigma, Q) = s(\pi(\sigma), \pi(Q))$.

Definition 76 (Permuted Strategy Profile)**.** For any strategy profile $\mathbf{s}$, we define $\pi(\mathbf{s})$ as a strategy profile with $\pi(\mathbf{s}) = (\pi(s_1), \pi(s_2), ..., \pi(s_n))$.

Note that $\pi^{-1}\pi Q = Q$ which implies $\pi^{-1}\pi(\mathbf{s}) = \mathbf{s}$.

Definition 77 (Permutation Strategy Profile)**.** We define a strategy profile $\mathbf{s}$ as a permutation strategy profile if there exists a permutation $\pi : \Sigma \to \Sigma$ such that $\mathbf{s} = \pi(\mathbf{T})$.

Note that if agents play $\pi(\mathbf{T})$, then the signal strategy of each agent is π, and so the distribution of report profiles is $\bar{\theta}_n\omega = \pi\omega$.

There exists a natural bijection between permutation strategy profiles and $|\Sigma|\times|\Sigma|$ permutation matrices. If the permutation strategy profile is constructed by permutation π, the only non-zero entries of the corresponding permutation matrix θ_π are $\theta_\pi(\pi(\sigma),\sigma) = 1$ for all $\sigma \in \Sigma$. For a transition matrix θ, if θ is not a permutation matrix, we would like to give a definition for when a transition matrix θ is what we call τ-*close* to a permutation given any $\tau > 0$. This definition is motivated by the following claim and will be described after it.

Claim 78. For any transition matrix $\theta_{m\times m}$ where the sum of every column is 1, θ is a permutation matrix iff for any row of θ, there at most one non-zero entry.

Now we give a definition for τ-*close*.

Definition 79 (τ-close)**.** We say a signal strategy θ is τ-*close* to a permutation if for any row of θ, there is at most one entry that is greater than τ.

Thus a permutation stragety is 0-close to a permutation. For any stategy profile s, if the average signal strategy of s is τ-*close* to a permutation matrix, we say s is τ-*close* to a permutation profile as well.

Recall that f-divergence([4]) is used to measuring the "difference" between distributions. One important property of the f-divergence family is information monotonicity: for any two distributions, if we post-process each distribution in the same way, the two distributions will become "closer" because of the information loses.

The information monotonicity of f-divergence implies that:

Fact 80. Given SNIFE prior Q, for any θ that is not a permutation, there exists two private signals $\sigma_1 \neq \sigma_2$ such that $D_f(\theta\mathbf{q}_{\sigma_1}, \theta\mathbf{q}_{\sigma_2}) < D_f(\mathbf{q}_{\sigma_1}, \mathbf{q}_{\sigma_2})$

Proof. First notice that when θ is not a permutation, based on Claim 78, there exists a row of θ such that the row has at least two positive entries, in other words, there exists $\sigma, \sigma', \sigma''$ such that $\theta(\sigma,\sigma'), \theta(\sigma,\sigma'') > 0$. Based on the non-zero and fine-grained

assumptions of Q, there exists $\sigma_1 \neq \sigma_2$ such that $\theta(\sigma, \sigma')\mathbf{p}(\sigma'), \theta(\sigma, \sigma'')\mathbf{p}(\sigma'') > 0$ and $\frac{\mathbf{p}(\sigma')}{\mathbf{p}(\sigma'')} \neq \frac{\mathbf{q}(\sigma')}{\mathbf{q}(\sigma'')}$ where $\mathbf{p} = \mathbf{q}_{\sigma_1}, \mathbf{q} = \mathbf{q}_{\sigma_2}$. When $\theta(\sigma, \sigma')\mathbf{p}(\sigma'), \theta(\sigma, \sigma'')\mathbf{p}(\sigma'') > 0$, we have $\theta(\sigma, \cdot)\mathbf{p} > 0$. By Lemma 6, we have $D_f(\theta\mathbf{q}_{\sigma_1}, \theta\mathbf{q}_{\sigma_2}) < D_f(\mathbf{q}_{\sigma_1}, \mathbf{q}_{\sigma_2})$ □

4.2.4 Mechanism design tools

Hellinger-divergence and strictly proper scoring rules are two of the main tools we will use in our mechanism design. Starting with [45], proper scoring rules have become a common ingredient in mechanisms for unverifiable information elicitation (e.g. [51, 67]). Hellinger-divergence is a type of f-divergence ([4]). F-divergence is always used in measuring the "difference" between distributions. One important property of f-divergence is information monotonicity: For any two distributions, if we use the same way to post-process each distribution, the two distributions will become "closer" because of potential information loses. The reason we pick Hellinger-divergence rather than other f-divergence is that we need **square root triangle inequality** of Hellinger-divergence (which we will describe later).

Hellinger-divergence Hellinger-divergence is a special case of f-divergence.

$D^* : \Delta_\Sigma \times \Delta_\Sigma \to \mathbb{R}$ is a non-symmetric measure of difference between distribution $\mathbf{p} \in \Delta_\Sigma$ and distribution $\mathbf{q} \in \Delta_\Sigma$ and is defined to be

$$D^*(\mathbf{p}, \mathbf{q}) = \sum_\sigma (\sqrt{\mathbf{p}(\sigma)} - \sqrt{\mathbf{q}(\sigma)})^2.$$

We highlight two important properties of Hellinger-divergence: one is *Information Monotonicity* which other f-divergences also have; another is *square root triangle inequality*.

(1) **Information Monotonicity**: For any $\mathbf{p}, \mathbf{q}$, and transition matrix $\theta \in \mathbb{R}^{|\Sigma| \times |\Sigma|}$ where $\theta(\sigma, \sigma')$ is the probability that we map σ' to σ, we have $D^*(\mathbf{p}, \mathbf{q}) \geq D^*(\theta\mathbf{p}, \theta\mathbf{q})$.

When θ is a permutation, $D^*(\mathbf{p}, \mathbf{q}) = D^*(\theta\mathbf{p}, \theta\mathbf{q})$.

(2) **Square root triangle inequality**: $|\sqrt{D^*(\mathbf{p}, \mathbf{q})} - \sqrt{D^*(\mathbf{p}, \mathbf{q}')}| < \sqrt{D^*(\mathbf{q}', \mathbf{q})}$ for any $\mathbf{p}, \mathbf{q}, \mathbf{q}'$

Proper scoring rules are a key tool in the design of mechanisms [51] in the BTS framework. In such mechanism, agents are asked to report their private information and forecast for other agents and paid based on a "prediction score" and an "information score". The prediction score is usually calculated by a proper scoring rule and the information score is customized.

Prediction Score via Proper Scoring Rules Agents will receive a prediction score based on how well their prediction predicts a randomly chosen agent's reported signal. Say an agent i reports prediction $\hat{\mathbf{p}}_i$ then a random agent, call him agent j, is chosen, agent i will receive a prediction score $PS(\hat{\sigma}_j, \hat{\mathbf{p}}_i)$ where PS is a proper scoring rule. Note that any proper scoring rule works. $PS(\hat{\sigma}_j, \hat{\mathbf{p}}_i)$ is maximized if and only if agent i's reported prediction $\hat{\mathbf{p}}_i$ is his expected likelihood for $\hat{\sigma}_j$. Agent i cannot pretend to have a different expected likelihood without reducing his expectation for his prediction score.

4.3 The Disagreement mechanism

4.3.1 Buiding block—Divergence-Based BTS

In this section, we introduce a building block of our Disagreement Mechanism—Divergence-Based BTS [55]. It follows the BTS framework and still pays agents an "information score" and a "prediction score". The main idea of Divergence-Based BTS is that the mechanism punishes the **Inconsistency** of agents—the "difference" between two random agents' predictions when they report the same signal. The common prior assumption tells us agents cannot agree to disagree. That is, if two agents receive the same private information, they must have the same "belief" about

the world. In our setting, if agents tell the truth, whenever two agents report the same signal, they will report the same prediction as well. Thus, everyone telling the truth is a consistent strategy. Since Divergence-Based BTS punishes inconsistency, the truth-telling strategy will be encouraged in Divergence-Based BTS.

Divergence-Based BTS [55] $\mathcal{M}$: Let $\alpha, \beta > 0$ be parameters and let PS be a strictly proper scoring rule, then we define $\mathcal{M}(\alpha, \beta, PS)$[1] as follows:

1. Each agent i reports a signal and a prediction $r_i = (\hat{\sigma}_i, \hat{\mathbf{p}}_i)$

2. For each agent i and agent j, we define a prediction score that depends on agent i's prediction and agent j's report signal

$$score_P(r_i, r_j) = PS(\hat{\sigma}_j, \hat{\mathbf{p}}_i),$$

and an information score

$$score_I(r_i, r_j) = \begin{cases} 0 & \hat{\sigma}_i \neq \hat{\sigma}_j \\ -(PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_j) - PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_i)) & \hat{\sigma}_i = \hat{\sigma}_j \end{cases}$$

3. Each agent i is matched with a random agent j. The payment for agent i is

$$payment_{\mathcal{M}(\alpha,\beta,\mathcal{PS})}(i, \mathbf{r}) = \alpha score_P(r_i, r_j) + \beta score_I(r_i, r_j).$$

Theorem 81. *[55]*

For any $\alpha, \beta > 0$ and any strictly proper scoring rule PS, $\mathcal{M}(\alpha, \beta, PS)$ has truth-telling as a strict Bayesian-Nash equilibrium whenever the prior Q is informative and symmetric.

[1]This mechanism is a little bit different with Divergence-Based BTS mechanism [55]. Divergence-Based BTS uses specific proper scoring rule (log scoring rule). But it is easy to see using general proper scoring rules still keeps the strictly truthful property of Divergence-Based BTS.

We introduce the proof in appendix (Section A.1.5).

Main Drawback of Divergence-Based BTS The main drawback is that there may be many other equilibria with inconsistency score 0. Agents can simply report the a priori most popular signal and predict that everyone does the same. This strategy is a consistent equilibrium and gives agents the maximum possible payoff since their predictions are perfect. In particular, for any non-trivial prior, this strategy pays *strictly more* than the truth-telling equilibrium—so that it Pareto dominates truth-telling.

The above extreme example provides a effortless and meaningless equilibrium but is preferred by agents in Divergence-Based BTS. To deal with this problem, one key observation is that in the meaningless equilibrium mentioned above, the unitary predictions implies their report profiles have little information. At a high level, the "disagreement" between agents represents the amount of information their report profiles have. Motivated by this extreme example, we design a new mechanism—the Disagreement Mechanism—that encourages "disagreement".

4.3.2 The Disagreement mechanism and main theorem

In this section, we will describe our Disagreement Mechanism and state our main theorem. To design our mechanism, we start with the Divergence-Based BTS and (a) first use a typical trick to create a zero-sum game which has the same equilibria as the Divergence-Based BTS; (b) pay each agent an extra score that only depends on other agents which will not change the structure of the equilibria. We want this extra score to represent "classification score" (See Figure 4.1).

Disagreement Mechanism $\mathcal{M}+(\alpha, \beta, PS(\cdot,\cdot))$ $\mathbf{r} = \{r_1, r_2, ..., r_n\}$ is all agents' report profiles where for any r, $r_i = (\hat{\sigma}_i, \hat{\mathbf{p}}_i)$.

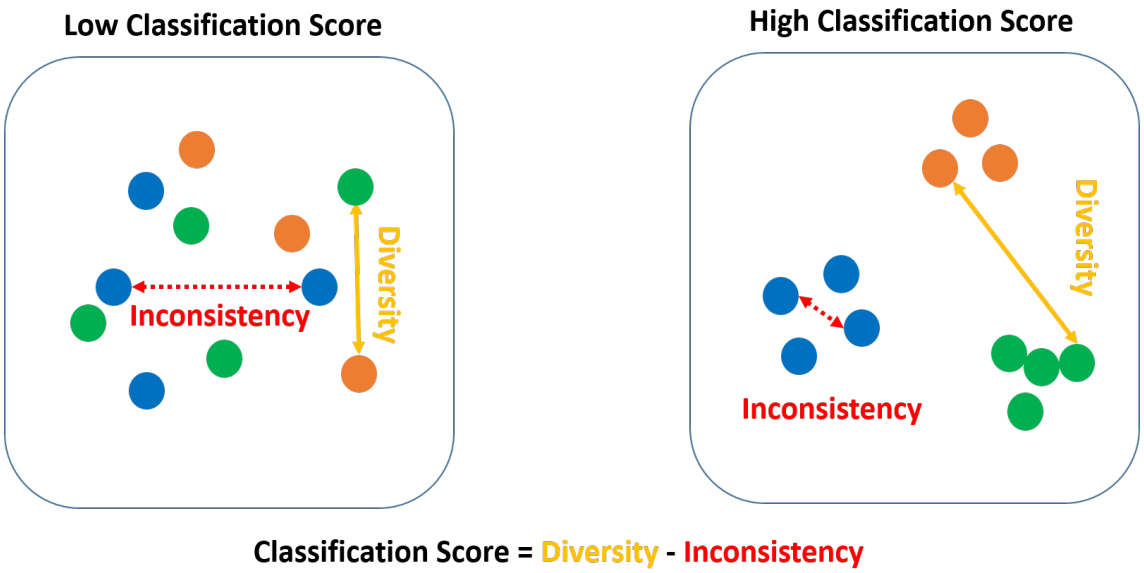

Figure 4.1: An illustration of Classification Score

Each point represents an agent's report profile—the *color* represents the *signal* the agent reports; the *position* represents the *prediction* the agent reports. We informally define **Inconsistency** as the *average* disagreement between every two agents' predictions when they report the *same* signal and **Diversity** as the *average* disagreement between every two agents' predictions when they report *different* signals. We informally define **Classification Score** as Diversity minus Inconsistency. Note that the report profiles in the right figure will have a much higher classification score than those in the left figure since the right figure has high Diversity and low Inconsistency.

1. *Zero-sum Trick*: Divide the agents into two non-empty groups—group A and group B. Each group of agents plays the game (mechanism) $\mathcal{M}$ that is restricted in their own group. For group A, each agent i_A receives a

$$\begin{aligned} score_{\mathcal{M}}(i_A, \mathbf{r}) =& payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i_A, \mathbf{r}_A) \\ & - \frac{1}{|A|} \sum_{j_B \in B} payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(j_B, \mathbf{r}_B) \end{aligned}$$

   Where $payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i_A, \mathbf{r}_A)$ is agent i_A's payment when he is paid by mechanism $\mathcal{M}(\alpha, \beta, PS(\cdot,\cdot))$ given group A's report profiles $\mathbf{r}_A$ and that he can only be paired with a random peer from group A (we have similar explanation for $payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(j_B, \mathbf{r}_B)$). For agents in group B, we use the analogous way to score them.

2. *Additional Classification Reward*: Each agent i is matched with two random agents j, k chosen from all agents (including group A and group B), the payment for agent i is

$$payment_{\mathcal{M}+(\alpha,\beta,PS(\cdot,\cdot)}(i, \mathbf{r}) = score_{\mathcal{M}}(i, \mathbf{r}) + score_C(r_j, r_k)$$

   where

$$score_C(r_j, r_k) = \begin{cases} D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) & \hat{\sigma}_j \neq \hat{\sigma}_k \\ -\sqrt{D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)} & \hat{\sigma}_j = \hat{\sigma}_k \end{cases}$$

   recall that D^* denotes the Hellinger Divergence.

Theorem 82. *For any number of signals m, given any SNIFE prior, if the number of agents $n \geq 3$, then in $\mathcal{M}+(\alpha, \beta, PS(\cdot,\cdot))$ with $\frac{\alpha}{\beta} < \frac{1}{4m}$,*

1. *(Truthful) truth-telling is a strict Bayesian Nash equilibrium;*

2. *(Focal) in any permutation equilibrium,* every agent *has equal expected payment with truth-telling; and in any* symmetric *equilibrium that is not a permutation equilibrium,* every agent*'s expected payment is strictly less than that of truth-telling.*

3. *(Robust Focal) any* symmetric *equilibrium that pays within γ_1 of truth-telling must be $\tau_1(\gamma_1)$ close to a permutation strategy profile; and moreover*

4. *(Tight) no detail free mechanism can have truth-telling as an equilibrium that has strictly higher agent-welfare than all other permutation equilibria.*

where $\tau_1(\gamma_1) = O(\sqrt[3]{\gamma_1})$, (the constants we omit only depend on the first two moments of prior Q)[2].

We extend our results to asymmetric equilibria when the number of agents is sufficiently large in Section A.1.2.

4.3.3 Proof highlights

In this section we give a few proof highlights.

First note that to show each agent's expected payment in a symmetric equilibrium is less than that of truth-telling, we only need to show the sum of all agents' expected payments—agent welfare—is less than that of truth-telling since everyone plays the same strategy in a symmetric equilibrium.

We first show that the agent welfare of our *Disagreement Mechanism* is *Diversity* minus *Inconsistency*, which follows by a straightforward computation. It remains to show that *Diversity* minus *Inconsistency* has the aforementioned properties.

Best Prediction Strategy Profiles: We call a strategy profile a *best prediction strategy profile* if for any i, agent i reports a prediction that maximizes his prediction

[2]Actually $\tau_1(\gamma_1) = \frac{1}{c_1}\sqrt[3]{\frac{\gamma_1}{c_2,c_3,c_4}}$

score. By some calculations, we know agent i's *best prediction* is $\theta_{-i}\mathbf{q}_{\sigma_i}$ given σ_i is his private signal and recall that $\theta_{-i} = \frac{\sum_{j\neq i}\theta_i}{n-1}$ where $(\theta_1, \theta_2, ..., \theta_n)$ is the signal strategy. We call this strategy profile a *symmetric* best prediction strategy profile if there exists a signal strategy θ such that $\theta_i = \theta$ for any i. Based on the definition of permutation strategy profile, it is clear that any permutation strategy profile is a symmetric *best prediction* strategy profile.

Consider two agents who report different signals. If they use a permutation strategy profile π then their predictions will be $\pi\mathbf{q}_\sigma, \pi\mathbf{q}_{\sigma'}$ given their private signals are $\sigma \neq \sigma'$. If they use a symmetric best prediction strategy, then their reported predictions will be $\theta\mathbf{q}_\sigma, \theta\mathbf{q}_{\sigma'}$. In the first case, the Hellinger divergence between the two agents' reported predictions is $D^*(\pi\mathbf{q}_\sigma, \pi\mathbf{q}_{\sigma'}) = D^*(\mathbf{q}_\sigma, \mathbf{q}_{\sigma'})$ while in the second case, the Hellinger divergence between the two agents' reported predictions is $D^*(\theta\mathbf{q}_\sigma, \theta\mathbf{q}_{\sigma'}) \leq D^*(\mathbf{q}_\sigma, \mathbf{q}_{\sigma'}) = D^*(\pi\mathbf{q}_\sigma, \pi\mathbf{q}_{\sigma'})$. The inequality follows from the information monotonicity of Hellinger divergence. Thus, the two agents' predictions in the second case is "closer" than those in the first case. So a permutation strategy profile is more diverse than any other symmetric best prediction strategy, and additionally has no inconsistency. To make permutation strategy profiles beat symmetric best prediction strategy profiles, it is enough to just pay agents the additional diversity reward.

General Equilibria: However, **the biggest challenge** is that there exists equilibria that are not best prediction strategy profiles. Thus, *it is not enough to just pay agents an additional diversity reward.* To deal with this challenge, we replace diversity by *classification score.* To show that classification score works, we map each equilibrium s^* to a strategy profile s^*_{BP} that belongs to *best prediction strategy profiles.* The *technical heart* of the proof bounds the classification score of an equilibrium strategy profile s^* by the diversity of its corresponding best prediction strategy profile s^*_{BP}. Once we finish this, we can bound the classification score of any equilibrium

strategy profile by the classification score of permutation strategy profiles (note that for permutation strategy profiles, the classification score is equal to the diversity since they are consistent strategy profiles) and complete the proof.

Asymmetric Equilibria: In the more complicated asymmetric case, the difficulty is that even if agents play best prediction strategy profiles, we cannot use information monotonicity to prove permutation strategy profiles gain the strictly highest classification score. However, if the number of agents is large enough, we will see any strategy profile that belongs to *best prediction strategy profiles* family is "almost symmetric". Using "almost symmetric" result, we can generalize the above framework to approximate work for asymmetric case.

Finally, we show that equilibrium that having the classification score close to that of truth-telling, must be close to a permutation equilibrium.

Tightness Result: The intuitive explanation for this tightness result is that the agents can collude to relabel the signals and the mechanism has no way to defend against this relabelling without knowing some information about agents' common prior. The key idea to prove that result is what we refer to as **Indistinguishable Scenarios**, that is, for the scenario A where agents collude to relabel the signals, there always exists another scenario B where agents tell the truth such that no detail free and truthful mechanism can distinguish A and B.

4.4 Mapping Bayesian truth serum into our information theoretic framework

Bayesian Truth Serum (BTS) [51] rewards the agents whose answer is "surprisingly popular". In this section, we will show that in BTS, essentially each agent is paid the mutual information between her information and the aggregated information conditioning a random peer's information which matches our Mutual Information

Paradigm. We show this via the connection we found between the log scoring rule and Shannon mutual information—the accuracy gain equals the information gain. Mapping Bayesian Truth Serum into our information theoretic framework substantially simplifies the proof in Prelec [51] via directly applying the information-monotone property of Shannon mutual information.

4.4.0.1 Prior work

Prelec [51] proposes the Bayesian Truth Serum mechanism in the single-task setting. In addition to the common prior and the symmetric prior assumptions, two additional assumptions are required:

Assumption 83 (Conditional Independence)**.** *We define the state of the world as a random variable $W : \Omega \mapsto \Delta_\Sigma$ such that given that $W = \omega$, agents' private signals are independently and identically distributed. That is, for every i, agent i receives signal σ with probability $\omega(\sigma)$.*

Assumption 84 (Large Group)**.** *The number of agents is infinite.*

We define a random variable $\hat{W} : \Omega \mapsto \Delta_\Sigma$ such that its outcome is the distribution over agents' reported signals. The distribution over $\hat{W}$ dependes on all agents' strategies. With the large group assumption, when agents tell the truth, $\hat{W} = W$.

BTS uses $\hat{W}$ as the posterior distribution and uses agents' forecasts as the prior distribution, and then rewards agents for giving signal reports that are "unexpectedly common" with respect to this distribution. Intuitively, an agent will believe her private signal is underestimated by other agents which means she will believe the actual fraction of her own private signal is higher than the average of agents' forecasts.

Prelec also proposes the signal-prediction framework for the design of detail free mechanism in the single-task setting.

Signal-prediction framework [51] Given a setting (n, Σ) with a symmetric common prior Q, the signal-prediction framework defines a game in which each agent i is asked to report his private signal $\sigma_i \in \Sigma$ and his prediction $\mathbf{p}_i \in \Delta_\Sigma$, a distribution over Σ, where $\mathbf{p}_i = \mathbf{q}_{\sigma_i}$. For any $\sigma \in \Sigma$, $\mathbf{p}_i(\sigma)$ is agent i's (reported) expectation for the fraction of other agents who has received σ given he has received σ_i. However, agents may not tell the truth. In this framework, the report space $\mathcal{R} = \Sigma \times \Delta_\Sigma$. We define a report profile of agent i as $r_i = (\hat{\sigma}_i, \hat{\mathbf{p}}_i) \in \mathcal{R}$ where $\hat{\sigma}_i$ is agent i's reported signal and $\hat{\mathbf{p}}_i$ is agent i's reported prediction.

We would like to encourage truth-telling $\mathbf{T}$, namely that agent i reports $\hat{\sigma}_i = \sigma_i, \hat{\mathbf{p}}_i = \mathbf{q}_{\sigma_i}$. To this end, agent i will receive some payment $\nu_i(\hat{\sigma}_i, \hat{\mathbf{p}}_i, \hat{\sigma}_{-i}, \hat{\mathbf{p}}_{-i})$ from the mechanism.

Mechanism Bayesian Truth Serum (BTS(α)) [51] The Bayesian Truth Serum (BTS) follows the signal-prediction framework. Here, we introduce the payment of BTS. Each agent i has two scores: a **prediction score** and an **information score**. BTS pays each agent

$$\textit{prediction score} + \alpha \cdot \textit{information score}$$

where $\alpha > 1$ To calculate the scores, for every agent i, the mechanism chooses a reference agent $j \neq i$ uniformly at random. Agent i's prediction score is

$$score_{Pre}(r_i, r_j) := L(\hat{\sigma}_j, \hat{\mathbf{p}}_i) - \log fr(\hat{\sigma}_j | \hat{\boldsymbol{\sigma}}_{-j}) = \log \hat{\mathbf{p}}_i(\hat{\sigma}_j) - \log fr(\hat{\sigma}_j | \hat{\boldsymbol{\sigma}}_{-j})$$

Note that only the log scoring rule part $L(\hat{\sigma}_j, \hat{\mathbf{p}}_i)$ is related to agent i's report. Based on the property of the log scoring rule, for agent i, in order to maximize her prediction score, the best $\hat{\mathbf{p}}_i(\sigma)$ should be her posterior expectation of the fraction of the agents who *report* σ rather than *receive*.

Agent i's information score is

$$score_{Im}(r_i, r_j) := \log \frac{fr(\hat{\sigma}_i|\hat{\boldsymbol{\sigma}}_{-i})}{\hat{\mathbf{p}}_j(\hat{\sigma}_i)} = \log fr(\hat{\sigma}_i|\hat{\boldsymbol{\sigma}}_{-i}) - \log \hat{\mathbf{p}}_j(\hat{\sigma}_i)$$

where $fr(\hat{\sigma}_i|\hat{\boldsymbol{\sigma}}_{-i})$ is the fraction of all reported signals $\hat{\boldsymbol{\sigma}}_{-i}$ (excluding agent i) that agree with agent i's reported signal $\hat{\sigma}_i$, which can be seen as the posterior expectation of the fraction of agents who report $\hat{\sigma}_i$ conditioning on all agents' reports, while $\hat{\mathbf{p}}_j(\hat{\sigma}_i)$ is agent j's posterior expectation of that fraction conditioning on agent j's private signal. Intuitively, the signals that actually occur more than other agents believe they will receive a higher information score.

Now we restate the main theorem concerning Bayesian Truth Serum:

Theorem 85. *[51] With the common prior, the symmetric prior, the conditional independence, and the large group assumptions, BTS(α) is detail free, (i) truthful and (ii) the expected average information score when everyone tells the truth is higher than that in any other equilibrium. Moreover, (iii) for $\alpha > 1$, BTS is focal.*

Prelec [51] uses some clever algebraic calculations to prove the main results. In the next section, we will apply our "accuracy gain=information gain" observation to map Bayesian Truth Serum [51] into our information theoretical framework and show results (ii) and (iii) via applying the data processing inequality of Shannon mutual information. We put Prelec [51]'s proof for results (i) in appendix since it is already sufficiently simple and not very related to our framework.

4.4.0.2 Using our information theoretic framework to analyze BTS

A key observation of BTS is that when agents report the optimal predictions, the average information score is exactly the "accuracy gain"—the accuracy of the posterior prediction for a *random agent's report* conditioning on *all agents' reports*, minus the accuracy of a *random reference agent j*'s posterior prediction for the *ran-*

dom agent's report conditioning on agent j's private signal. Based on Lemma 24, this accuracy gain equals the Shannon mutual information between a *random agent's reported signal* and *all agents' reports* conditioning on the *random reference agent* j's private signal $\Psi_j = \sigma_j$. Therefore, the expected information score can be represented as the form of Shannon mutual information. We have similar analysis for the prediction score. We formally state the above observation in Lemma 86. Aided by this lemma, we will show results (ii) and (iii) via applying the information-monotone property of Shannon mutual information.

Lemma 86. *In BTS, when agents tell the truth, each agent i's expected information score and prediction score are*

$$I(W; \Psi_i|\Psi_j), \qquad -I(W; \Psi_j|\Psi_i)$$

respectively, $\forall j \neq i$. When the agents play a non-truthful equilibrium, we denote random variable $\hat{\Psi}$ as the reported signal of an agent who is picked uniformly at random, the expected average information score and prediction score are

$$I(\hat{W}; \hat{\Psi}|\Psi_j), \qquad -I(\hat{W}; \hat{\Psi}|\Psi_i)$$

respectively, $\forall i, j$.

Proof. When agents tell the truth, each agent i's expected information score is

$$\begin{aligned}
&\mathbb{E}_{\Psi_i,\Psi_j,W} L(\Psi_i, Pr[\mathbf{\Psi}_i|W]) - L(\Psi_i, Pr[\mathbf{\Psi}_i|\Psi_j]) \\
&= \mathbb{E}_{\Psi_i,\Psi_j,W} L(\Psi_i, Pr[\mathbf{\Psi}_i|W, \Psi_j]) - L(\Psi_i, Pr[\mathbf{\Psi}_i|\Psi_j]) \quad \text{(Conditional independence)} \\
&= I(W; \Psi_i|\Psi_j) \quad \text{(Theorem 24 / Expected accuracy gain equals information gain)}
\end{aligned}$$

when she is paired with reference agent j. Since the prior is symmetric, $I(W; \Psi_i|\Psi_j)$ is independent of the identity of j if $j \neq i$.

In any equilibrium $\mathbf{s}$, based on the properties of proper scoring rules, each agent j will always maximize his expected prediction score by truthfully reporting his predictions. Moreover, for agent j, his reference agent is picked uniformly at random. Therefore,

$$\hat{\mathbf{p}}_j(\hat{\sigma}) = Pr[\hat{\Psi} = \hat{\sigma} | \Psi_j = \sigma_j]$$

where $\hat{\Psi}$ is the reported signal of an agent who is picked uniformly at random.

Then we can replace W, Ψ_i by $\hat{W}, \hat{\Psi}$ in the above equations and prove that the expected average information score is

$$I(\hat{W}; \hat{\Psi} | \Psi_j).$$

The analysis for the expected prediction score is the same as the above analysis except that we need to exchange i and j.

□

Proof of Theorem 85 (ii), (iii). Based on Lemma 86, when agents play an equilibrium, the expected average information score equals

$$\begin{aligned} I(\hat{W}; \hat{\Psi} | \Psi_j) &= \sum_{\sigma_j} \Pr[\Psi_j = \sigma_j] I(\hat{W}; \hat{\Psi} | \Psi_j = \sigma_j) \\ &\leq \sum_{\sigma_j} \Pr[\Psi_j = \sigma_j] I(\hat{W}, W; \hat{\Psi} | \Psi_j = \sigma_j) \quad \text{(Data processing inequality)} \end{aligned}$$

Note that, when the number of agents is infinite, since every agent's strategy is independent with each other, we can see W determines $\hat{W}$ [3]. Therefore,

[3]When $W = \omega$, $\hat{W} = \frac{1}{n} \sum_i M_i^T \omega$ where M_i^T is the transpose matrix of the transition matrix corresponded to agent i's strategy for signal reporting, and the distribution ω is represented by a column vector.

$$\sum_{\sigma_j} \Pr[\Psi_j = \sigma_j] I(\hat{W}, W; \hat{\Psi} | \Psi_j = \sigma_j)$$

$$= \sum_{\sigma_j} \Pr[\Psi_j = \sigma_j] I(W; \hat{\Psi} | \Psi_j = \sigma_j)$$

$$\leq \sum_{\sigma_j} \Pr[\Psi_j = \sigma_j] I(W; \Psi_i | \Psi_j = \sigma_j), \forall i \neq j$$

(Data processing inequality and the symmetric prior assumption)

$$= I(W; \Psi_i | \Psi_j), \forall i \neq j$$

Thus, the expected average information score is maximized when everyone tells the truth.

It is left to show for $\alpha > 1$, in BTS(α), the agent-welfare is maximized by truth-telling over all equilibria. Lemma 86 shows that when the prior is symmetric, the sum of the expected prediction scores equals the sum of the expected information scores in any equilibrium. Thus, when $\alpha > 1$, the agent welfare is proportional to the sum of the expected information scores which is maximized by truth-telling over all equilibria.

□

It is natural to ask if we replace the $-log$ in BTS's information score by other convex functions, what property of BTS we can still preserve. The following theorem shows that even though we may not ganrantee the truthful property of BTS, the average expected information score is still monotone with the amount of information for any convex function we use.

Theorem 87. *If we replace the information score in BTS by $f(\frac{\hat{p}_j(\hat{\sigma}_i)}{fr(\hat{\sigma}_i|\hat{\boldsymbol{\sigma}}_{-i})})$ where f is a convex function, result (ii)—the expected average information score when everyone tells the truth is higher than that in any other equilibrium—is preserved.*

Proof. When agents tell the truth, each agent i's expected information score is

$$\begin{aligned}
&\mathbb{E}_{\Psi_i,\Psi_j,W} f(\frac{Pr[\Psi_i|\Psi_j]}{Pr[\Psi_i|W]}) \\
&= \mathbb{E}_{\Psi_i,\Psi_j,W} f(\frac{Pr[\Psi_i|\Psi_j]}{Pr[\Psi_i|W,\Psi_j]}) && \text{(Conditional independence)} \\
&= \mathbb{E}_{\Psi_i,\Psi_j,W} f(\frac{Pr[\Psi_i|\Psi_j]Pr[W|\Psi_j]}{Pr[\Psi_i,W|\Psi_j]}) \\
&= MI^f(W;\Psi_i|\Psi_j)
\end{aligned}$$

In any equilibrium $\mathbf{s}$, based on the properties of proper scoring rules, each agent j will always maximize their prediction by truthfully report their predictions, thus,

$$\hat{\mathbf{p}}_j(\hat{\sigma}) = Pr[\hat{\Psi} = \hat{\sigma}|\Psi_j = \sigma_j].$$

Then we can replace W, Ψ_i by $\hat{W}, \hat{\Psi}$ in the above equations and prove that the expected average information score is

$$MI^f(\hat{W};\hat{\Psi}|\Psi_j).$$

With the similar proof of Theorem 85, the theorem follow immediately from the data processing inequality of f-mutual information. □

CHAPTER V

Expertise Elicitation

5.1 Related work

Model perspective Prior work has modeled heterogeneous expertise where different agents receive a different number of signals [25] or expertise is embedding in several dimensions [20, 76, 65, 44]; however in these works lower expertise/effort along with a certain dimension only leads to a more noisy signal. In contrast, our model allows such signals to be systematically biased.

Mechanism design perspective The most related work with the current chapter is Prelec, Seung, and McCoy [52] which uses Bayesian Truth Serum [51] to incentivize agents to report their signal and selects the most surprising signal (measured by occurring more than its average prediction) as the final answer. McCoy and Prelec [44] follow Prelec, Seung, and McCoy [52] to propose a probabilistic model to learn the expertise of agents. Riley [58] compares the peer prediction decision rule (similar to Prelec, Seung, and McCoy [52]) and the majority vote rule and exhibits cases where each outperforms the other. The current chapter differs with Prelec, Seung, and McCoy [52] in the model and assumptions as well as the possible applications. Prelec, Seung, and McCoy [52] only focus on the single-task setting and assume that agents receive the signals endogenously (without effort). In contrast, this chapter

considers both single and multiple task settings and the model used in this chapter handles both exogenous and indigenous signals.

The mechanism design framework in the current chapter extends the information theoretic framework MIP. Agarwal et al. [2] propose a mechanism that works for the heterogeneous participants in the multi-task setting. Mandal et al. [42] consider the heterogeneous tasks setting. They both do not assume the hierarchy of the information and cannot be applied to identify and elicit expertise.

Algorithmic perspective Several works [76, 20, 65, 23] provide clever methods to learn the expertise as well as the ground truth of the crowdsourcing tasks. The algorithm in the current chapter differs in two main aspects: (1) The current chapter uses a different expertise model which can successfully capture the possibly hierarchical relationship between different information/expertise as well as the most valuable information; (2) the current chapter combines the algorithm with an incentive mechanism that endogenously controls the quality and structure of the input, rather than making exogenous assumptions about the quality of the input.

5.2 Multi-task setting

In this section, we will apply the HMIP framework in the multi-task setting where each agent receives a random batch of *a priori similar* tasks

5.2.1 Backgrounds and assumptions

In multi-task setting, the major challenge solved in previous peer prediction literature is that agents may "get something for nothing" by always answering the same answer (e.g. always saying good in peer grading).

In the setting where agents are assigned ≥ 2 tasks, Dasgupta and Ghosh [18], Kong and Schoenebeck [34], and Shnayder et al. [63] solve this challenge by assuming

agents are homogeneous and rewarding agents not only for their agreements but also for the diversity of their answers. If an agent answers the same answer all the time (no diversity), she will be paid nothing. Kong and Schoenebeck [34] show that this idea essentially means rewarding each agent MI^{tvd}(her information; her peer's information).

When agents are heterogeneous, Mandal et al. [42] ask agents to answer a sufficient number of tasks and then classify their answers into different clusters to learn their levels and pay them.

Assumption 88 (a priori similar)**.** *All tasks are a priori similar for all agents. That is, tasks are i.i.d samples for all agents. For every agent, before she invests any effort, for each m, for all tasks, she has the same prior belief for the signals she and other agents will receive by performing m.*

Prior work [63, 18, 34] also makes this assumption; however in their setting it is much stronger than in ours. For example, it insists that the only "signal" included in a prompt is for the correct answer. In reality, some false answers are more appealing than others (see Example 1 where Kansas is an unlikely answer). In our model, these appealing false answers can be modeled as "cheap" information instead of being assumed away.

Note that in the multi-task setting, we allow agents to have different priors and only require that for every agent, her prior satisfies our assumptions.

5.2.2 Known information structure and a small number of tasks

In order to avoid agents "getting something for nothing" by reporting the cheap signals instead of the expensive signals (e.g. giving a high quality grade when there are no typos in Example 2), we reward agents the information score of expensive signals according to not only their agreements but also the diversity of their answers *conditioning on the tasks which have the same cheap signals. (e.g. the essays which*

all have no typos). We will show this idea is essentially the application of HMIP framework when MI^f is chosen to be MI^{tvd}.

Assumption 89 (Positively correlated signals)**.** *We assume that for every method m, each agent i, every $\sigma \neq \sigma'$, every possible $\{\sigma^{m'}\}_{m' \prec m}$, every subset $M' \subset \{m' | m' \prec m\}$, Ψ_{-i}^m is* positively correlated *with $\Psi_i^m = \sigma$:*

$$\Pr[\Psi_{-i}^m = \sigma | \Psi_i^m = \sigma] > \Pr[\Psi_{-i}^m = \sigma],$$

$$\Pr[\Psi_{-i}^m = \sigma | \Psi_i^m = \sigma'] < \Pr[\Psi_{-i}^m = \sigma],$$

conditioning on $\{\Psi_{-i}^{m'}\}_{m' \in M'} = \{\sigma^{m'}\}_{m' \in M}$.

Dasgupta and Ghosh [18] and Shnayder et al. [63] both make this assumption as well. It means that receiving σ by performing m will increase each agent's belief for how many other agents receive σ by performing m. It is a substantially weaker assumption than that agents always believe they are in the majority.

> In the peer grading example, this assumption means that for every agent, receiving ☺ for quality signal will increase her belief for the probability other agents receive ☺ for quality signal.

Assumption 90 (Conditional independence)**.** *For each agent i who performs method m_i, we assume that for every possible $\{\sigma^{m'}\}_{m' \prec m}$, every subset $M' \subset \{m' | m' \prec m\}$, for each $m \preceq m_i$, Ψ_i^m contains all information agent i has that is related to Ψ_{-i}^m, in other words, conditioning on Ψ_i^m, $\{\Psi_i^{m'}\}_{m' \preceq m_i, m' \neq m}$ are independent with Ψ_{-i}^m, conditioning on $\{\Psi_{-i}^{m'}\}_{m' \in M'} = \{\sigma^{m'}\}_{m' \in M'}$*[1].

> In the peer grading example, this assumption means that for every agent, if she has already thought the writing is good, her quality signal will not affect her opinion for the writing.

[1]Note that if agents receive the same signal by performing the same method, both Assumption 89 and Assumption 90 will hold.

With this assumption, when an agent needs to report her information that is related to Ψ_{-i}^m, assuming she has already performed method m, it's sufficient for her to only report Ψ_i^m.

Multi-task Hierarchical Mutual Information Mechanism (Multi-HMIM($\{\alpha_m\}_m$))

Report Each agent i is assigned a random batch of tasks (at least two). For each task t which is assigned to agent i, she is asked to report both the method $m_i(t)$ she performed on task t and method $m_i(t)$'s output $\psi_i^{m_i(t)}(t)$; for each $m \neq m_i(t)$, agent i is asked to optionally report her signal $\psi_i^m(t)$. We denote her actual report for her performed method and signal for every method m by $\hat{m}_i(t)$ and $\hat{\psi}_i^m(t)$ respectively.

Information Score For each method m, the mechanism collects agent i's method m signals and records them via a T dimensional vector $\hat{\boldsymbol{\psi}}_i^m$.

$$\text{The } t^{th} \text{ coordinate of } \hat{\boldsymbol{\psi}}_i^m \text{ is } \begin{cases} \hat{\psi}_i^m(t), & \text{if agent } i \text{ provides the} \\ & \text{method } m\text{'s output } \hat{\psi}_i^m(t) \\ & \text{for task } t; \\ \emptyset, & \text{otherwise} \end{cases}$$

We define $\hat{\boldsymbol{\psi}}_{-i}^m$ as a vector where the t^{th} coordinate of $\hat{\boldsymbol{\psi}}_{-i}^m$ is

$$\begin{cases} \hat{\psi}_{-i}^m(t), & \text{we arbitrarily pick an agent (who is not agent } i\text{)} \\ & \text{whose performed method is } \succeq m \text{ for task } t \\ & \text{and provides method's } m\text{'s output for task } t; \\ & \text{we denote his report by } \hat{\psi}_{-i}^m(t); \\ \emptyset, & \text{such agent does not exist} \end{cases}$$

Agent i is paid by her information score

$$\sum_m 2\alpha_m Corr(\hat{\boldsymbol{\psi}}_i^m; \hat{\boldsymbol{\psi}}_{-i}^m | \{\hat{\boldsymbol{\psi}}_{-i}^{m'}\}_{m' \prec m})$$

and $Corr(\cdot)$[2] is a random algorithm defined in Algorithm 1.

We design $Corr(\hat{\boldsymbol{\psi}}_i^m; \hat{\boldsymbol{\psi}}_{-i}^m | \{\hat{\boldsymbol{\psi}}_{-i}^{m'}\}_{m' \prec m})$ to be an unbiased estimator of $MI^{tvd}(\hat{\Psi}_i^m; \hat{\Psi}_{-i}^m | \{\hat{\Psi}_{-i}^{m'}\}_{m' \prec m})$[3] if $\hat{\Psi}_i^m$ and $\hat{\Psi}_{-i}^m$ are *positively correlated*. Thus, in Multi-HMIM, agents are essentially paid based on the (conditional) mutual information by picking a special f-mutual information—MI^{tvd}, if agents are honest since we have assumed that agents' honest signals are positively correlated. This makes our Multi-HMIM a special case of HMIP framework.

Definition 91 (Amount of information in Multi-HMIM)**.** In Multi-HMIM, when agent i performs method m_i, the amount of information acquired with the effort is defined as

[2] $Corr(\cdot;\cdot)$ is essentially the same concept as the payment schemes in Dasgupta and Ghosh [18], Kong and Schoenebeck [34], and Shnayder et al. [63]. $Corr(\cdot;\cdot|\cdot)$ is a new concept in this chapter.

[3] In the current chapter, $\hat{\boldsymbol{\psi}}_i^m$ means vector, $\hat{\Psi}_i^m$ means random variable.

$$
\begin{aligned}
&AOI(m_i, \text{Multi-HMIM}(\{\alpha_m\}_m)) \\
:=&\sum_{t\in[T]} \max_{f_m:\Pi_{\ell\preceq m_i}\Sigma_\ell\mapsto\Sigma_m} \alpha_m MI^{tvd}(f_m(\{\Psi_i^\ell\}_{\ell\preceq m_i}); \Psi_{-i}^m|\{\Psi_{-i}^{m'}\}_{m'\prec m}).
\end{aligned}
$$

$\max_{f_m:\Pi_{\ell\preceq m_i}\Sigma_\ell\mapsto\Sigma_m}$ means agent i optimize her expected information score over all report strategies that maps her received signals ($\{\Psi_i^\ell\}_{\ell\preceq m_i}$) to her reported signal for method m.

Like we did in the analysis of HMIP, we need to guarantee that for agent i whose performed method is m_i, the amount of her received information defined by the above definition should be her optimal payment in Multi-HMIM, given that Multi-HMIM has access to all levels of honest signals reported by other agents. Note that the building block $Corr$ in our mechanism is an unbiased estimator of MI^{tvd} only if the signals are positively correlated. Thus, in order to make the above guarantee, we make an additional assumption—positively correlated guess: agents' optimal guesses for each method m's output are positively correlated with m's real output.

Assumption 92 (Positively correlated guess)**.** *For agent i whose performed method is m_i, for all m, for all subset $M' \subset \{m'|m' \prec m\}$, there exists $f^*_{m,M'}$ such that*

$$
f^*_{m,M'} \in \arg\max_{f_m:\Pi_{\ell\preceq m_i}\Sigma_\ell\mapsto\Sigma_m} MI^{tvd}(f_m(\{\Psi_i^\ell\}_{\ell\preceq m_i}); \Psi_{-i}^m|\{\Psi_{-i}^{m'}\}_{m'\in M'})
$$

and $f^*_{m,M'}(\{\Psi_i^\ell\}_{\ell\preceq m_i})$ is positively correlated with Ψ_{-i}^m.

Definition 93 (Prudent strategy in Multi-HMIM)**.** For each agent i, we say she plays prudent strategy in Multi-HMIM($\{\alpha_m\}_m$) if she (a) performs method m_i^* for all her tasks such that

$$
m_i^* = \arg\max_{m_i} \left(AOI(m_i, \text{Multi-HMIM}(\{\alpha_m\}_m)) - h_i(m_i)\right);
$$

(b) reports her method m_i^* honestly and reports her all received signals honestly for all her tasks.

Definition 94 (Potent coefficients for Multi-HMIM)**.** Given the priors $\{Q_m\}_m$, we say the coefficients $\{\alpha_m\}_m$ are potent for Multi-HMIM($\{\alpha_m\}_m$) if given the coefficients $\{\alpha_m\}_m$, for every maximal m, for every task t, *among the agents who are assigned task* t, there exists at least **two** agents whose prudent strategy in Multi-HMIM($\{\alpha_m\}_m$) are performing method m.

Definition 95 (Truthful strategy in Multi-HMIM)**.** For each agent i, we say she plays truthful strategy if for each task t, she honestly report her method $m_i(t)$ for task t and for each $m \prec m_i(t)$, either she chooses to not report or she reports honestly.

We allow agents to guess the signals they did not receive. Thus, in the definition of prudent strategy and truthful strategy, we only require agents to honestly report the signals they receive and do not put any restriction on their guesses.

Here we propose a new mechanism design goal: we say a mechanism is (strictly) *truthful* if for each agent, when she believes other agents play a truthful strategy, she can (strictly) maximize her expected utility by playing a truthful strategy.

The truthful property is incomparable with the potent property. A potent mechanism incentivizes the efforts of agents but it requires agents to believe other agents play prudent strategy. A truthful mechanism may not be able to incentivize efforts of agents but it incentivizes truthful report by only requiring agents to believe other agents either report honestly or choose to not report.

Theorem 96. *With Assumption 88, 89, 90, Multi-HMIM($\{\alpha_m\}_m$) is truthful; moreover, when $\{\alpha_m\}_m$ are potent for Multi-HMIM($\{\alpha_m\}_m$), Multi-HMIM($\{\alpha_m\}_m$) is potent and truthful.*

In order to show the truth property of Multi-HMIM, we will show for each agent, given that other agents play truthful strategy, (1) conditioning on using pure effort

strategy, she can maximize her payment as well as her utility by reporting all her received information honestly; (2) pure effort strategy gives her better utility than mixed effort strategy. We can apply Theorem 43 directly and use the information monotonicity of MI^{tvd} to prove part (1) directly. In order to show part (2), we need to solve the *mixed effort strategy problem* in the multi-task setting—agents put high level effort only for partial number of tasks but claim that they spend high level effort all the time. Note that even though agents can expend lower effort in randomizing between performing a low level method and a high level method than purely performing high level method, they also obtain lower payment since they have less "agreement" with high level information provided by other people. It turns out that the convexity of the f-mutual information—including MI^{tvd}—implies that agents cannot obtain higher *utility*—which is the payment minus the cost—by playing a mixed effort strategy. The potent property immediately follows from the truthful property and the condition that the coefficients are potent . We defer the formal proof to appendix.

Algorithm 1: Building Block $Corr$

procedure $Corr(\boldsymbol{v}_1; \boldsymbol{v}_2)$ ▷ e.g. $\boldsymbol{v}_1 = (☺, \emptyset, ☺, ☺, ☺)$, $\boldsymbol{v}_2 = (☺, ☺, ☺, ☺, \emptyset)$

  if either $\boldsymbol{v}_1$ or $\boldsymbol{v}_2$ has fewer than two non-empty entries **then return** 0

  else

    $B \subset [M] \leftarrow$ the set of entries where both $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are not empty ▷ $B \leftarrow \{1, 3, 4\}$

    if $B = \emptyset$ **then return** 0

    else

      for $t_B \in B$ **do** ▷ *We call* t_B *a* ***reward task***

        $v_1(t_1) \leftarrow$ a random non-empty entry in $\boldsymbol{v}_1$

        ▷ $v_1(t_1) \leftarrow ☺$

        $v_2(t_2) \leftarrow$ a random non-empty entry in $\boldsymbol{v}_2$, $t_2 \neq t_1$ ▷ $v_2(t_2) \leftarrow ☺$

        $Corr_{t_B} \leftarrow \mathbb{1}(v_1(t_B) = v_2(t_B)) - \mathbb{1}(v_1(t_1) = v_2(t_2))$ ▷ $Corr_{t_B} \leftarrow 0$

      return $\sum_{t_B \in B} Corr_{t_B}$ and "success" ▷ *Return 0*

---

procedure $Corr(\boldsymbol{v}_1; \boldsymbol{v}_2 | V)$ ▷ e.g. $\boldsymbol{v}_1 = (☺, ☺, ☹, ☺, ☺)$, $\boldsymbol{v}_2 = (☺, ☺, ☹, ☺, ☹)$, $V = \{v\}$, $v = (☺, ☺, ☹, ☺, ☹)$

  $C \leftarrow$ the set of entries where every $v \in V$ is not empty ▷ $C \leftarrow \{1, 2, 3, 4, 5\}$

  if $C = \emptyset$ **then return** $Corr(\boldsymbol{v}_1; \boldsymbol{v}_2)$

  else

    $t^*_C \leftarrow$ a random element in C ▷ $t^*_C \leftarrow 2$

    $D \leftarrow \emptyset$

    for $t \in [T]$ **do**

      if for every $v \in V$, $v(t) = v(t^*_C)$ **then**

        put t in D

        ▷ $D = \{1, 2, 4\}$, $\boldsymbol{v}_1(D) = \boldsymbol{v}_2(D) = (☺, ☺, ☺)$

    return $Corr(\boldsymbol{v}_1(D); \boldsymbol{v}_2(D))$

    ▷ Return $Corr(\boldsymbol{v}_1(D); \boldsymbol{v}_2(D)) = 0$ and "success"

5.2.3 Learning information structure with a large number tasks

Assumption 97 (δ_0-gap)**.** *For each m, we assume that for every $i \neq j$, each $m' \neq m$*

$$MI^f(\Psi_i^m; \Psi_j^m) > \frac{1}{\delta_0} \qquad MI^f(\Psi_i^m; \Psi_j^{m'}) < \frac{1}{\delta_0}$$

The above assumption guarantees that when we can accurately learn the f-mutual information between two agents' answer vectors, we can accurately classify the answer vectors and then learn the maximal method's outputs correctly.

Learning based Multi-HMIM($\mathcal{RULE}$)

Report Each agent i is assigned T tasks and asked to perform the same method for all tasks. For agent i who performs method m_i, she is asked to report her own answer vector

$$\boldsymbol{\psi}_i^{m_i} = (\psi_i^{m_i}(1), \psi_i^{m_i}(2), ..., \psi_i^{m_i}(T))$$

and, for each method $m \neq m_i$, is asked to optionally report her answer vector $\boldsymbol{\psi}_i^m$. We denote the set of methods whose outputs are reported by agent i as M_i and the actual answer vector she reports for each method $\ell \in M_i$ as $\hat{\boldsymbol{\psi}}_i^\ell$. Agent i can name the methods freely[4].

Learning Information Structure We define the distance between $\hat{\boldsymbol{\psi}}_i^m$ and $\hat{\boldsymbol{\psi}}_j^{m'}$ as $\frac{1}{MI^f(\hat{\Psi}_i^m; \hat{\Psi}_j^{m'})}$. The mechanism starts to cluster answer vectors. A set of answer vectors are clustered into one cluster if and only if their pairwise distance is less than δ_0. A cluster may have ≥ 1 answer vector(s).

For two clusters m_1, m_2, $m_1 \succ m_2$ if and only if there exists an agent who's own answer vector is in cluster m_1 and also provides an answer vector which is

[4]The mechanism will ignore the name of the methods and only record the relationship that the other answer vectors reported by agent i have lower level than agent i's own answer vector.

classified in cluster m_2. The mechanism picks positive real values for the type payment scale α_m according to a rule $\mathcal{RULE}$.

Information Score The mechanism learns the information structure using all agents' reports excluding agent i. We denote the set of clusters by M_{-i}. For each cluster $m \in M_{-i}$, the mechanism randomly picks an answer vector, denoted $\hat{\boldsymbol{\psi}}_{-i}^{m}$, from it.

Agent i is paid her information score:

$$\sum_{m \in M_{-i}} \alpha_m MI^f(\{\hat{\Psi}_i^{\ell}\}_{\ell \in M_i}; \hat{\Psi}_{-i}^{m} | \{\hat{\Psi}_{-i}^{m'}\}_{m' \prec m, m' \in M_{-i}})$$

which can be calculated accurately when the number of tasks is large.

We define $\boldsymbol{\alpha}(\mathcal{RULE}) := \{\alpha_m(\mathcal{RULE})\}_m$ as the coefficients determined by $\mathcal{RULE}$, given that the mechanism has access to all levels of honest answer vectors. Here the amount of information and prudent strategy are defined by the same way in HMIP, except that the coefficients are $\boldsymbol{\alpha}(\mathcal{RULE})$.

Definition 98 (Prudent strategy in learning based Multi-HMIM)**.** For each agent i, we say she plays a prudent strategy in learning based Multi-HMIM($\mathcal{RULE}$) if she chooses to (a) perform method m_i^* such that

$$m_i^* \in \arg\max_{m_i} \left(AOI(m_i, \text{HMIP}(MI^f, \{\alpha_m(\mathcal{RULE})\}_m)) - h_i(m_i) \right);$$

(b) report all received information honestly.

We also define potent $\mathcal{RULE}$ such that $\boldsymbol{\alpha}(\mathcal{RULE})$ is potent in the definition in HMIP.

Definition 99 (Potent rule for learning based Multi-HMIM)**.** Given the priors $\{Q_m\}_m$, we say the rule $\mathcal{RULE}$ that determines the coefficients is potent for learning based

Multi-HMIM($\mathcal{RULE}$) if given $\mathcal{RULE}$, for every maximal m, there exists at least **two** agents whose prudent strategy in learning based Multi-HMIM($\mathcal{RULE}$) are performing method m.

Theorem 100. *With Assumption 88, Learning based multi-HMIM is dominant truthful.*

Moreover, with Assumption 97, when the rule $\mathcal{RULE}$ is potent , Learning based multi-HMIM is potent , dominant truthful and will output the hierarchical information structure as well as the maximal level(s) answer vector given that agents play prudent strategy.

Learning based multi-HMIM can be mapped to HMIP since when we have a large number of tasks, we can calculate the mutual information directly by first calculating the joint distribution over agents' answers. We can also learn the information structure based on the gap assumption (Assumption 97) and cluster the agents correctly. Note that even if the mechanism clusters incorrectly, the mechanism is still dominant truthful since even each agent is paid by the mutual information between her information and "wrong" information, the information monotonicity still incentivize the agent to report all information she has. Thus we do not need the gap assumption for the dominant truthfulness. With the gap assumption, we can cluster agents correctly and use Theorem 43 to show the potent property. We defer the formal proof to appendix. Moreover, we want to emphasize that our mechanisms work even if every agent only has a piece of correct information for the information structure.

5.3 Single-task setting

In this section, we apply the HMIM framework to the single task setting.

5.3.1 Backgrounds and assumption

In the single task setting without known prior, previous peer prediction works all assume the common prior assumption and follow the framework proposed in Prelec [51]—asking agents not only her signal but also her prediction. In order to achieve truthfulness for ≥ 3 agents, Radanovic and Faltings [55] and Kong and Schoenebeck [35] punish each agent if her predictions differs from the predictions of other agents who report the same signals with her, and reward each agent for the accuracy of her prediction. Therefore, for each agent, in order to maximize her accuracy reward, she will honestly report her predictions. In order to avoid the punishment for the "inconsistency", she will honestly report her received signals as well because of the following commonly assumed assumption.

Assumption 101 (common prior and stochastic relevance)**.** *We assume that for every two agents agent i and agent j, they will have the same belief for the distribution of the signals received by other agents if and only if they receive the same signals.*

5.3.2 Applying HMIP in the single-task setting

We naturally follow the previous "signal-prediction" framework and "punishing inconsistency" idea in the hierarchical information case. We ask agents their received signals and predictions for different levels. We pay each agent the accuracy of her forecasts. The high expertise agents have accurate predictions for even high cost information reports while the low expertise agents only has accurate prediction for low cost information. Therefore, high expertise agents will be paid more.

Single-HMIM$(PS, \{\alpha_m\}_m)$

Report (signals, predictions) Each agent i who performs method m_i is asked to report her received signals $\{\sigma_i^m\}_{m \preceq m_i}$ and her forecast $p_i^{m_i}$ for $\Psi_{-i}^{m_i}$. For each $m \neq m_i$, she is asked to optionally report her forecast p_i^m for Ψ_{-i}^m. We denote

her report for her received signals as $\{\hat{\sigma}_i^m\}_{m \preceq \hat{m}_i}$ and her prediction report as $\{\hat{p}_i^m\}_{m \in M_i}$ where M_i is the set of methods whose outputs are predicted by agent i.

Prediction Score We define M_{-i} as the set of methods whose outputs are reported by an agent who is not agent i. For each $m \in M_{-i}$, we pick an arbitrary reference agent $j \neq i$ whose performed method is higher than m and denote his report for method m's output by $\hat{\sigma}^m$. Agent i's prediction score is $\sum_{m \in M_{-i} \cap M_i} \alpha_m PS(\sigma^m, \hat{p}_i^m)$.

Information Score If there is no other agent who reports the same signals as agent i, then agent i's information score is 0. Otherwise, arbitrarily pick a reference agent $j \neq i$ from the agents who report the same signals as agent i. Agent i's information score is minus the inconsistency between her prediction report and agent j's prediction report, that is,

$$-\left(\sum_{m \in M_i \cap M_j} \alpha_m (PS(\hat{p}_j^m, \hat{p}_j^m) - PS(\hat{p}_j^m, \hat{p}_i^m))\right).$$

In Single-HMIM, the payment of each agent is

$$\alpha * \text{Information Score} + \beta * \text{Prediction Score}.$$

Definition 102 (Truthful strategy in Single-HMIM)**.** For each agent i whose performed method is m_i, we say she plays truthful strategy if she honestly report her received signals $\{\sigma_i^m\}_{m \preceq m_i}$ and her forecast for $\Psi_{-i}^{m_i}$ and for each $m \neq m_i$, either she chooses to not report or she reports her forecast for Ψ_{-i}^m honestly.

We denote $p_{m_i}^m$ as agent i's honest forecast for Ψ_{-i}^m given that she performs method m_i.

Definition 103 (Amount of information in Single-HMIM)**.** For each agent i who performs method m_i, her acquired amount of information is defined as

$$AOI(m_i, \text{Single-HMIM}(PS, \{\alpha_m\}_m)) := \sum_m \alpha_m \mathbb{E}_{Q_m}[PS(\sigma^m, p^m_{m_i})].$$

Later in the proof of Theorem 106, we will see the amount of information is also the optimal payment of agent i who performs method m_i in Single-HMIM, given that Single-HMIM has access to all levels of honest signals reported by other agents.

Definition 104 (Prudent strategy in Single-HMIM)**.** For each agent i, we say she plays a prudent strategy in Single-HMIM$(PS, \{\alpha_m\}_m)$ if she chooses to (a) perform method m_i^* such that

$$m_i^* \in \arg\max_{m_i} \left(AOI(m_i, \text{Single-HMIM}(PS, \{\alpha_m\}_m)) - h_i(m_i)\right);$$

(b) play a truthful strategy.

Definition 105 (Potent coefficients for Single-HMIM)**.** Given the priors $\{Q_m\}_m$, we say the coefficients $\{\alpha_m\}_m$ are potent for Single-HMIM$(PS,\{\alpha_m\}_m)$ if given the coefficients $\{\alpha_m\}_m$, for every maximal m, there exists at least **two** agents whose prudent strategy in Single-HMIM$(PS,\{\alpha_m\}_m)$ is performing method m.

Recall that a mechanism is (strictly) *truthful* if for each agent, when she believes other agents play a truthful strategy, she can (strictly) maximize her expected utility by playing a truthful strategy.

Theorem 106. *With Assumption 101, single-HMIM is strictly truthful; moreover, when the coefficients is potent for single-HMIM, single-HMIM is potent and strictly truthful.*

The strictly truthful property follows from the common prior and stochastic relevance assumption using a proof similar to that in Radanovic and Faltings [55] and

[35]. The potent property follows from the definition of prudent strategy and potent coefficients. We defer the formal proof to appendix.

CHAPTER VI

Forecast Elicitation and An Information Aggregation Problem: Co-training

6.1 Related work

Learning Co-training/multiview learning is a problem that asks to aggregate two views of data into a prediction for the latent label and was first proposed by Blum and Mitchell [9] and explored by many works (e.g. Dasgupta, Littman, and McAllester [19] and Collins and Singer [14]). Xu, Tao, and Xu [73] and Li, Yang, and Zhang [39] give surveys on this literature. Although co-training is an important learning problem, it lacks a unified theory and a solid theoretic guarantee for the general model. Most traditional co-training methods usually require additional restrictions on the hypothesis space (e.g. weakly good hypotheses) to address the "naive agreement" issue and fail to deal with soft hypotheses whose output is not a discrete signal and thus cannot fully aggregate the two sources of information. Becker [7] deals with a feature learning problem which is very similar to the co-training problem. Becker [7] designs the optimization goal as maximizing the Shannon mutual information between the outputs of two functions. However, Becker [7] only considers hard (not soft) hypotheses and lacks a solid theoretic analysis for the maximizer. Kakade and Foster [30] consider the multi-view regression and maximize the correlation between

the two hypotheses. Their method captures the "mutual information" idea (in fact, correlation is a special f-mutual information [34]) but their model has a very specific set up and the analysis cannot be extended to other co-training problems.

In contrast, we propose a simple, powerful and general information theoretic framework, f-mutual information gain, that has a solid theoretic guarantee, works for soft hypothesis and addresses the "naive agreement" issue without any additional assumption.

Natarajan et al. [46], Sukhbaatar and Fergus [64] and many other work (e.g. [6, 33, 62]) consider the learning with noisy labels problem. Natarajan et al. [46] consider binary labels and calibrate the original loss function such that the Bayesian posterior predictor that forecasts ground truth Y is a maximizer of the calibrated loss. Sukhbaatar and Fergus [64] extend this work to the multiclass setting. These works require additional estimation steps to learn the transition probability that transits the ground truth labels to the noisy labels and fix this transition probability in their calibration step. In contrast, by mapping this problem into our framework (Section 6.6.4), we do not need the additional estimation steps to make the calibrated forecaster part of a maximizer of our optimization problem, and can incorporate any kind of side information to learn the calibrated forecaster and true transition probability simultaneously.

Moreover, our results can handle more complicated setting where each instance is labeled by multiple labels. Rather than preprocessing the labels by a particular algorithm (e.g. majority vote, weighted average, spectral method) and assuming some information structure model among the crowds [56], our framework is model-free and can learn the best calibrated forecaster (predictor P_A) and the best processing algorithm (predictor P_B) simultaneously.

Raykar et al. [57] also *jointly* learn the calibrated forecaster and the distribution over the crowd-sourced feedback and ground truth labels. Raykar et al. [57] uses

the maximum likelihood estimator and assumes a simple generative model for the distribution over the crowdsourced feedback and the ground truth labels, which is conditioning the ground truth label, the crowdsourced feedback is drawn from a binomial distribution, while our framework is model-free. We also extend the maximum likelihood estimator method in Raykar et al. [57] to a general family of estimators, PS-gain estimators, based on the family of proper scoring rules, which also *jointly* learn the calibrated forecaster and the distribution. We will show the range of applications of PS-gain is more limited compared with the range of applications of f-mutual information gain (see Section 6.6.3 for more details). Cid-Sueiro [13] also uses proper scoring rules to design the loss functions that address the learning with noisy labels problem. However, Cid-Sueiro [13] designs a different family of loss functions from the PS-gain and cannot jointly learn the calibrated forecaster and the distribution.

Generative Adversarial Networks (GAN) [26] combine game theory and learning theory to make innovative progress. We also combine game theory and learning theory by proposing a peer prediction game between two predictors. The game in GAN is a zero-sum competitive game while the game in the current chapter is collaborative.

Several learning problems (e.g. finding the pose of an object in an image [8], blind source separation [12], feature selection [50]) use mutual information maximization (infomax) as their optimization goal. Some of these problems require data labeled with ground truth and some of them have a very different problem set up than our work.

We borrow the techniques about the duality of f-divergence from Nguyen, Wainwright, and Jordan [49, 48]. Nguyen, Wainwright, and Jordan [49] show a correspondence between the f-divergence and the surrogate loss in the *binary supervised learning* setting and Nguyen, Wainwright, and Jordan [48] propose a way to estimate the f-divergence between two high dimensional random variables. We apply the duality of f-divergence to an unsupervised learning problem and not restricted to the

binary setting.

We also differ from the crowdsourcing literature that infers ground truth answers from agents' reports (e.g. [76, 32, 75, 17]) in the sense that their agents' reports are a simple choice (e.g. A, B, C, D) while in our setting, the report can come from a space larger than the space of ground truth answers, perhaps even a very high dimensional vector.

Mechanism design Our mechanism design setting differ from the traditional peer prediction literature (e.g.[45, 51, 18, 34, 63, 42]) since we are eliciting forecast rather than a simple signal. We can discretize the forecast report and apply the traditional peer prediction literature results. However, this will only provide approximated truthfulness and fail to design focal mechanisms which pay truth-telling *strictly* better than any other non-permutation equilibrium since the forecast is discretized, while our mechanisms are focal for $\geq$2 tasks setting.

Witkowski et al. [72] consider the forecast elicitation situation and assume that they have an unbiased estimator of the optimal forecast while we assume an additional conditional independence assumption but do not need the unbiased estimator.

Liu and Chen [41] connect mechanism design with learning by using the learning methods to design peer prediction mechanisms. In the setting where several agents are asked to label a batch of instances, Liu and Chen [41] design a peer prediction mechanism where each agent is paid according to her answer and a reference answer generated by a classification algorithm using other agents' reports. Instead of using learning methods to design the peer prediction mechanisms, our work uses peer prediction mechanism design techniques to address a learning problem. Moreover, our mechanism design problem has a very different set up from theirs. Agarwal and Agarwal [1] connect learning theory with information elicitation by showing the equivalence between the calibrated surrogate losses in *supervised* learning and the eliciting

of some certain properties of the underlying conditional label distribution. Both our learning problem and mechanism design problem have a very different set up from theirs.

Independent work Like the current chapter, McAllester [43] also use Shannon mutual information to propose an information theoretic training objective that can deal with soft hypotheses/classifiers. We use a more general information measure, f-mutual information, which has Shannon mutual information as a special case, and we also propose an innovative connection between co-training and peer prediction.

6.2 Preliminaries

Given a finite set $[N] := \{1, 2, ..., N\}$, for any function $\phi : [N] \mapsto \mathbb{R}$, we use $(\phi(y))_{y\in[N]}$ to represent the vector $(\phi(1), \phi(2), ..., \phi(N)) \in \mathbb{R}^N$.

6.2.1 f-divergence, f-mutual information and Fenchel's duality

f-divergence [3, 16] Recall that f-divergence $D_f : \Delta_\Sigma \times \Delta_\Sigma \mapsto \mathbb{R}$ is a non-symmetric measure of the difference between distribution $\mathbf{p} \in \Delta_\Sigma$ and distribution $\mathbf{q} \in \Delta_\Sigma$ and is defined to be

$$D_f(\mathbf{p}, \mathbf{q}) = \sum_{\sigma\in\Sigma} \mathbf{p}(\sigma) f\left(\frac{\mathbf{q}(\sigma)}{\mathbf{p}(\sigma)}\right)$$

where $f : \mathbb{R} \mapsto \mathbb{R}$ is a convex function and $f(1) = 0$.

Definition 107 (Fenchel Duality [60])**.** Given any function $f : \mathbb{R} \mapsto \mathbb{R}$, we define its convex conjugate $f^\star$ as a function that also maps $\mathbb{R}$ to $\mathbb{R}$ such that

$$f^\star(x) = \sup_t tx - f(t).$$

Fact 108 (Dual version of f-divergence [49, 48])**.**

$$D_f(\mathbf{p},\mathbf{q}) \geq \sup_{u\in\Sigma} \mathbb{E}_{\mathbf{p}} u - \mathbb{E}_{\mathbf{q}} f^\star(u) = \sup_{u\in\mathcal{G}} \sum_\sigma u(\sigma)\mathbf{p}(\sigma) - \sum_\sigma f^\star(u(\sigma))\mathbf{q}(\sigma)$$

where $\mathcal{G}$ is a set of functions that maps $\sigma \in \Sigma$ to $\mathbb{R}$.

The equality holds if and only if $u(\sigma) = u^*(\sigma) \in \partial f(\frac{\mathbf{p}(\sigma)}{\mathbf{q}(\sigma)})$.

We call $(u^*, f^\star(u^*))$ *a pair of best distinguishers.*

We define $K(X = x, Y = y)$ as the ratio between $U_{X,Y}(x, y)$ and $V_{X,Y}(x, y)$—

$$K(X = x, Y = y) := \frac{\Pr[X = x, Y = y]}{\Pr[X = x]\Pr[Y = y]} = \frac{\Pr[Y = y|X = x]}{\Pr[Y = y]} = \frac{\Pr[X = x|Y = y]}{\Pr[X = x]}.$$

$K(X = x, Y = y)$ represents the "**pointwise mutual information**(PMI)" between $X = x$ and $Y = y$.

Fact 108 directly implies:

Fact 109 (Dual version of f-mutual information)**.**

$$MI^f(X;Y) \geq \sup_{u\in\mathcal{G}} \mathbb{E}_{U_{X,Y}} u - \mathbb{E}_{V_{X,Y}} f^\star(u)$$

where $\mathcal{G}$ is a set of functions that maps (x, y) to $\mathbb{R}$.

The equality holds if and only if $u(x, y) = u^*(x, y) \in \partial f(K(X = x, Y = y))$.

6.2.2 Property of the pointwise mutual information

We will introduce a simple property of the pointwise mutual information that we will use multiple times in the future. In addition to several different formats of the pointwise mutual information (e.g. joint distribution/product of the marginal distributions, posterior/prior), if there exists a latent random variable Y such that random variable X_A and random variable X_B are independent conditioning on Y,

f-divergence	$f(t)$	$u^*(x,y) = \partial f(K(x,y))$	$f^\star(u^*(x,y))$
Total Variation Distance	$\lvert t-1\rvert$	$\text{sign}(\log K(x,y))$	$\text{sign}(\log K(x,y))$
KL divergence	$t\log t$	$1+\log K(x,y)$	$K(x,y)$
Reverse KL	$-\log t$	$-\frac{1}{K(x,y)}$	$-1+\log K(x,y))$
Pearson χ^2	$(t-1)^2$	$2(K(x,y)-1)$	$(K(x,y))^2-1$
Squared Hellinger	$(\sqrt{t}-1)^2$	$1-\sqrt{\frac{1}{K(x,y)}}$	$\sqrt{K(x,y)}-1$

Table 6.1: Reference for common f-divergences and corresponding pairs of best distinguishers $(u^*(x,y), f^\star(u^*(x,y))$ of f-mutual information.

$$K(x,y) = K(X=x, Y=y) := \frac{\Pr[X=x,Y=y]}{\Pr[X=x]\Pr[Y=y]} = \frac{\Pr[Y=y|X=x]}{\Pr[Y=y]} = \frac{\Pr[X=x|Y=y]}{\Pr[X=x]}.$$

we can also represent the pointwise mutual information between X_A and X_B by the "agreement" between the "relationship" between X_A and Y, and the "relationship" between X_B and Y.

Claim 110. When random variables X_A, X_B are independent conditioning on Y,

$$\begin{aligned} K(X_A=x_A, X_B=x_B) &= \sum_y \Pr[Y=y] K(X_A=x_A, Y=y) K(X_B=x_B, Y=y) \\ &= \sum_y \Pr[Y=y|X_A=x_A] K(X_B=x_B, Y=y) \\ &= \sum_y \frac{\Pr[Y=y|X_A=x_A]\Pr[Y=y|X_B=x_B]}{\Pr[Y=y]}. \end{aligned}$$

We defer the proof to appendix.

6.3 General Model and Assumptions

Let X_A, X_B, Y be three random variables and we define prior Q as the joint distribution over X_A, X_B, Y. We want to forecast the ground truth Y whose realization is a signal in a finite set Σ. X_A, X_B are two sources of information that are related to Y. X_A's realization is a signal in a finite set Σ_A. X_B's realization is a signal in a finite

set Σ_B. We may have access to both of the realizations of X_A and X_B or only one of them. Thus, we need to learn the relationship between X_A, X_B and Y to forecast Y. It's impossible to learn by only accessing the samples of X_A, X_B without additional assumption. We make the following conditional independence assumption:

Assumption 111 (Conditional independence)**.** *We assume that conditioning on Y, X_A, and X_B are independent.*

Intuitively, Y can be seen as the "intersection" between X_A and X_B. We call Z a *solution* if conditioning on Z, X_A, and X_B are independent. Y is a solution. However, there are a lot of solutions. For example, conditioning on X_A or X_B, X_A and X_B are independent, which means X_A and X_B are both solutions. Thus, we have additional restriction on the prior—well-defined prior and stable prior.

6.3.1 Well-defined and stable prior

We will need restrictions on the prior when we analyze the strictness of our learning algorithm/mechanism. Readers can skip this section without losing the core idea of our results.

To infer the relationship between Y and X_A, X_B with only samples of X_A, X_B, we cannot do better than to just solve the system of equations (6.1), given the joint distribution over X_A, X_B—Q. Our goal is to obtain the Bayesian posterior predictor. Thus, we list a system that the Bayesian posterior predictor satisfies. The following system equations involve variables $\{\mathbf{a}^{x_A}, \mathbf{b}^{x_B} \in \Delta_\Sigma\}_{x_A \in \Sigma_A, x_B \in \Sigma_B}$, and $\mathbf{r} \in \Delta_\Sigma$. We insist $a_y^{x_A} = \Pr[Y = y | X_A = x_A]$, $b_y^{x_B} = \Pr[Y = y | X_B = x_B]$ and $r_y = \Pr[Y = y]$ is a solution and we call it the *desired* solution.

$$\mathcal{S}(\{\mathbf{a}^{x_A},\mathbf{b}^{x_B}\}_{x_A\in\Sigma_A,x_B\in\Sigma_B},\mathbf{r}) \tag{6.1}$$
$$:=\left\{\sum_{y\in\Sigma}\frac{a_y^{x_A}b_y^{x_B}}{r_y}-K(X_A=x_A,X_B=x_B)\right\}_{x_A\in\Sigma_A,x_B\in\Sigma_B}=0$$

Claim 110 shows the above system has the desired solution.

Note that any permutation of a solution is still a valid solution[1]. Since we cannot do better than to solve the above system, if the above system only has one "unique" solution, in the sense that any two solutions are permuted version of each other, we call the prior Q a well-defined prior. Formally,

Definition 112 (Well-defined)**.** A prior Q is well-defined if for any two solutions $\{\mathbf{a}^{x_A},\mathbf{b}^{x_B}\}_{x_A\in\Sigma_A,x_B\in\Sigma_B}$, $\mathbf{r}$ and $\{\mathbf{c}^{x_A},\mathbf{d}^{x_B}\}_{x_A\in\Sigma_A,x_B\in\Sigma_B}$, $\mathbf{r}'$ of the system of equations (6.1), there exists a permutation $\pi:\Sigma\mapsto\Sigma$ such that $\mathbf{r}=\pi\mathbf{r}'$ for any x_A,x_B, $\mathbf{a}^{x_A}=\pi\mathbf{c}^{x_A}$, $\mathbf{b}^{x_B}=\pi\mathbf{d}^{x_B}$.

The well-defined prior exist since intuitively, if $|\Sigma_A|$ and $|\Sigma_B|$ are high and $|\Sigma|$ is low, it is likely Y is the "unique intersection" since the number of constraints of the system will be much greater than the number of variables.

We say a prior is stable if fixing part of the desired solution of the system (6.1), in order to make it still a solution of the system, other parts of the desired solution should also be fixed.

Definition 113 (Stable)**.** A prior Q is stable if fixing $a_y^{x_A}=\Pr[Y=y|X_A=x_A]$ and $r_y=\Pr[Y=y]$, the system (6.1) $\mathcal{S}(\{\mathbf{a}^{x_A},\mathbf{b}^{x_B}\}_{x_A\in\Sigma_A,x_B\in\Sigma_B},\mathbf{r})=0$ has unique solution $\mathbf{b}^{x_A}$ such that $b_y^{x_B}=\Pr[Y=y|X_B=x_B]$; and fixing $b_y^{x_B}=\Pr[Y=y|X_B=x_B]$ and $r_y=\Pr[Y=y]$, the system (6.1) $\mathcal{S}(\{\mathbf{a}^{x_A},\mathbf{b}^{x_B}\}_{x_A\in\Sigma_A,x_B\in\Sigma_B},\mathbf{r})=0$ has unique solution $\mathbf{a}^{x_A}$ such that $a_y^{x_A}=\Pr[Y=y|X_A=x_A]$.

[1]We may be able to distinguish a solution with its permuted version if we have some side information (e.g. the prior of Y/a few (x_A,x_B,y) samples).

We require stable priors when we design *strictly* truthful mechanisms.

6.3.2 Predictors

This section gives the definition of predictors. We have two sets of samples $S_A := \{x_A^\ell\}_{\ell \in \mathcal{L}_A}$ and $S_B := \{x_B^\ell\}_{\ell \in \mathcal{L}_B}$ which are i.i.d samples of X_A and X_B respectively. For $\ell \in \mathcal{L}_A \cap \mathcal{L}_B$, (x_A^ℓ, x_B^ℓ)s are i.i.d samples of the joint random variable (X_A, X_B).

A predictor $P_A : \Sigma_A \mapsto \Delta_\Sigma$ for X_A maps $x_A \in \Sigma$ to a forecast $P_A(x_A)$ for ground truth Y. We similarly define the predictors for X_B. We define *the Bayesian posterior predictor* as the predictor that maps any input information $X = x$ to its Bayesian posterior forecast for $Y = y$—$Pr(Y = y|X = x)$.

With the conditional independence assumption, we have

$$\begin{aligned}
\Pr[Y|X_A, X_B] =& \frac{\Pr[Y, X_A, X_B]}{\Pr[X_A, X_B]} \\
=& \frac{\Pr[Y]\Pr[X_A|Y]\Pr[X_B|Y]}{\Pr[X_A, X_B]} && \text{(conditional independence)} \\
=& \frac{\Pr[Y|X_A]\Pr[Y|X_B]}{K(X_A, X_B)\Pr[Y]} \\
& && (K(X_A, X_B) \text{ is the pointwise mutual information.})
\end{aligned}$$

When we have access to both the sources where $X_A = x_A$ and $X_B = x_B$, given the prior of the ground truth Y, we can construct an aggregated forecast for $Y = y$ using P_A, P_B:

$$\frac{P_A(x_A)P_B(x_B)}{\Pr[Y = y]} * \text{normalization}$$

In this case, if both P_A and P_B are the Bayesian posterior predictor, the aggregated forecast is the Bayesian posterior predictor as well. Thus, it's sufficient to only train P_A and P_B. In the rest sections, we will show how to train P_A and P_B (Section 6.4),

given the two sets of samples S_A and S_B, as well as how to incentivize high quality predictors from the crowds (Section 6.5).

6.4 Co-training: find the common ground truth

We have a set of candidates $\mathcal{H}_A$ for the predictor for X_A and a set of candidates $\mathcal{H}_B$ for the predictor for X_B. We sometimes call each predictor candidate *a hypothesis*. Given the two sets of samples $S_A = \{x_A^\ell\}_{\ell \in \mathcal{L}_A}$ and $S_B = \{x_B^\ell\}_{\ell \in \mathcal{L}_B}$, our goal is to figure out the best hypothesis in $\mathcal{H}_A$ and the best hypothesis in $\mathcal{H}_B$ simultaneously. Thus, we need to design proper "loss function" such that the best hypotheses minimize the loss. In fact, we will show how to design a proper "reward function" such that the best hypotheses maximize the reward.

6.4.1 f-mutual information gain

f-mutual information gain $MIG^f(R)$ (Figure 1.7)

Hypothesis We are given $\mathcal{H}_A = \{h_A : \Sigma_A \mapsto \Delta_\Sigma\}$, $\mathcal{H}_B = \{h_B : \Sigma_B \mapsto \Delta_\Sigma\}$: the set of hypotheses/predictor candidates for X_A and X_B, respectively.

Gain Given reward function $R : \Delta_\Sigma \times \Delta_\Sigma \mapsto \mathbb{R}$,

for each $\ell \in \mathcal{L}_A \cap \mathcal{L}_B$, reward "the amount of agreement" between the two predictor candidates' predictions for task ℓ—

$$R(h_A(x_A^\ell), h_B(x_B^\ell));$$

for each distinct pair $(\ell_A, \ell_B), \ell_A \in \mathcal{L}_A, \ell_B \in \mathcal{L}_B, \ell_A \neq \ell_B$, punish both predictor candidates "the amount of agreement" between their predictions for a pair of distinct tasks (ℓ_A, ℓ_B)—

$$f^\star(R(h_A(x_A^{\ell_A}), h_B(x_B^{\ell_B})).$$

The f-mutual information gain $MIG^f(R)$ that is corresponding to the reward function R is

$$MIG^f(R(h_A,h_B))_{|S_A,S_B} = \frac{1}{|\mathcal{L}_A \cap \mathcal{L}_B|} \sum_{\ell \in \mathcal{L}_A \cap \mathcal{L}_B} R(h_A(x_A^\ell), h_B(x_B^\ell))$$
$$- \frac{1}{|\mathcal{L}_A||\mathcal{L}_B| - |\mathcal{L}_A \cap \mathcal{L}_B|^2} \sum_{\ell_A \in \mathcal{L}_A, \ell_B \in \mathcal{L}_B, \ell_A \neq \ell_B} f^\star(R(h_A(x_A^{\ell_A}), h_B(x_B^{\ell_B})))$$

Lemma 114. *The expected total f-mutual information gain is maximized over all possible R, h_A, and h_B if and only if for any $(x_A, x_B) \in \Sigma_A \times \Sigma_B$,*

$$R(h_A(x_A), h_B(x_B)) \in \partial f(K(x_A, x_B)).$$

The maximum is

$$MI^f(X_A; X_B).$$

Proof. $(x_A^\ell, x_B^\ell)_\ell$ are i.i.d. realizations of (X_A, X_B). Therefore, the expected f-mutual information gain is

$$\mathbb{E}_{U_{X_A,X_B}} R - \mathbb{E}_{V_{X_A,X_B}} f^\star(R)$$

The results follow from Fact 109. □

Although any reward function corresponds to an f-mutual information gain function, we need to properly design the reward function R such that, fixing R, there exist hypotheses to maximize the corresponding f-mutual information gain $MIG^f(R)$ to the f-mutual information between the two sources. We will use the intuition from Lemma 114 to design such reward functions R in the next section.

6.4.2 Finding the common ground truth: maximizing the f-mutual information gain

In this section, we will construct a special reward function R^f and then show that the maximizers of the corresponding f-mutual information gain $MIG^f(R^f)$ are the Bayesian posterior predictors.

Definition 115 (R^f)**.** We define reward function R^f as a function that maps the two hypotheses' outputs $\mathbf{p}_1, \mathbf{p}_2 \in \Delta_\Sigma$ and the vector $\mathbf{p} \in \Delta_\Sigma$ to

$$R^f(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}) := g\bigg(\sum_y \frac{\mathbf{p}_1(y)\mathbf{p}_2(y)}{\mathbf{p}(y)}\bigg)$$

where $g(t) \in \partial f(t), \forall t$. When f is differentiable,

$$R^f(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}) := f'\bigg(\sum_y \frac{\mathbf{p}_1(y)\mathbf{p}_2(y)}{\mathbf{p}(y)}\bigg).$$

With this definition of the reward function, fixing $\mathbf{p} \in \Delta_\Sigma$ which can be seen as the prior over Y, the "amount of agreement" between two predictions $\mathbf{p}_1, \mathbf{p}_2$ are an increasing function g of

$$\sum_y \frac{\mathbf{p}_1(y)\mathbf{p}_2(y)}{\mathbf{p}(y)},$$

which is intuitive and reasonable. The increasing function g is the derivative of the convex function f. By carefully choosing convex function f, we can use any increasing function g here.

Example 116. Here we present some examples of the f-mutual information gain $MIG^f(R^f)$ with reward function R^f, associated with different f-divergences. We use Table 1 as reference for $\partial f(\cdot)$ and $f^\star(\partial f(\cdot))$.

Total variation distance:

$$\frac{1}{|\mathcal{L}_A \cap \mathcal{L}_B|} \sum_{\ell \in \mathcal{L}_A \cap \mathcal{L}_B} sign\Big(log[\sum_y \frac{h_A(x_A^\ell)(y)h_B(x_B^\ell)(y)}{\mathbf{p}(y)}]\Big)$$
$$- \frac{1}{|\mathcal{L}_A||\mathcal{L}_B| - |\mathcal{L}_A \cap \mathcal{L}_B|^2} \sum_{\ell_A \in \mathcal{L}_A, \ell_B \in \mathcal{L}_B, \ell_A \neq \ell_B} sign\Big(log[\sum_y \frac{h_A(x_A^{\ell_A})(y)h_B(x_B^{\ell_B})(y)}{\mathbf{p}(y)}]\Big)$$

KL divergence:

$$\frac{1}{|\mathcal{L}_A \cap \mathcal{L}_B|} \sum_{\ell \in \mathcal{L}_A \cap \mathcal{L}_B} \Big(1 + log[\sum_y \frac{h_A(x_A^\ell)(y)h_B(x_B^\ell)(y)}{\mathbf{p}(y)}]\Big)$$
$$- \frac{1}{|\mathcal{L}_A||\mathcal{L}_B| - |\mathcal{L}_A \cap \mathcal{L}_B|^2} \sum_{\ell_A \in \mathcal{L}_A, \ell_B \in \mathcal{L}_B, \ell_A \neq \ell_B} \Big(\sum_y \frac{h_A(x_A^{\ell_A})(y)h_B(x_B^{\ell_B})(y)}{\mathbf{p}(y)}\Big)$$

Pearson:

$$\frac{1}{|\mathcal{L}_A \cap \mathcal{L}_B|} \sum_{\ell \in \mathcal{L}_A \cap \mathcal{L}_B} 2 * \Big(\sum_y \frac{h_A(x_A^\ell)(y)h_B(x_B^\ell)(y)}{\mathbf{p}(y)} - 1\Big)$$
$$- \frac{1}{|\mathcal{L}_A||\mathcal{L}_B| - |\mathcal{L}_A \cap \mathcal{L}_B|^2} \sum_{\ell_A \in \mathcal{L}_A, \ell_B \in \mathcal{L}_B, \ell_A \neq \ell_B} \Big((\sum_y \frac{h_A(x_A^{\ell_A})(y)h_B(x_B^{\ell_B})(y)}{\mathbf{p}(y)})^2 - 1\Big)$$

Theorem 117. *With the conditional independent assumption on X_A, X_B, Y, given the samples S_A, S_B, given a convex function f, we define the optimization goal as the expected f-mutual information gain with reward function R^f —*

$$MIG^f(h_A, h_B, \mathbf{p}) := \mathbb{E}_{X_A, X_B} MIG^f(R^f(h_A, h_B, \mathbf{p}))_{|S_A, S_B}$$

and optimize over all possible hypotheses $h_A : \Sigma_A \mapsto \Delta_\Sigma$, $h_B : \Sigma_B \mapsto \Delta_\Sigma$ and distribution vectors $\mathbf{p} \in \Delta_\Sigma$. We have

Solution→Maximizer: *any solution Z corresponds to a maximizer of $MIG^f(h_A, h_B, \mathbf{p})$[2]: for any solution Z,*

$$h_A^*(x_A) := (\Pr[Z = y|X_A = x_A])_y \qquad h_B^*(x_B) := (\Pr[Z = y|X_B = x_B])_y$$[3]

and the prior over Z, $\Pr[Z = y]_y$, is the maximizer of $MIG^f(h_A, h_B, \mathbf{p})$ and the maximum is $MI^f(X_A; X_B)$;

Maximizer→(Permuted) Ground truth *when the prior is well-defined, f is differentiable, and f' is invertible, any maximizer of $MIG^f(h_A, h_B, \mathbf{p})$ corresponds to the (possibly permuted) ground truth Y: for any maximizer $(h_A^*(\cdot), h_B^*(\cdot), \mathbf{p}^*)$ of $MIG^f(h_A, h_B, \mathbf{p})$, there exists a permutation π such that*

$$h_A^*(x_A) := (\Pr[\pi(Y) = y|X_A = x_A])_y \qquad h_B^*(x_B) := (\Pr[\pi(Y) = y|X_B = x_B])_y$$

and $\mathbf{p}^* = \Pr[\pi(Y) = y]_y$.

The above theorem does not investigate computation complexity (this may be affected by the choice of f), data requirement and the choice of the hypothesis class in practical implementation. To implement our f-mutual information gain framework in practice, we implicitly assume that for high dimensional X_A, X_B, there exists a trainable set of hypotheses (e.g. neural networks) that is sufficiently rich to contain the Bayesian posterior predictor but not everything to cause over-fitting. The most apparent empirical direction will be running experiments on real data by training two neural networks to test our algorithms.

Proof for Theorem 117. Lemma 114 shows that the expected f-mutual information

[2]Given the prior over Y, we can fix $\mathbf{p}$ as the prior over Y. Without knowing the prior over Y, $\mathbf{p}$ becomes a variable of the optimization goal and helps us learn the prior over Y.

[3]Recall that we use $(\phi(y))_{y\in[N]}$ to represent the vector $(\phi(1), \phi(2), ..., \phi(N)) \in \mathbb{R}^N$.

gain is maximized if and only if for any (x_A, x_B),

$$R^f(h_A^*(x_A), h_B^*(x_B), \mathbf{p}^*) \in \partial f(K(x_A, x_B)).$$

(1) *Solution→Maximizer:* For any solution Z, we can construct

$$h_A^*(x_A) := (\Pr[Z = y|X_A = x_A])_y \qquad h_B^*(x_B) := (\Pr[Z = y|X_B = x_B])_y$$

and $\mathbf{p}^* = \Pr[Z = y]_y$. Then

$$\begin{aligned} R^f(h_A^*(x_A), h_B^*(x_B), \mathbf{p}^*) &\in \partial f\left(\sum_y \frac{\Pr[Z = y|X_A = x_A]\Pr[Z = y|X_B = x_B]}{\Pr[Z = y]}\right) \\ &= \partial f(K(x_A, x_B) \qquad \text{(Claim 110)} \end{aligned}$$

Thus, based on Lemma 114, any solution Z corresponds to a maximizer of the optimization goal.

(2)*Maximizer→(Permuted) Ground truth:* For any maximizer $(h_A^*(\cdot), h_B^*(\cdot), \mathbf{p}^*)$ of the optimization goal, when f is differentiable, Lemma 114 shows that

$$R^f(h_A^*(x_A), h_B^*(x_B), \mathbf{p}^*) = f'(K(x_A, x_B)).$$

When f' is invertible, we have

$$\sum_y \frac{h_A^*(x_A)(y)h_B^*(x_B)(y)}{\mathbf{p}^*(y)} = K(x_A, x_B)$$

for all x_A, x_B.

Thus, $\{(h_A^*(x_A), h_B^*(x_B), \mathbf{p}^*)\}_{x_A, x_B}$ is actually the solution of the system (6.1).

When the prior is well-defined, there exists a permutation π such that

$$h_A^*(x_A) := (\Pr[\pi(Y) = y | X_A = x_A])_y \qquad h_B^*(x_B) := (\Pr[\pi(Y) = y | X_B = x_B])_y$$

and $\mathbf{p}^* = \Pr[\pi(Y) = y]_y$ where Y is the ground truth.

□

6.5 Forecast elicitation without verification

In this section, we consider the setting where the predictions are provided by human beings and we want to incentivize high quality forecast by providing an instant reward without instant access to the ground truth.

There is a forecasting task. Alice and Bob have private information $X_A, X_B = x_A \in \Sigma_A, x_B \in \Sigma_B$ correspondingly and are asked to forecast the ground truth $Y = y$. We denote $(\Pr[Y = y | X_A = x_A])_y$, $(\Pr[Y = y | X_B = x_B])_y$ by $\mathbf{p}_{x_A}$, $\mathbf{p}_{x_B}$ correspondingly. Alice and Bob are asked to report their Bayesian forecast $\mathbf{p}_{x_A}$, $\mathbf{p}_{x_B}$. We denote their actual reports by $\hat{\mathbf{p}}_{x_A}$ and $\hat{\mathbf{p}}_{x_B}$. Without access to the realization of Y, we want to incentivize both Alice and Bob play *truth-telling* strategies—honestly reporting their forecast $\mathbf{p}_{x_A}$, $\mathbf{p}_{x_B}$ for Y.

We define the *strategy* of Alice as a mapping s_A from x_A (private signal) to a probability distribution over the space of all possible forecast for random variable Y. Analogously, we define Bob's strategy s_B. Note that essentially each (possibly mixed) strategy s_A can be seen as a (possibly random) predictor P_A where $P_A(x_A)$ is a random forecast drawn from distribution $s_A(x_A)$. In particular, the truthful strategy corresponds to the Bayesian posterior predictor.

We say agents play a *permutation strategy profile* if there exists permutation $\pi : \Sigma \mapsto \Sigma$ such that each agent always reports $\pi\mathbf{p}$ given her truthful report is $\mathbf{p}$.

Note that without any side information about Y, we cannot distinguish the sce-

nario where agents are honest and the scenario where agents play a permutation strategy profile. Thus, it is too much to ask truth-telling to be strictly better than any other strategy profile. The focal property defined in the following paragraph is the optimal property we can obtain.

Mechanism Design Goals

(Strictly) Truthful Mechanism $\mathcal{M}$ is (strictly) truthful if truth-telling is a (strict) equilibrium.

Focal Mechanism $\mathcal{M}$ is focal if it is strictly truthful and each agent's expected payment is maximized if agents tell the truth; moreover, when agents play a non-permutation strategy profile, each agent's expected payment is strictly less.

We consider two settings:

Multi-task Each agent is assigned several independent a priori similar forecasting tasks in a random order and is asked to report her forecast for each task.

Single-task All agents are asked to report their forecast for the same single task.

In the single-task setting, it's impossible to design focal mechanisms since agents can collaborate to pick an arbitrary $y^* \in \Sigma$ and pretend that they know $Y = y^*$. However, we will show we can design strictly truthful mechanism in the single-task setting. In the multi-task setting, since agents may be assigned different tasks and the tasks show in random order, they cannot collaborate to pick an arbitrary $y^* \in \Sigma$ for each task. In fact, we will show if the number of tasks is greater or equal to 2, we can design a family of focal mechanisms.

Achieving the focal goal in the multi-task setting is very similar to what we did in finding the common ground truth. Note that in the forecast elicitation problem, incentivizing a truthful strategy is equivalent to incentivizing the Bayesian posterior

predictor. Thus, we can directly use the f-mutual information gain as the reward in multi-task setting. Achieving the strictly truthful goal in the single-task setting is more tricky and we will return to it later.

6.5.1 Multi-task: focal forecast elicitation without verification

We assume Alice is assigned tasks set $\mathcal{L}_A$ and Bob is assigned tasks set $\mathcal{L}_B$. For each task ℓ, Alice's private information is x_A^ℓ and Bob's private information is x_B^ℓ. The ground truth of this task is y^ℓ.

Multi-task common ground mechanism $MCG(f)$ Given the prior distribution over Y, a convex and *differentiable* function f whose convex conjugate is $f^\star$,

Report for each task $\ell \in \mathcal{L}_A$, Alice is asked to report $\mathbf{p}_{x_A{}^\ell} := (\Pr[Y = y|x_A^\ell])_y$; for each task $\ell \in \mathcal{L}_B$, Bob is asked to report $\mathbf{p}_{x_B{}^\ell} := (\Pr[Y = y|x_B^\ell])_y$. We denote their actual reports by $\hat{\mathbf{p}}^\ell_{x_A{}^\ell}$ and $\hat{\mathbf{p}}^\ell_{x_B{}^\ell}$.

Payment For each $\ell \in \mathcal{L}_A \cap \mathcal{L}_B$, reward both Alice and Bob "the amount of agreement" between their forecast in task ℓ—

$$R(\hat{\mathbf{p}}^\ell_{x_A{}^\ell}, \hat{\mathbf{p}}^\ell_{x_B{}^\ell});$$

for each pair of distinct tasks $(\ell_A, \ell_B), \ell_A \in \mathcal{L}_A, \ell_B \in \mathcal{L}_B, \ell_A \neq \ell_B$, punish both Alice and Bob "the amount of agreement" between their forecast in distinct tasks (ℓ_A, ℓ_B)—

$$f^\star(R(\hat{\mathbf{p}}^{\ell_A}_{x_A{}^{\ell_A}}, \hat{\mathbf{p}}^{\ell_B}_{x_B{}^{\ell_B}}).$$

In total, both Alice and Bob are paid

$$\frac{1}{|\mathcal{L}_A \cap \mathcal{L}_B|} \sum_{\ell \in \mathcal{L}_A \cap \mathcal{L}_B} R(\hat{\mathbf{p}}^{\ell}_{x_A{}^{\ell}}, \hat{\mathbf{p}}^{\ell}_{x_B{}^{\ell}}) - \frac{1}{|\mathcal{L}_A||\mathcal{L}_B| - |\mathcal{L}_A \cap \mathcal{L}_B|^2} \sum_{\ell_A \in \mathcal{L}_A, \ell_B \in \mathcal{L}_B, \ell_A \neq \ell_B} f^{\star}(R(\hat{\mathbf{p}}^{\ell_A}_{x_A{}^{\ell_A}}, \hat{\mathbf{p}}^{\ell_B}_{x_B{}^{\ell_B}})$$

where

$$R(\mathbf{p}_1, \mathbf{p}_2) := f'(\sum_y \frac{\mathbf{p}_1(y)\mathbf{p}_2(y)}{\Pr[Y = y]}).$$

We do not want agents to collaborate with each other based on the index of the task or other information in addition to the private information. Thus, we make the following assumption to guarantee the index of the task is meaningless for all agents.

Assumption 118 (A priori similar and random order)**.** *For each task ℓ, fresh i.i.d. realizations of $(X_A, X_B, Y) = (x_A^{\ell}, x_B^{\ell}, y^{\ell})$ are generated. All tasks appear in a random order, independently drawn for each agent.*

Theorem 119. *With the conditional independence assumption, and a priori similar and random order assumption, when the prior Q is stable and well-defined, given the prior distribution over the Y, given a differential convex function f whose derivative f' is invertible, if $\max\{|\mathcal{L}_A|, |\mathcal{L}_B|\} \geq 2$, then $MCG(f)$ is focal.*

When both Alice and Bob are honest, each of them's expected payment in $MCG(f)$ is

$$MI^f(X_A; X_B).$$

Like Theorem 117, in order to show Theorem 119, we need to first introduce a lemma which is very similar with Lemma 114.

Lemma 120. *With the conditional independence assumption, the expected total payment is maximized over Alice and Bob's strategies if and only if $\forall \ell_1 \in \mathcal{L}_A, \ell_2 \in \mathcal{L}_B$,*

for any $(x_A^{\ell_1}, x_B^{\ell_2}) \in \Sigma_A \times \Sigma_B$,

$$R(\hat{\mathbf{p}}^{\ell_1}_{x_A{}^{\ell_1}}, \hat{\mathbf{p}}^{\ell_2}_{x_B{}^{\ell_2}}) = f'(K(x_A^{\ell_1}, x_B^{\ell_2})).$$

The maximum is

$$MI^f(X_A; X_B).$$

The proofs of Lemma 120 and Theorem 119 are very similar with Lemma 114 and Theorem 117. We defer the formal proofs to the appendix.

6.5.2 Single-task: strictly truthful forecast elicitation without verification

This section introduces the strictly truthful mechanism in the single-task setting. If we know the realization y of Y, we can simply apply a proper scoring rule and pay Alice and Bob $PS(y, \hat{\mathbf{p}}_{x_A})$ and $PS(y, \hat{\mathbf{p}}_{x_B})$ respectively. Then according to the property of the proper scoring rule, Alice and Bob will honestly report their truthful forecast to maximize their expected payment. However, we do not know the realization of Y. In the information elicitation without verification setting where Alice and Bob are required to report their information, Miller, Resnick, and Zeckhauser [45] propose the "peer prediction" idea, that is, pays Alice the accuracy of the forecast that predicts Bob's information conditioning Alice's information—

$$PS\Big(\hat{x}_B, (\Pr[X_B = x_B | \hat{x}_A])_y\Big)$$

where $\hat{x}_A$ and $\hat{x}_B$ are Alice and Bob's reported information. It's easy to see the peer prediction mechanism in Miller, Resnick, and Zeckhauser [45] is truthful. With a similar "peer prediction" idea, we propose a strictly truthful mechanism in forecast elicitation.

Common ground mechanism Given the prior distribution over Y,

Report Alice and Bob are required to report $\mathbf{p}_{x_A}$, $\mathbf{p}_{x_B}$. We denote their actual reports by $\hat{\mathbf{p}}_{x_A}$ and $\hat{\mathbf{p}}_{x_B}$.

Payment Both Alice and Bob are paid

$$\log \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\hat{\mathbf{p}}_{x_B}(y)}{\Pr[Y=y]}.$$

Theorem 121. *With the conditional independence assumption (and when the prior is stable), given the prior distribution over the Y, the common ground mechanism is (strictly) truthful; moreover, when both Alice and Bob are honest, each of them's expected payment in the common ground mechanism is the Shannon mutual information between their private information*

$$I(X_A; X_B) = MI^{KL}(X_A; X_B).$$

Proof. When both Alice and Bob are honest, their payment is $\log K(x_A, x_B)$ according to Claim 110. Their expected payment will be

$$\sum_{x_A,x_B} \Pr[x_A, x_B] \log K(x_A, x_B) = \sum_{x_A,x_B} \Pr[x_A, x_B] \log \frac{\Pr[x_A, x_B]}{\Pr[x_A]\Pr[x_B]} = MI^{KL}(X_A; X_B)$$

Given that Bob honestly reports $\hat{\mathbf{p}}_{x_B} = \mathbf{p}_{x_B}$, we would like to show that the expected payment of Alice is less than $MI^{KL}(X_A; X_B)$ regardless of the strategy

Alice plays. The expected payment of Alice is

$$\begin{aligned}
&\sum_{x_A,x_B} \Pr[X_A = x_A, X_B = x_B] \log \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y=y]} \\
=&\sum_{x_A,x_B} \Pr[X_A = x_A, X_B = x_B] \log \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y=y]} \Pr[X_B = x_B] \\
&- \sum_{x_A,x_B} \Pr[X_A = x_A, X_B = x_B] \log \Pr[X_B = x_B] \\
=&\sum_{x_A,x_B} \Pr[X_A = x_A, X_B = x_B] \log \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y=y]} \Pr[X_B = x_B] - C \\
&\qquad\qquad (C \text{ is a constant that does not depend on Alice's strategy}) \\
=&\sum_{x_A,x_B} \Pr[X_A = x_A] \Pr[X_B = x_B | X_A = x_A] \log \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y=y]} \Pr[X_B = x_B] - C
\end{aligned}$$

Moreover, fixing $X_A = x_A$

$$\begin{aligned}
&\sum_{x_B} \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y=y]} \Pr[X_B = x_B] \\
=&\sum_{x_B} \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y) \Pr[X_B = x_B, Y = y]}{\Pr[Y=y]} \\
=&\sum_{x_B} \sum_y \hat{\mathbf{p}}_{x_A}(y) \Pr[X_B = x_B | Y = y] \\
=&\sum_y \hat{\mathbf{p}}_{x_A}(y) = 1
\end{aligned}$$

Thus, $\sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y=y]} \Pr[X_B = x_B]$ can be seen as a forecast for $X_B = x_B$. Since $LSR(\mathbf{p}, \mathbf{q}) = \sum_\sigma \mathbf{p}(\sigma) \log \mathbf{q}(\sigma) \leq \sum_\sigma \mathbf{p}(\sigma) \log \mathbf{p}(\sigma) = LSR(\mathbf{p}, \mathbf{p})$ for any $\mathbf{p}, \mathbf{q} \in \Delta_\Sigma$, we have

$$\sum_{x_A,x_B} \Pr[X_A = x_A] \Pr[X_B = x_B | X_A = x_A] \log \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y = y]} \Pr[X_B = x_B] - C \tag{6.2}$$

$$\leq \sum_{x_A,x_B} \Pr[X_A = x_A] \Pr[X_B = x_B | X_A = x_A] \log \Pr[X_B = x_B | X_A = x_A] - C$$

$$= \sum_{x_A,x_B} \Pr[X_A = x_A] \Pr[X_B = x_B | X_A = x_A] \log \Pr[X_B = x_B | X_A = x_A]$$

$$- \sum_{x_A,x_B} \Pr[X_A = x_A, X_B = x_B] \log \Pr[X_B = x_B]$$

$$= \sum_{x_A,x_B} \Pr[X_A = x_A, X_B = x_B] \log \frac{\Pr[X_B = x_B | X_A = x_A]}{\Pr[X_B = x_B]}$$

$$= I(X_A; X_B)$$

It remains to analyze the strictness of the truthfulness. We need to show for any x_A, given that Alice receives $X_A = x_A$, she will obtain strictly less payment via reporting $\hat{\mathbf{p}}_{x_A} \neq \mathbf{p}_{x_A}$.

Given that Alice receives $X_A = x_A$, her expected payment is

$$\sum_{x_B} \Pr[X_B = x_B | X_A = x_A] \log \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y = y]} \Pr[X_B = x_B] - C$$

(see equation (6.2))

$$\leq \sum_{x_B} \Pr[X_B = x_B | X_A = x_A] \log \Pr[X_B = x_B | X_A = x_A] - C \tag{6.3}$$

Note that $\sum_\sigma \mathbf{p}(\sigma) \log \mathbf{q}(\sigma) < \sum_\sigma \mathbf{p}(\sigma) \log \mathbf{p}(\sigma)$ when $\mathbf{q} \neq \mathbf{p}$. When the prior is stable, since $\hat{\mathbf{p}}_{x_A} \neq \mathbf{p}_{x_A}$, then $\mathbf{p}_{x_B}, \hat{\mathbf{p}}_{x_A}, (\Pr[Y = y])_y$ is not the solution of system (6.1). This implies that there exists x_B such that

$$\Pr[X_B = x_B | X_A = x_A] \neq \sum_y \frac{\hat{\mathbf{p}}_{x_A}(y)\mathbf{p}_{x_B}(y)}{\Pr[Y = y]} \Pr[X_B = x_B].$$

Thus, the inequality (6.3) must be strict. Therefore, when the prior is stable, the common ground mechanism is strictly truthful.

□

6.6 PS-gain

In this section, we will extend the maximum likelihood estimator method in Raykar et al. [57] to a general family of optimization goals—PS-gain and compare the general family with our f-mutual information gain. We will see the application of PS-gain requires either one of the information sources to be low dimensional or that we have a simple generative model for the distribution over one of the information sources and ground truth label. Thus, the range of applications of PS-gain is more limited compared with the range of applications of f-mutual information gain.

In Raykar et al. [57], X_A is a feature vector which has multiple crowdsourced labels X_B. We have access to $(x_A^\ell, x_B^\ell)_{\ell\in\mathcal{L}}$ which are i.i.d samples of (X_A, X_B). Raykar et al. [57] also have the conditional independence assumption.

6.6.1 Maximum likelihood estimator (MLE)

Let Θ_A, Θ_B be two parameters that control the distribution over X_A and Y and the distribution over X_B and Y respectively.

With the conditional independence assumption, we have

$$\begin{aligned}
\log \Pr[(x_A^\ell, x_B^\ell)_{\ell\in\mathcal{L}}|\Theta_A, \Theta_B] &= \log \Pi_{\ell\in\mathcal{L}} \Pr[X_B = x_B^\ell | X_A = x_A^\ell, \Theta_A, \Theta_B] \\
&= \log \Pi_{\ell\in\mathcal{L}} \sum_y \Pr[X_B = x_B^\ell | Y = y, \Theta_B] \Pr[Y = y | X_A = x_A^\ell, \Theta_A] \\
&= \sum_{\ell\in\mathcal{L}} \log \left(\sum_y \Pr[X_B = x_B^\ell | Y = y, \Theta_B] \Pr[Y = y | X_A = x_A^\ell, \Theta_A] \right)
\end{aligned}$$

The MLE is a pair of parameters Θ_A^*, Θ_B^* that maximizes the expected

$$\log \Pr[(x_A^\ell, x_B^\ell)_{\ell\in\mathcal{L}}|\Theta_A, \Theta_B] = \sum_{\ell\in\mathcal{L}} \log\left(\sum_y \Pr[X_B = x_B^\ell|Y = y, \Theta_B]\Pr[Y = y|X_A = x_A^\ell, \Theta_A]\right).$$

Raykar et al. [57] use the MLE to estimate the parameters. In order to compare this MLE method with our f-mutual information gain framework, we map this MLE method into our language and provide a theoretical analysis for the condition when MLE is meaningful.

LSR-gain/MLE

Hypothesis We are given $\mathcal{H}_A = \{h_A : \Sigma_A \mapsto \Delta_\Sigma\}$, $\mathcal{V}_B = \{v_B : \Sigma_B \mapsto [0,1]^{|\Sigma|}\}$: the set of hypotheses candidates for X_A and X_B, respectively. Note that v_B maps $x_B \in \Sigma_B$ into a vector in $[0,1]^{|\Sigma|}$ rather than a distribution vector.

Gain We see

$$(v_B(x_B) \cdot h_A(x_A^\ell))_{x_B}$$

as a forecast for random variable X_B conditioning on $X_A = x_A^\ell$ and we reward the hypotheses LSR-gain—the accuracy of this forecast via log scoring rule (LSR):

$$\sum_{\ell\in\mathcal{L}} LSR\Big(x_B^\ell, (v_B(x_B) \cdot h_A(x_A^\ell))_{x_B}\Big) = \sum_{\ell\in\mathcal{L}} \log\Big(v_B(x_B^\ell) \cdot h_A(x_A^\ell)\Big)$$

We use $\mathbf{v} \cdot \mathbf{v}'$ to represent the dot product between two vectors.

Note that by picking $\mathcal{H}_A$ as the set of mappings—associated with a set of parameters $\{\Theta_A\}$—that map $X_A = x_A$ to $(\Pr[Y = y|X_A = x_A^\ell, \Theta_A])_y$ and picking $\mathcal{V}_B$ as the set of mappings—associated with a set of parameters $\{\Theta_B\}$—that map $X_B = x_B$ to $(\Pr[X_B = x_B|Y = y, \Theta_B])_y$, maximizing LSR-gain is equivalent to obtaining MLE.

The idea of LSR-gain is very similar with the original peer prediction idea introduced in Section 6.5.2 as well as our common ground mechanism.

Theorem 122. *When* $\sum_{x_B \in \Sigma_B} v_B(x_B) = (1, 1, .., 1)$ *for all* $v_B \in \mathcal{V}_B$*, the ground truth* Y *corresponds to a maximizer of LSR-gain:*

$$v_B^*(x_B) = (\Pr[X_B = x_B | Y = y])_y \qquad h_A^*(x_A) = (\Pr[Y = y | X_A = x_A])_y.$$

The maximum is the conditional Shannon entropy $H(X_B|X_A)$*.*

Remark 123. Note that without the restriction: $\sum_{x_B \in \Sigma_B} v_B(x_B) = (1, 1, .., 1)$ for all $v_B \in \mathcal{V}_B$,

$$v_B^*(x_B) = (\Pr[X_B = x_B | Y = y])_y \qquad h_A^*(x_A) = (\Pr[Y = y | X_A = x_A])_y$$

is not a maximizer and we will have a meaningless maximizer $v_B(x_B) = (1, 1, .., 1), \forall x_B$ and $h_A(x_A) = (1, 0, ..., 0), \forall x_A$.

By picking $\mathcal{V}_B$ as the set of mappings—associated with a set of parameters $\{\Theta_B\}$—that map $X_B = x_B$ to $(\Pr[X_B = x_B | Y = y, \Theta_B])_y$, the restriction $\sum_{x_B \in \Sigma_B} v_B(x_B) = (1, 1, .., 1)$ for all $v_B \in \mathcal{V}_B$ satisfies naturally. However, it requires the knowledge of the generative distribution model over X_B and Y with parameter Θ_B. Raykar et al. [57] assume a simple distribution model between X_B and Y with parameter Θ_B—conditioning the ground truth label, the crowdsourced feedback X_B is drawn from a binomial distribution, such that $\Pr[X_B = x_B | Y = y, \Theta_B]$ has a simple explicit form.

Proof of Theorem 122.

$$
\begin{aligned}
&\mathbb{E}\sum_{\ell\in\mathcal{L}}\log\Big(v_B(x_B^\ell)\cdot h_A(x_A^\ell)\Big)\\
=&\sum_{x_A\in\Sigma_A,x_B\in\Sigma_B}\Pr[X_A=x_A,X_B=x_B]\log\Big(v_B(x_B)\cdot h_A(x_A)\Big)\\
=&\sum_{x_A\in\Sigma_A,x_B\in\Sigma_B}\Pr[X_A=x_A]\Pr[X_B=x_B|X_A=x_A]\log\Big(v_B(x_B)\cdot h_A(x_A)\Big)\\
=&\sum_{x_A\in\Sigma_A,x_B\in\Sigma_B}\Pr[X_A=x_A]LSR\Big((\Pr[X_B=x_B|X_A=x_A])_{x_B},(v_B(x_B)\cdot h_A(x_A))_{x_B}\Big)
\end{aligned}
$$

Fixing $X_A=x_A$, since $\sum_{x_B\in\Sigma_B}v_B(x_B)=(1,1,...,1)$ for all $v_B\in\mathcal{V}_B$, we have

$$\sum_{x_B}\Big(v_B(x_B)\cdot h_A(x_A)\Big)=\sum_y h_A(x_A)(y)=1$$

Since $LSR(\mathbf{p},\mathbf{q})\le LSR(\mathbf{p},\mathbf{p})$ for any $\mathbf{p},\mathbf{q}\in\Delta_\Sigma$, we have

$$
\begin{aligned}
&\mathbb{E}\sum_{\ell\in\mathcal{L}}\log\Big(v_B(x_B)\cdot h_A(x_A)\Big)\\
=&\sum_{x_A\in\Sigma_A,x_B\in\Sigma_B}\Pr[X_A=x_A]LSR\Big((\Pr[X_B=x_B|X_A=x_A])_{x_B},(v_B(x_B)\cdot h_A(x_A))_{x_B}\Big)\\
\le&\sum_{x_A\in\Sigma_A,x_B\in\Sigma_B}\Pr[X_A=x_A]LSR\Big((\Pr[X_B=x_B|X_A=x_A])_{x_B},(\Pr[X_B=x_B|X_A=x_A])_{x_B}\Big)\\
=&\sum_{x_A\in\Sigma_A,x_B\in\Sigma_B}\Pr[X_A=x_A]\Pr[X_B=x_B|X_A=x_A]\log\Pr[X_B=x_B|X_A=x_A]\\
=&H(X_B|X_A)\\
=&\sum_{x_A\in\Sigma_A,x_B\in\Sigma_B}\Pr[X_A=x_A]\Pr[X_B=x_B|X_A=x_A]\qquad\text{(conditional independence)}\\
&\log\Big(\sum_y\Pr[X_B=x_B|Y=y]\Pr[Y=y|X_A=x_A]\Big)
\end{aligned}
$$

Thus,

$$v_B^*(x_B) = (\Pr[X_B = x_B | Y = y])_y \qquad h_A^*(x_A) = (\Pr[Y = y | X_A = x_A])_y$$

is a maximizer and the maximum is the conditional Shannon entropy $H(X_B|X_A)$.

□

6.6.2 Extending LSR-gain to PS-gain

The property $LSR(\mathbf{p}, \mathbf{q}) = \sum_\sigma \mathbf{p}(\sigma) \log \mathbf{q}(\sigma) \leq \sum_\sigma \mathbf{p}(\sigma) \log \mathbf{p}(\sigma) = LSR(\mathbf{p}, \mathbf{p})$ for any $\mathbf{p}, \mathbf{q} \in \Delta_\Sigma$ is also valid for all proper scoring rules. Thus, we can naturally extend the MLE to PS-gain by replacing the LSR by any given proper scoring rule PS.

PS-gain

Hypothesis We are given $\mathcal{H}_A = \{h_A : \Sigma_A \mapsto \Delta_\Sigma\}$, $\mathcal{V}_B = \{v_B : \Sigma_B \mapsto [0,1]^{|\Sigma|}\}$: the set of hypotheses candidates for X_A and X_B, respectively.

Gain We see

$$(v_B(x_B) \cdot h_A(x_A^\ell))_{x_B}$$

as a forecast for random variable X_B conditioning on $X_A = x_A^\ell$ and we reward the hypotheses PS-gain—the accuracy of this forecast via a given proper scoring rule PS:

$$\sum_{\ell \in \mathcal{L}} PS\Big(x_B^\ell, (v_B(x_B) \cdot h_A(x_A^\ell))_{x_B}\Big)$$

Note that the general PS-gain may involve the calculations of $(v_B(x_B) \cdot h_A(x_A^\ell))_{x_B}$ while LSR-gain only requires the value of $v_B(x_B^\ell) \cdot h_A(x_A^\ell)$. Thus, unlike LSR-gain,

the general PS-gain may be only applicable for low dimensional X_B, even if we assume a simple generative distribution model over X_B and Y.

Theorem 124. *Given a proper scoring rule PS, when $\sum_{x_B \in \Sigma_B} v_B(x_B) = (1, 1, ..., 1)$ for all $v_B \in \mathcal{V}_B$, the ground truth Y corresponds to a PS-gain maximizer:*

$$v_B^*(x_B) = (\Pr[X_B = x_B | Y = y])_y \qquad h_A^*(x_A) = (\Pr[Y = y | X_A = x_A])_y.$$

The proof is the same with Theorem 122 except that we replace $LSR(\mathbf{p}, \mathbf{q}) \leq LSR(\mathbf{p}, \mathbf{p})$ by $PS(\mathbf{p}, \mathbf{q}) \leq PS(\mathbf{p}, \mathbf{p})$ for any $\mathbf{p}, \mathbf{q} \in \Delta_\Sigma$.

6.6.3 Comparing PS-gain with f-mutual information gain

Generally, f-mutual information gain can be applied to a more general setting.

PS-gain requires the restriction $\sum_{x_B \in \Sigma_B} v_B(x_B) = (1, 1, ..., 1)$ for all $v_B \in \mathcal{V}_B$. Thus, PS-gain requires the full knowledge of v_B for all $v_B \in \mathcal{V}_B$ to check whether it satisfies the restriction, while for the f-mutual information gain, it is sufficient to just have the access to the outputs of the hypothesis: $\{h_B(x_B^\ell)\}_{\ell \in \mathcal{L}_B}$. Therefore, in the mechanism design part, we can only use f-mutual information gain to design focal mechanisms since we only have the outputs from agents.

Moreover, $\sum_{x_B \in \Sigma_B} v_B(x_B) = (1, 1, ..., 1)$ is also hard to check when $|\Sigma_B|$ is very large. For example, when x_B is a 100×100 black-and-white image, $|\Sigma_B| = 2^{100}$ and checking $\sum_{x_B \in \Sigma_B} v_B(x_B) = (1, 1, ..., 1)$ requires 2^{100} time. Normalizing v_B such that it satisfies the condition also requires 2^{100} time. Thus, when $|\Sigma_B|$ is very large, we need a simple generative distribution model between X_B and Y with parameter Θ_B such that we can pick $\mathcal{V}_B$ as the set of mappings—associated with a set of parameters $\{\Theta_B\}$—that map $X_B = x_B$ to $(\Pr[X_B = x_B | Y = y, \Theta_B])_y$, to make the restriction $\sum_{x_B \in \Sigma_B} v_B(x_B) = (1, 1, .., 1)$ for all $v_B \in \mathcal{V}_B$ satisfy naturally. When we have the simple generative distribution model, we can use LSR-gain. The general PS-gain

may involve the calculations of the $|\Sigma_B|$ dimensional vector—$(v_B(x_B) \cdot h_A(x_A^\ell))_{x_B}$—for each x_A^ℓ. Thus, the general PS-gain may be only applicable to low dimensional X_B.

In the learning with noisy labels problem, the distribution between X_B and Y can be represented by a simple transition matrix and X_B is low dimensional. Therefore, both PS-gain and f-mutual information gain can be applied to the learning with noisy labels problem.

Therefore, the application of PS-gain requires either one of the information sources to be low dimensional or that we have a simple generative model for the distribution over one of the information sources and ground truth label, while f-mutual information gain does not have the restrictions.

6.6.4 Applications

In our startup running example, we consider the situation where one source of information is the features and another source of information is the crowdsourced feedback. In fact, our results apply to all kinds of information sources. For example, we can make both sources features or crowdsourced feedback. Different setups for the information sources and predictor candidates can bring different applications of our results.

Let's consider the "learning with noisy labels" problem where the labels in the training data are a noisy version of the ground truth labels Y and the noise is independent. We can map this problem into our framework by letting X_B be the noisy label of features X_A. That is , X_B is a noisy version of Y. Our framework guarantees that the Bayesian posterior predictor that forecasts Y using X_A must be part of a maximizer of the optimization problem. However, there are many other maximizers. For example, since X_A and X_B are independent conditioning X_B. The Bayesian posterior predictor that forecasts X_B using X_A is also part of a maximizer, since the

scenario $Y = X_B$ also satisfies the conditional independence assumption. If X_B has much higher dimension than Y, we do not have this issue. But X_B has the same signal space with Y in the learning with noisy label problem. Thus, it's impossible to eliminate other maximizers without any side information here. With some side information (e.g. a candidate set $\mathcal{F}$—like linear regressions—that only contains our desired maximizer.), it's possible to obtain the Bayesian posterior predictor that forecasts Y using X_A. Note that our framework does not require a pre-estimation on the transition probability that transits the ground truth label Y to the noisy ground truth label X_B, since our framework has this transition probability, which corresponds to the predictor P_B, as parameters as well and learn the correct forecaster P_A and the transition probability P_B simultaneously.

Ratner et al. [56] propose a method to collect massive labels by asking the crowds to write heuristics to label the instances. Each instance is associated with many noisy labels outputted by the heuristics. In their setting, the crowds use a different source of information from the learning algorithm (e.g. the learning algorithm uses the biology description of the genes and the crowds use the scientific papers about the gene). Thus, the conditional independence assumption is natural here and we can map this setting's training problem into our framework. Ratner et al. [56] preprocess the collected labels to approximate ground truth by assuming a particular information structure model on the crowds. Our framework is model-free and does not need to preprocess the collected labels since we can learn the best forecaster (predictor P_A) and the best processing/aggregation algorithm (predictor P_B) simultaneously.

Moreover, since the highest evaluation value of the predictors P_A, P_B is the f-mutual information between X_A and X_B, our results provide a method to calculate the f-mutual information between any two sources of information X_A, X_B of any format. Kong and Schoenebeck [34] propose a framework for designing information elicitation mechanisms that reward truth-telling by paying each agent the f-mutual

information between her report and her peers' report. Thus, the f-mutual information gain method can be combined with this framework to design information elicitation mechanisms when the information has a complicated format.

CHAPTER VII

Conclusion and Future work

This thesis addresses two central problems in crowdsourcing, information elicitation and information aggregation, in the context where the ground truth is unknown, by distilling the essence of the central problems to the design of proper information measurements. Aided by the finding of two (weakly) information-monotone measurements—f-mutual information, Bregman-mutual information, a variety of novel information elicitation mechanisms, information aggregation algorithms are designed and a natural connection between information elicitation and information aggregation is built.

When people and information are homogeneous, this thesis proposes a simple yet powerful information theoretic paradigm—the *Mutual Information Paradigm (MIP)*—for designing information elicitation mechanisms that are truthful, focal, and, detail-free. Moreover, some of the mechanisms based on this paradigm are additionally minimal and dominantly truthful. Aided by the mutual information paradigm, this thesis exhibits two families of novel mechanisms that are *dominantly truthful*, detail free, *and* minimal in the multi-task setting when the number of questions is large—the f-mutual information mechanism and the Bregman mutual information mechanism. This thesis also employs the information theory tools in a more subtle way to exhibit the first strictly truthful, focal, detail free mechanism which applies to a small num-

ber of people in the single-question setting—Disagreement mechanism. Moreover, this thesis also unifies several important previous works by mapping them into the MIP framework.

This thesis also addresses a main problem, how to elicit expertise without verification, in crowd-sourcing situations, where agents have different levels of expertise and the lack of effort can systemically bias agents reports. This thesis creates a model of expertise based on a natural assumption that more sophisticated agents know the beliefs of less sophisticated agents. Within the model, this thesis provides a mechanism design framework the *Hierarchical Mutual Information paradigm (HMIP)* and apply HMIP in three different settings creating three mechanisms: Multi-HMIM, Learning Based Multi-HMIM, and Single-HMIM.

Finally, this thesis builds a natural connection between information elicitation and information aggregation by addressing two related problems: (1) co-training: how to learn to forecast ground truth using two conditionally independent sources, without access to any data labeled with ground truth; (2) forecast elicitation without verification: how to elicit high quality forecasts from the crowds without verification, by the same information theoretic approach, the MIP framework.

7.1 Future directions

A big goal in the future is to build an unsupervised/decentralized information trading and information aggregation system in real-world by combining this thesis's information theoretic approach, blockchain technique and unsupervised learning theory. To achieve this goal, many interesting future directions can be explored.

Experiments For the information elicitation mechanisms, the most apparent future direction is to test the mechanisms by performing real-world experiments. For example, for the expertise elicitation mechanisms, we can design the experiments by

simplifying the information structure by roughly dividing it into two levels where the higher level requires more time. Then we can test our Multi-HMIM or Single-HMIM in the peer grading scenario or any other situations where certain agents are only given 15 seconds to grade a work and others are expected to do a good job. We value the information *conditioning on* the reports provided by the "15 seconds" agents. We could also use machine learning to obtain the lower level information. For the forecast elicitation problem, to test the mechanisms proposed in this thesis, we do not need that every two agents' information is conditionally independent. In fact, for each agent, we only need to find a single reference agent for her such that the reference agent's information is conditionally independent with hers. Then we can run our mechanisms on the agent and her reference agent. In practice, we can pair the agents with some side information and make sure each pair of agents' information is conditionally independent.

For the co-training problem, as usual in the related literature, this thesis reduces the problem to an optimization problem and do not investigate the computation complexity and data requirement. The most apparent direction will be running experiments on real data by training two neural networks.

Robust to adversary In order to run the system in real-world, the existence of adversary who has other incentives must be considered. Adversarial mechanism design and adversarial algorithm design are two important future directions. Hopefully, those two problems can also be connected and addressed by the same theoretic approach.

Information cost elicitation Another future direction is the information cost elicitation: tuning the coefficients of the mechanisms such that the payment matches the actual effort required by agents pay and the cost needed to elicit high quality information is minimized.

Sample complexity, empirical risk The theoretic analysis of the sample complexity needed in calculating the information measures and the empirical risk in addressing the co-training problem are other interesting theoretic future directions. Another direction would be the analysis of the influence of the choice of the convex function f on the convergence rate.

Hopefully, after exploring the above theoretic and experimental future directions, the real-world unsupervised/decentralized information trading and information aggregation system can be built and applied to many contexts like online education (e.g. MOOC, peer grading), sharing economy, products pricing.

APPENDICES

APPENDIX A

Additional proofs

A.1 Disagreement mechanism

A.1.1 Proof for main theorem

In this section, we are going to show the first three parts of Theorem 82. Part 4 (tight property) is implied by the impossibility results in Section 2.5.

Lemma 125. *The* Disagreement Mechanism *has the same equilibria as the Divergence-based BTS.*

Theorem 82 Part 1: $\mathcal{M}+(\alpha, \beta, PS(\cdot, \cdot))$ is truthful.

Proof for Theorem 82 Part 1. Radanovic and Faltings [55] have already show $\mathcal{M}(\alpha, \beta, PS(\cdot, \cdot))$ has truth-telling as a strict equilibrium for any SNIFE prior in Theorem 81. Since $\mathcal{M}+(\alpha, \beta, PS(\cdot, \cdot))$ does not change the equilibrium structure of $\mathcal{M}(\alpha, \beta, PS(\cdot, \cdot))$ according to Claim 125, we have $\mathcal{M}+(\alpha, \beta, PS(\cdot, \cdot))$ has truth-telling as a strict equilibrium for any SNIFE prior as well. □

We finish our proof for the first part of the main theorem. For other parts, We first give technical definitions for *Diversity* and *Inconsistency* and then prove that the average agent-welfare in the *Disagreement Mechanism* is *Diversity* − *Inconsistency*.

We first introduce a short hand which will simplify the formula for *Diversity* and *Inconsistency*.

$$\int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) \triangleq \int_{\hat{\sigma}_j,\hat{\mathbf{p}}_j,\hat{\sigma}_k,\hat{\mathbf{p}}_k} Pr_{(\hat{\sigma}_j,\hat{\mathbf{p}}_j)\leftarrow s_j(\sigma_j)}(\hat{\sigma}_j,\hat{\mathbf{p}}_j) Pr_{(\hat{\sigma}_k,\hat{\mathbf{p}}_k)\leftarrow s_k(\sigma_k)}(\hat{\sigma}_k,\hat{\mathbf{p}}_k)$$

where s_j is the strategy of agent j and $s_j(\sigma_j)$ is a distribution over agent j's report profile $(\hat{\sigma}_j,\hat{\mathbf{p}}_j)$ given agent j receives private signal σ_j and uses strategy s_j, and similarly for agent k. This defines the natural measure on the reports of agents j and k given that they play strategies s_j and s_k and a fixed prior Q (which is implicit), and allows us to succinctly describe probabilities of events in this space.

We define *Diversity* as the expected Hellinger divergence D^* between two random agents when they report different signals, so

$$\textit{Diversity} = \sum_{\substack{j \\ k\neq j}} \sum_{\sigma_j,\sigma_k} Pr(j,k) Pr(\sigma_j,\sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) \delta(\hat{\sigma}_j \neq \hat{\sigma}_k) D^*(\hat{\mathbf{p}}_j,\hat{\mathbf{p}}_k)$$

where $Pr(j,k)$ is the probability agents j,k are picked, and $Pr(\sigma_j,\sigma_k)$ is the probability that agent j receives private signal σ_j and agent k receives private signal σ_k.

Similarly, we can write down the technical definition for *Inconsistency*. But here we do not use Hellinger divergence as the "difference" function in $\sum_{u,v\in U,\mathcal{C}_r(u)=\mathcal{C}_r(v)} D(u,v)$, we use square root of the Hellinger divergence which is the Hellinger distance as the "difference" function. The reason is we want to use the convexity of the Hellinger divergence and the triangle inequality of the Hellinger distance. We will describe the

details in the future. For now we give a technical definition for *Inconsistency*:

$$Inconsistency = -\sum_{\substack{j \\ k \neq j}} \sum_{\sigma_j, \sigma_k} Pr(j,k) Pr(\sigma_j, \sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) \delta(\hat{\sigma}_j = \hat{\sigma}_k) \sqrt{D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)}$$

Now we define the *ClassificationScore* as the expected average extra score $score_C$:

$$ClassificationScore = \sum_{\substack{i \\ j \neq i}} \sum_{\substack{k \neq i,j \\ \sigma_i, \sigma_j, \sigma_k}} Pr(i) Pr(\sigma_i) Pr(j,k) Pr(\sigma_j, \sigma_k | \sigma_i) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) score_C(r_j, r_k)$$

Claim 126. *ClassificationScore* = *Diversity* − *Inconsistency*.

Claim 127. Every permutation strategy profile has the same *ClassificationScore*, *Diversity*, and *Inconsistency* as truth-telling.

Claim 128. The average agent-welfare in our *Disagreement Mechanism* is the *ClassificationScore*.

A.1.1.1 Proof outline for main theorem

First note that the average agent-welfare is the *ClassificationScore*. We want to show that: if the number of agents is greater than 3, then any *symmetric* equilibrium that is not permutation equilibrium must have *ClassificationScore* strictly less than truth-telling; and any *symmetric* equilibrium that has *ClassificationScore* close to truth-telling must be close to a permutation equilibrium.

To prove our main theorem, we first introduce the concept of *TotalDivergence* and then we use this value as a bridge. Recall that we defined
$Diversity = \sum_{j,k \neq j, \sigma_j, \sigma_k} Pr(j,k) Pr(\sigma_j, \sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) \delta(\hat{\sigma}_j \neq \hat{\sigma}_k) D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)$, now we

define a similar concept

$$\textit{TotalDivergence} = \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)$$

First note that total divergence is robust to summing over j, k or $j \neq k$ since when $j = k$, $D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) = 0$.

We can see *TotalDivergence* $\geq$ *Diversity* since *TotalDivergence* also includes the divergence between the agents who report the same signals. We show that the equality holds if and only if *Inconsistency* $= 0$:

Claim 129. For any strategy profile s, *Diversity*(s) = *TotalDivergence*(s)
$\Leftrightarrow$ *Inconsistency*$(s) = 0$

Corollary 130.

$$\begin{aligned} &\textit{ClassificationScore(truth-telling)} \\ =&\textit{Diversity(truth-telling)} \\ =&\textit{TotalDivergence(truth-telling)} \end{aligned}$$

Proof. At the truth-telling equilibrium, $\hat{\sigma}_i = \sigma_i, \hat{\mathbf{p}}_i = \mathbf{q}_{\sigma_i}$ for any i, so the inconsistency score of truth-telling is 0 since $\hat{\sigma}_j = \hat{\sigma}_k \Rightarrow \sigma_j = \sigma_k \Rightarrow \hat{\mathbf{p}}_j = \hat{\mathbf{p}}_k \Rightarrow D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) = 0$ which implies this corollary. □

Now we begin to state our proof outline: For any equilibrium s, we define a modified strategy for s:

We define s_{BP} what we call a *best prediction strategy* of s as a strategy where each agent uses the same signal strategy which he uses in s but plays his *best prediction* which maximizes the prediction score. In this case, based on Claim 74, for any i, agent i plays $\theta_{-i}\mathbf{q}_{\sigma_i}$. In the symmetric case, agents i play $\theta\mathbf{q}_{\sigma_i}$.

The results or our main theorem follows from two technical lemmas:

(1) ***ClassificationScore***$(s) \leq$ ***TotalDivergence***(s_{BP})**.**[Lemma 131]. This is our main lemma and we defer the proof of main lemma to Section A.1.1.2. Once we show it, we can directly prove that the focal property of Disagreement Mechanism. Note that this result is valid for any equilibrium s—symmetric or asymmetric—and still a main ingredient when we extend the focal property to asymmetric case.

(2) ***TotalDivergence***$(truthtelling) \approx$ ***TotalDivergence***$(s_{BP}) \Rightarrow \theta \approx \pi$ when s is a symmetric equilibrium with signal strategy θ. [Lemma 136] This, informally, means that if a symmetric equilibrium pays close to truth-telling, it must be close to a permutation equilibrium, and thus pays about the same as truth-telling.

We will show *TotalDivergence*$(truthtelling)$ is

$$\sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\mathbf{q}_{\sigma_j},\mathbf{q}_{\sigma_k})$$
$$= \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta_\pi\mathbf{q}_{\sigma_j},\theta_\pi\mathbf{q}_{\sigma_k})$$

where $Pr(j,k)Pr(\sigma_j,\sigma_k)$ is the probability that agent j,k are picked and agent j receives private signal σ_j; agent k receives private signal σ_k.

We will also show *TotalDivergence*(s_{BP}) is

$\sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta\mathbf{q}_{\sigma_j},\theta\mathbf{q}_{\sigma_k})$.

A.1.1.2 Proof for main lemma

In this section, we will prove the main lemma the classification score of non-permutation equilibrium s is less than the total divergence of the report profiles when agents report their best predictions given they still use the signal strategy of s. We first show the inequality and then show that if the equality holds, then s is

consistent and $s = s_{BP}$.

Lemma 131 (Main Lemma)**.** *For any equilibrium s, if s_{BP} is a best prediction strategy of s, we have*

$$\mathit{ClassificationScore}(s) \leq \mathit{TotalDivergence}(s_{BP})$$

If the equality holds, then we have $\mathit{Inconsistency}(s) = 0$ and $s = s_{BP}$.

In order to show the inequality, we first show

$$\mathit{TotalDivergence}(s) - \mathit{TotalDivergence}(s_{BP}) \leq \mathit{Inconsistency}(s)$$

once we show this, since we have $\mathit{ClassificationScore} = \mathit{Diversity} - \mathit{Inconsistency}$ and $\mathit{Diversity} \leq \mathit{TotalDivergence}$, our main lemma $\mathit{ClassificationScore}(s) \leq \mathit{TotalDivergence}(s_{BP})$ will follow since

$$\begin{aligned}\mathit{ClassificationScore}(s) =& \mathit{Diversity}(s) - \mathit{Inconsistency}(s)\\ \leq& \mathit{TotalDivergence}(s) - \mathit{Inconsistency}(s) \leq \mathit{TotalDivergence}(s_{BP})\end{aligned} \tag{A.1}$$

To prove $\mathit{TotalDivergence}(s) - \mathit{TotalDivergence}(s_{BP}) \leq \mathit{Inconsistency}(s)$, we will write it in a explicit form:

$$\mathit{TotalDivergence}(s) - \mathit{TotalDivergence}(s_{BP}) \tag{A.2}$$

$$= \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k})(D^*(\hat{\mathbf{p}}_j,\hat{\mathbf{p}}_k) - D^*(\theta_{-j}\mathbf{q}_{\sigma_j},\theta_{-k}\mathbf{q}_{\sigma_k})) \tag{A.3}$$

It is difficult to compare $D^*(\hat{\mathbf{p}}_j,\hat{\mathbf{p}}_k)$ and $D^*(\theta_{-j}\mathbf{q}_{\sigma_j},\theta_{-k}\mathbf{q}_{\sigma_k})$ directly. To deal with

this problem, we introduce a new value $D^*(\hat{\mathbf{p}}_j, \theta_{-k}\mathbf{q}_{\sigma_k})$ and write (A.2) as

$$\sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)* \\ \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) \left(D^*(\hat{\mathbf{p}}_j,\hat{\mathbf{p}}_k) - D^*(\hat{\mathbf{p}}_j,\theta_{-k}\mathbf{q}_{\sigma_k}) + D^*(\hat{\mathbf{p}}_j,\theta_{-k}\mathbf{q}_{\sigma_k}) - D^*(\theta_{-j}\mathbf{q}_{\sigma_j},\theta_{-k}\mathbf{q}_{\sigma_k})\right) \tag{A.4}$$

We will first give the analysis for $D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) - D^*(\hat{\mathbf{p}}_j, \theta_{-k}\mathbf{q}_{\sigma_k})$, then we will see $D^*(\hat{\mathbf{p}}_j, \theta_{-k}\mathbf{q}_{\sigma_k}) - D^*(\theta_{-j}\mathbf{q}_{\sigma_j}, \theta_{-k}\mathbf{q}_{\sigma_k})$ is similar.

Remember that both $D^*(a,\cdot)$ and $D^*(\cdot,b)$ are convex functions. So $D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) - D^*(\hat{\mathbf{p}}_j, \theta_{-k}\mathbf{q}_{\sigma_k})$ can be seen as $g(\hat{\mathbf{p}}_k) - g(\theta_{-k}\mathbf{q}_{\sigma_k})$ where $g(\cdot)$ is convex function $D^*(\hat{\mathbf{p}}_j, \cdot)$.

Recall that

$$Inconsistency = \sum_{\substack{j \\ k\neq j}} \sum_{\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k})\delta(\hat{\sigma}_j = \hat{\sigma}_k)\sqrt{D^*(\hat{\mathbf{p}}_j,\hat{\mathbf{p}}_k)}$$

We hope we can obtain a upper bound for $g(\hat{\mathbf{p}}_k) - g(\theta_{-k}\mathbf{q}_{\sigma_k})$ that relates to agent k's neighbors' best response predictions. Here agent k's *neighbors* mean the agents who report the same signal with agent k and *best response prediction* means the reported prediction at equilibrium.

Now we begin to analyze the relationship between $\hat{\mathbf{p}}_k$ and $\theta_{-k}\mathbf{q}_{\sigma_k}$. Recall that each agent's payment depends on his prediction score and information score. $\theta_{-k}\mathbf{q}_{\sigma_k}$ maximizes the prediction score while $\hat{\mathbf{p}}_k$ maximizes the payment. The information score depends on agent k's neighbors' reported predictions $\{\hat{\mathbf{p}}_l | l \neq k\}$. So we can see $\hat{\mathbf{p}}_k$ is related to both his best prediction $\theta_{-k}\mathbf{q}_{\sigma_k}$ and his neighbors' reported predictions $\{\hat{\mathbf{p}}_l | l \neq k\}$. Actually we will show that $\hat{\mathbf{p}}_k$ can be computed as a linear combination of $\theta_{-k}\mathbf{q}_{\sigma_k}$ and $\{\hat{\mathbf{p}}_l | l \neq k\}$, which is based on the fact that every proper scoring rule is linear for the first entry (we will discuss the detail in the following proof). Once we

have this result, we can construct a linear system about agents' reported predictions $\{\hat{\mathbf{p}}_i|i\}$ and their best predictions. This linear system helps us obtain a upper bound for $g(\hat{\mathbf{p}}_k) - g(\theta_{-k}\mathbf{q}_{\sigma_k})$ which upper-bounds the distance between agent k's best response prediction and his neighbors' best response predictions.

Equilibrium Analysis We will analyze the equilibrium in our *Truthful Mechanism* which is also the equilibrium in our *Disagreement Mechanism*. We first show, in Claim 132, that at equilibrium, an agent's reported prediction only depends on his private signal and reported signal. Then we use this property to construct a linear system and via this linear system, we obtain a upper bound for $g(\hat{\mathbf{p}}_k) - g(\theta_{-k}\mathbf{q}_{\sigma_k})$ in Claim 133.

Claim 132. At any equilibrium $s = (s_1, ..., s_n)$, for each agent i, fix s_{-i}, agent i's private signal $\sigma_i \in \Sigma$ and reported signal $\hat{\sigma}_i \in \Sigma$, then there exists a unique prediction which is agent i's best response.

We define this unique prediction as $\hat{\mathbf{p}}(i, \sigma_i, \hat{\sigma}_i)$

In other words, $s_i(\sigma_i)$ is a distribution over at most m vectors: $\{(\hat{\sigma}_i, \hat{\mathbf{p}}(i, \sigma_i, \hat{\sigma}_i))|\hat{\sigma}_i \in \Sigma\}$ and

$$Pr_{(\hat{\sigma}_i, \hat{\mathbf{p}}_i) \leftarrow s_i(\sigma)}(\hat{\sigma}_i, \hat{\mathbf{p}}_i) = \begin{cases} \theta_i(\hat{\sigma}_i, \sigma_i) & \hat{\mathbf{p}}_i = \hat{\mathbf{p}}(i, \sigma_i, \hat{\sigma}_i) \\ 0 & \hat{\mathbf{p}}_i \neq \hat{\mathbf{p}}(i, \sigma_i, \hat{\sigma}_i) \end{cases}$$

Proof. For any agent i, assume his private signal is σ_i and he reports $\hat{\sigma}_i$ at equilibrium $(s_1, s_2, ..., s_n)$. Now we will prove there is a unique prediction that maximize agent i's payment.

$$\arg\max_{\hat{\mathbf{p}}} \mathbb{E}[payment(i, \mathcal{M}+)|\sigma_i] \tag{A.5}$$

$$= \arg\max_{\hat{p}} \alpha PS(\theta_{-i}\mathbf{q}_{\sigma_i}, \hat{\mathbf{p}}) \tag{A.6}$$

$$+ \beta \sum_{j \neq i} Pr(j) \sum_{\sigma_j} Pr(\sigma_j|\sigma_i) \int_{\hat{\sigma}_j, \hat{\mathbf{p}}_j} Pr_{(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \leftarrow s_j(\sigma_j)}(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \delta(\hat{\sigma}_i = \hat{\sigma}_j) PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}) \tag{A.7}$$

$$= \arg\max_{\hat{p}} \left(\alpha + \beta \sum_{j \neq i} Pr(j) \sum_{\sigma_j} Pr(\sigma_j|\sigma_i) \int_{\hat{\sigma}_j, \hat{\mathbf{p}}_j} Pr_{(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \leftarrow s_j(\sigma_j)}(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \delta(\hat{\sigma}_i = \hat{\sigma}_j) \right)$$
$$PS\left(\frac{\alpha\theta_{-i}\mathbf{q}_{\sigma_i} + \beta \sum_{j \neq i} Pr(j) \sum_{\sigma_j} Pr(\sigma_j|\sigma_i) \int_{\hat{\sigma}_j, \hat{\mathbf{p}}_j} Pr_{(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \leftarrow s_j(\sigma_j)}(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \delta(\hat{\sigma}_i = \hat{\sigma}_j) \hat{\mathbf{p}}_j}{\alpha + \beta \sum_{j \neq i} Pr(j) \sum_{\sigma_j} Pr(\sigma_j|\sigma_i) \int_{\hat{\sigma}_j, \hat{\mathbf{p}}_j} Pr_{(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \leftarrow s_j(\sigma_j)}(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \delta(\hat{\sigma}_i = \hat{\sigma}_j)}, \hat{\mathbf{p}}\right) \tag{A.8}$$

$$= \frac{\alpha\theta_{-i}\mathbf{q}_{\sigma_i} + \beta \sum_{j \neq i} Pr(j) \sum_{\sigma_j} Pr(\sigma_j|\sigma_i) \int_{\hat{\sigma}_j, \hat{\mathbf{p}}_j} Pr_{(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \leftarrow s_j(\sigma_j)}(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \delta(\hat{\sigma}_i = \hat{\sigma}_j) \hat{\mathbf{p}}_j}{\alpha + \beta \sum_{j \neq i} Pr(j) \sum_{\sigma_j} Pr(\sigma_j|\sigma_i) \int_{\hat{\sigma}_j, \hat{\mathbf{p}}_j} Pr_{(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \leftarrow s_j(\sigma_j)}(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \delta(\hat{\sigma}_i = \hat{\sigma}_j)} \tag{A.9}$$

In equation (A.6), the first part is the prediction score of agent i, the second part is part of the information score of agent i. Note that for the information score $PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}) - PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_j)$ of agent i, only $PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}})$ is related to agent i's reported prediction $\hat{\mathbf{p}}$ so we only consider this part to analyze the equilibrium. $Pr(j)$ is the probability that agent j is matched with agent i, $Pr(\sigma_j|\sigma_i)$ is the probability that agent j receives σ_j given agent i receives σ_i. Then given agent j's strategy s_j and private signal, we integrate over agent j possible report profiles and only consider the case $\hat{\sigma}_i = \hat{\sigma}_j$.

The second equality follows since proper scoring rule is linear for the first entry.

The last equality follows since we obtain the highest value only if $\hat{\mathbf{p}}$ equals the first entry based on the property of strict proper scoring rule.

□

The following claim tells us we can bound the distance between each agent's

best response prediction (the prediction which maximizes his total reward) and his best prediction (the prediction which maximizes his prediciton score) by the distance between his best response prediction and his neighbors' best response predictions.

Claim 133. For any convex function $g(\cdot)$, for any σ_i and $\hat{\sigma}_i$, we have

$$\begin{aligned}&\alpha Pr(\sigma_i)(g(\hat{\mathbf{p}}(i,\sigma_i,\hat{\sigma}_i)) - g(\theta_{-i}\mathbf{q}_{\sigma_i}))\\ \leq&\beta\sum_{j\neq i} Pr(j)\sum_{\sigma_j} Pr(\sigma_j,\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)(g(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_i)) - g(\hat{\mathbf{p}}(i,\sigma_i,\hat{\sigma}_i)))\end{aligned}$$

Proof. Based on Claim 132, we can rewrite (A.5)=(A.8) as a $n\times m\times m$ linear system about

$\{\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k)|k\in[1,n],\sigma_k\in\Sigma,\hat{\sigma}_k\in\Sigma\}$:

$$\hat{\mathbf{p}}(i,\sigma_i,\hat{\sigma}_i) = \arg\max_{\hat{\mathbf{p}}} \mathbb{E}[payment(i,\mathcal{M}+)|\sigma_i] \tag{A.10}$$

$$=\frac{\alpha\theta_{-i}\mathbf{q}_{\sigma_i} + \beta\sum_{j\neq i} Pr(j)\sum_{\sigma_j} Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_i)}{\alpha+\beta\sum_{j\neq i} Pr(j)\sum_{\sigma_j} Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)} \tag{A.11}$$

Fix i, let $\lambda_i = \frac{\alpha}{\alpha+\beta\sum_{j\neq i} Pr(j)\sum_{\sigma_j} Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)}$, $\lambda_{j,\sigma_j} = \frac{\beta\sum_{j\neq i} Pr(j)\sum_{\sigma_j} Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)}{\alpha+\beta\sum_{j\neq i} Pr(j)\sum_{\sigma_j} Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)}$ for $j\neq i$ and $\sigma_j\in\Sigma$, we have $\lambda_i + \sum_{j\neq i,\sigma_j}\lambda_{j,\sigma_j} = 1$

Based on the convexity of $g(\cdot)$, we have

$$\begin{aligned}g(\hat{\mathbf{p}}(i,\sigma_i,\hat{\sigma}_i)) &= g(\lambda_i\theta_{-i}\mathbf{q}_{\sigma_i} + \sum_{j\neq i,\sigma_j}\lambda_{j,\sigma_j}\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_i))\\ &\leq \lambda_i g(\theta_{-i}\mathbf{q}_{\sigma_i}) + \sum_{j\neq i,\sigma_j}\lambda_{j,\sigma_j} g(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_i))\end{aligned}$$

After substitutions, we multiply $\left(\alpha+\beta Pr(j)\sum_{j\neq i}\sum_{\sigma_j} Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)\right) Pr(\sigma_i)$ in both sides. Note that $Pr(\sigma_i)Pr(\sigma_j|\sigma_i) = Pr(\sigma_j,\sigma_i)$, then by manipulation, the claim follows.

□

Claim 133 gives an upper bound to $g(\hat{\mathbf{p}}_k) - g(\theta_{-k}\mathbf{q}_{\sigma_k})$ that is the distance between agent k's best response prediction and his neighbors' best response predictions. Now we continue the proof for our main lemma.

To bound

$$\sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)*$$
$$\int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k})(D^*(\hat{\mathbf{p}}_j,\hat{\mathbf{p}}_k) - D^*(\hat{\mathbf{p}}_j,\theta_{-k}\mathbf{q}_{\sigma_k}) + D^*(\hat{\mathbf{p}}_j,\theta_{-k}\mathbf{q}_{\sigma_k}) - D^*(\theta_{-j}\mathbf{q}_{\sigma_j},\theta_{-k}\mathbf{q}_{\sigma_k})) \tag{A.12}$$

We rewrite $\int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k})$ as $\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k)$ and $\hat{\mathbf{p}}_j$ as $\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j)$, $\hat{\mathbf{p}}_k$ as $\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k)$ which we can do because of Claim 132.

We first give an upper bound to

$$\sum_{j,k}\sum_{\sigma_j,\sigma_k,\hat{\sigma}_j,\hat{\sigma}_k} Pr(j,k)Pr(\sigma_j,\sigma_k)\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k)$$
$$(D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k)) - D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\theta_{-k}\mathbf{q}_{\sigma_k}))$$

The analysis for the second part is similar.

Based on Claim 133, we have

$$\sum_{j,k}\sum_{\sigma_j,\sigma_k,\hat{\sigma}_j,\hat{\sigma}_k} Pr(j,k)Pr(\sigma_j,\sigma_k)\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k) \tag{A.13}$$

$$* \left(D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k)) - D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\theta_{-k}\mathbf{q}_{\sigma_k})\right) \tag{A.14}$$

$$\leq \sum_{j,k}\sum_{\sigma_j,\sigma_k,\hat{\sigma}_j,\hat{\sigma}_k} Pr(j,k)Pr(\sigma_j,\sigma_k)\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k) \tag{A.15}$$

$$* \frac{\beta}{\alpha Pr(\sigma_k)} \sum_{l\neq k}\sum_{\sigma_l} Pr(l)Pr(\sigma_l,\sigma_k)\theta_l(\hat{\sigma}_k,\sigma_l) \tag{A.16}$$

$$* \left(D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(l,\sigma_l,\hat{\sigma}_k)) - D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k))\right) \tag{A.17}$$

Since $\frac{Pr(\sigma_j,\sigma_k)}{Pr(\sigma_k)} \leq 1$, we obtain (A.19) from (A.17).

$$\begin{aligned}
\text{(A.17)} \leq & \sum_{j,k}\sum_{\sigma_j,\sigma_k,\hat{\sigma}_j,\hat{\sigma}_k} Pr(j,k)\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k) \\
& \frac{\beta}{\alpha}\sum_{l\neq k}\sum_{\sigma_l} Pr(l)Pr(\sigma_l,\sigma_k)\theta_l(\hat{\sigma}_k,\sigma_l) && \text{(A.18)} \\
& * \left(D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(l,\sigma_l,\hat{\sigma}_k)) - D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k))\right) && \text{(A.19)} \\
\leq & \sum_{j,k}\sum_{\sigma_j,\sigma_k,\hat{\sigma}_j,\hat{\sigma}_k} Pr(j,k)\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k) \\
& \frac{\beta}{\alpha}\sum_{l\neq k}\sum_{\sigma_l} Pr(l)Pr(\sigma_l,\sigma_k)\theta_l(\hat{\sigma}_k,\sigma_l) && \text{(A.20)} \\
& * \left|\left(D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(l,\sigma_l,\hat{\sigma}_k)) - D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k))\right)\right| && \text{(A.21)}
\end{aligned}$$

Note that (A.21) and (A.19) are identical except for the value sign.

Then we obtain (A.22) from (A.21) since

$$\begin{aligned}
|D^*(x,y) - D^*(x,z)| &\leq (\sqrt{D^*(x,y)} + \sqrt{D^*(x,z)})|\sqrt{D^*(x,y)} - \sqrt{D^*(x,z)})| \\
&\leq 2|\sqrt{D^*(x,y)} - \sqrt{D^*(x,z)})|
\end{aligned}$$

The second inequality follows since $0 \leq D^* \leq 1$

$$\begin{aligned}(\text{A.21}) \leq & 2\sum_{j,k}\sum_{\sigma_j,\sigma_k,\hat{\sigma}_j,\hat{\sigma}_k} Pr(j,k)\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k) \\ & \frac{\beta}{\alpha}\sum_{l\neq k}\sum_{\sigma_l} Pr(l)Pr(\sigma_l,\sigma_k)\theta_l(\hat{\sigma}_k,\sigma_l) \\ & |(\sqrt{D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(l,\sigma_l,\hat{\sigma}_k))} - \sqrt{D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k))})| \quad (\text{A.22})\end{aligned}$$

Once we get (A.22), we can use the fact that $\sqrt{D^*}$ is metric which implies the triangle inequality; (A.23) follows.

$$\begin{aligned}(\text{A.22}) \leq & 2\sum_{j,k}\sum_{\sigma_j,\sigma_k,\hat{\sigma}_j,\hat{\sigma}_k} Pr(j,k)\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k) \\ & \frac{\beta}{\alpha}\sum_{l\neq k}\sum_{\sigma_l} Pr(l)Pr(\sigma_l,\sigma_k)\theta_l(\hat{\sigma}_k,\sigma_l)(\sqrt{D^*(\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k),\hat{\mathbf{p}}(l,\sigma_l,\hat{\sigma}_k))}) \quad (\text{A.23})\end{aligned}$$

Note that $\sum_{\sigma_j}\sum_{\hat{\sigma}_j}\theta_j(\hat{\sigma}_j,\sigma_j) = \sum_{\sigma_j} 1 = m$, also we have $\sum_j Pr(l) = \sum_j Pr(j) = 1$, $Pr(k,l) = Pr(j,k)$ then (A.25) follows.

$$\begin{aligned}(\text{A.23}) = & 2m\frac{\beta}{\alpha}\sum_{l}\sum_{k\neq l}\sum_{\sigma_k,\hat{\sigma}_k} Pr(k,l)\theta_k(\hat{\sigma}_k,\sigma_k)\sum_{\sigma_l} Pr(\sigma_l,\sigma_k)\theta_l(\hat{\sigma}_k,\sigma_l) & (\text{A.24}) \\ & * (\sqrt{D^*(\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k),\hat{\mathbf{p}}(l,\sigma_l,\hat{\sigma}_k))}) & (\text{A.25}) \\ = & 2m\frac{\beta}{\alpha}\sum_{k,l\neq k}\sum_{\sigma_k,\hat{\sigma}_k,\sigma_l} Pr(k,l)\theta_k(\hat{\sigma}_k,\sigma_k)Pr(\sigma_l,\sigma_k)\theta_l(\hat{\sigma}_k,\sigma_l) & (\text{A.26}) \\ & * (\sqrt{D^*(\hat{\mathbf{p}}(k,\sigma_k,\hat{\sigma}_k),\hat{\mathbf{p}}(l,\sigma_l,\hat{\sigma}_k))}) & (\text{A.27}) \\ = & 2m\frac{\beta}{\alpha} \times Inconsistency & (\text{A.28})\end{aligned}$$

The analysis for the second part

$$\sum_{j,k}\sum_{\sigma_j,\sigma_k,\hat{\sigma}_j,\hat{\sigma}_k} Pr(j,k)Pr(\sigma_j,\sigma_k)\theta_j(\hat{\sigma}_j,\sigma_j)\theta_k(\hat{\sigma}_k,\sigma_k)(D^*(\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_j),\theta_{-k}\mathbf{q}_{\sigma_k})-D^*(\theta_{-j}\mathbf{q}_{\sigma_j},\theta_{-k}\mathbf{q}_{\sigma_k}))$$

is similar, note that j and k are symmetric and $D^*(\cdot,\theta_{-k}\mathbf{q}_{\sigma_k})$ is a convex function. We can use Claim 133 and triangle inequality to bound the second part by $2m\frac{\beta}{\alpha}\times$ *Inconsistency*.

So if we set $2m\frac{\beta}{\alpha}<\frac{1}{2}$, then *TotalDivergence*$(s)-$*TotalDivergence*$(s_{BP})<$ *Inconsistency*, proving the inequality in our main lemma.

To prove that if the equality in our main lemma holds then $s=s_{BP}$, we first show that

Claim 134. The equality in *ClassificationScore*$(s)\leq$ *TotalDivergence*(s_{BP}) holds iff *Inconsistency*$(s)=0$.

Proof. Note that (A.1) tells us when *ClassificationScore*$(s)=$ *TotalDivergence*(s_{BP}), we have *Diversity*$(s)=$ *TotalDivergence*(s) which implies *Inconsistency*$(s)=0$ based on Claim 129. □

Then we will prove

Claim 135. If *Inconsistency*$(s)=0$ then $s=s_{BP}$

Proof. Recall in (A.10), we have for any i,

$$\begin{aligned}\hat{\mathbf{p}}(i,\sigma_i,\hat{\sigma}_i)&=\arg\max_{\hat{\mathbf{p}}}\mathbb{E}[payment(i,\mathcal{M}+)|\sigma_i]\\&=\frac{\alpha\theta_{-i}\mathbf{q}_{\sigma_i}+\beta Pr(j)\sum_{j\neq i}\sum_{\sigma_j}Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_i)}{\alpha+\beta Pr(j)\sum_{j\neq i}\sum_{\sigma_j}Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)}\end{aligned}\tag{A.29}$$

If *Inconsistency*$(s)=0$, we can see if $\theta_j(\hat{\sigma}_i,\sigma_j)>0$ we must have $\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_i)=\hat{\mathbf{p}}_i(i,\sigma_i,\hat{\sigma}_i)$. So we have $\theta_{-i}\mathbf{q}_{\sigma_i}=\hat{\mathbf{p}}_i(i,\sigma_i,\hat{\sigma}_i)$ for any i since we have $\alpha\theta_{-i}\mathbf{q}_{\sigma_i}=\alpha\hat{\mathbf{p}}_i(i,\sigma_i,\hat{\sigma}_i)$ if we multiply

$\alpha+\beta Pr(j)\sum_{j\neq i}\sum_{\sigma_j} Pr(\sigma_j|\sigma_i)\theta_j(\hat{\sigma}_i,\sigma_j)$ in both sides of equation (A.29) and combine the fact that $\hat{\mathbf{p}}(j,\sigma_j,\hat{\sigma}_i)=\hat{\mathbf{p}}_i(i,\sigma_i,\hat{\sigma}_i)$. Thus we have $s=s_{BP}$. □

A.1.1.3 Proof for main theorem part 2 and 3

Theorem 82 Part 2: $\mathcal{M}+(\alpha,\beta,PS(\cdot,\cdot))$ **has truth-telling as a focal equilibrium.** We use our main lemma *ClassificationScore*(s) < *TotalDivergence*(s_{BP}) directly to prove: **any *symmetric* non-permutation equilibrium's agent welfare (*ClassificationScore*) must be strictly less than truth-telling**

Notice that if all agents play a symmetric signal strategy θ, then for any j,k, $\theta_{-j}=\theta_{-k}=\theta$. For any symmetric non-permutation equilibrium s, it is possible that the signal strategy of s is not a permutation or it is a permutation θ_π but agents do not report $\pi\mathbf{q}_\sigma$ given σ is their private signal. So we consider two cases:

(a) We first consider the case that the signal strategy θ of s is a permutation matrix θ_π, but agents do not report $\pi\mathbf{q}_\sigma$.

$$
\begin{aligned}
\textit{ClassificationScore}(s) <& \textit{TotalDivergence}(s_{BP})\\
=& \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta_{-j}\mathbf{q}_{\sigma_j},\theta_{-k}\mathbf{q}_{\sigma_k})\\
=& \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta_\pi\mathbf{q}_{\sigma_j},\theta_\pi\mathbf{q}_{\sigma_k})\\
=& \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\mathbf{q}_{\sigma_j},\mathbf{q}_{\sigma_k})\\
=& \textit{TotalDivergence}(truthtelling)\\
=& \textit{ClassificationScore}(truthtelling)
\end{aligned}
$$

The first inequality follows from our main lemma. The inequality is strict for the following reason: when the signal strategy θ of s is a permutation matrix, s_{BP} is a

permutation strategy profile since for any i, agent i's best prediction is $\theta_{-i}\mathbf{q}_\sigma = \theta\mathbf{q}_\sigma$. Based on our main lemma if $\mathit{ClassificationScore}(s) = \mathit{TotalDivergence}(s_{BP})$, we have $s = s_{BP}$ which implies that s is a permutation strategy profile which is a contradiction to the fact s is a non-permutation strategy profile.

The second line follows since at s_{BP}, each agent's reported prediction only depends on his private signal.

The last equality follows from Corollary 130.

(b) We consider the case that the signal strategy θ of s is not a permutation matrix. The above proof still holds except in two places: one is the inequality in the first line may not be strict; another is the equality in the fourth line should be a strict inequality:

$$\begin{aligned} & \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta_\pi\mathbf{q}_{\sigma_j},\theta_\pi\mathbf{q}_{\sigma_k}) \\ < & \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\mathbf{q}_{\sigma_j},\mathbf{q}_{\sigma_k}) \end{aligned}$$

The inequality must be strict since based on Corollary 80, we know that if θ is not a permutation, and Q is fine-grained, then there exists $\sigma_1 \neq \sigma_2$ such that $D^*(\theta\mathbf{q}_{\sigma_1},\theta\mathbf{q}_{\sigma_2}) < D^*(\mathbf{q}_{\sigma_1},\mathbf{q}_{\sigma_2})$. Also based on non-zero assumption of Q, we have $Pr(\sigma_j = \sigma_1, \sigma_k = \sigma_2) > 0$.

So in both of the above two cases, we have

$$\mathit{ClassificationScore}(s) < \mathit{ClassificationScore}(\mathit{truthtelling})$$

if s is not a permutation equilibrium

Theorem 82 Part 3: $\mathcal{M}+(\alpha, \beta, PS(\cdot,\cdot))$ has truth-telling as a robust focal equilibrium:

TotalDivergence$(truthtelling) \approx$ ***TotalDivergence***$(s_{BP}) \Rightarrow \theta \approx \pi$ Now we start to prove that when a symmetric equilibrium has classification score that is close to that of truth-telling, its signal strategy θ is close to a permutation. We prove it by contradiction. We will assume θ is far from a permutation equilibrium, that is, recalling the definition of τ-close, we assume there exists a row of θ that has at least two large numbers. Under this assumption, we will show the total divergence of s_{BP} is far from classification score of truth-telling when n is sufficiently large. Formally, we assume that given any τ, there exists $u', v', w' \in \Sigma$ such that $\theta(u', v') > \tau, \theta(u', w') > \tau$. Under this assumption, we will prove that, when $n > N(\tau, Q)$, the total divergence of s_{BP} is $O(\tau^3)$ far from classification score of truth-telling.

Lemma 136. *Given any fixed τ, for any symmetric equilibrium s with signal strategy θ, if there exists $u', v', w' \in \Sigma$ such that $\theta(u', v') > \tau, \theta(u', w') > \tau$, then*

$$\textit{TotalDivergence}(truthtelling) - \textit{TotalDivergence}(s_{BP}) \geq c_2(\tau c_1)^3 c_4 c_3$$

Proof of Lemma 136. We first write *TotalDivergence* in an explicit form:

$$\sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\mathbf{q}_{\sigma_j}, \mathbf{q}_{\sigma_k}) - \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta\mathbf{q}_{\sigma_j}, \theta\mathbf{q}_{\sigma_k}) \tag{A.30}$$

Actually, We will show for any j, k,

$$\sum_{\sigma_j,\sigma_k} Pr(\sigma_j,\sigma_k)D^*(\mathbf{q}_{\sigma_j}, \mathbf{q}_{\sigma_k}) - \sum_{\sigma_j,\sigma_k} Pr(\sigma_j,\sigma_k)D^*(\theta\mathbf{q}_{\sigma_j}, \theta\mathbf{q}_{\sigma_k}) \tag{A.31}$$

is greater than $c_2(\tau c_1)^3 c_4 c_3$, which implies the result.

We want give a lower bound for (A.31). In order to obtain this lower bound, we are going to transform this value to $\sum_u \lambda_u g(x_u) - g(\sum_u \lambda_u x_u)$ where $g(\cdot)$ is a convex function. To obtain a lower bound of $\sum_u \lambda_u g(x_u) - g(\sum_u \lambda_u x_u)$, we have an observation:

For any convex function $g(\cdot)$, $g(\sum_u \lambda_u x_u)$ and $\sum_u \lambda_u g(x_u)$ are "very different" if there are two large coefficients λ_1 and λ_2 with the corresponding x_1 and x_2 that are "very different". Now we introduce a claim to show this observation.

Claim 137.

$$\sum_u \lambda_u g(x_u) - g(\sum_u \lambda_u x_u) \geq \frac{d_2(g)}{2} \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} ||x_1 - x_2||^2$$

where $d_2(g)$ is a lower bound of $g''(\cdot)$

Proof.

$$g\left(\sum_u \lambda_u x_u\right) \leq (\lambda_1 + \lambda_2) g\left(\frac{\lambda_1 x_1 + \lambda_2 x_2}{\lambda_1 + \lambda_2}\right) + \sum_{u>2} \lambda_u g(x_u) \leq \sum_u \lambda_u g(x_u)$$

So

$$
\begin{aligned}
&\sum_u \lambda_u g(x_u) - g(\sum_u \lambda_u x_u) \\
&\geq \sum_u \lambda_u g(x_u) - (\lambda_1+\lambda_2) g\left(\frac{\lambda_1 x_1 + \lambda_2 x_2}{\lambda_1+\lambda_2}\right) - \sum_{u>2} \lambda_u g(x_u) \\
&= \lambda_1 g(x_1) + \lambda_2 g(x_2) - (\lambda_1+\lambda_2) g(\frac{\lambda_1 x_1 + \lambda_2 x_2}{\lambda_1+\lambda_2}) \\
&= (\lambda_1+\lambda_2)(\frac{\lambda_1 g(x_1) + \lambda_2 g(x_2)}{\lambda_1+\lambda_2} - g(\frac{\lambda_1 x_1 + \lambda_2 x_2}{\lambda_1+\lambda_2})) \\
&\geq (\lambda_1+\lambda_2)\frac{d_2(g)}{2}\frac{\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^2}||x_1-x_2||^2 \\
&= \frac{d_2(g)}{2}\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}||x_1-x_2||^2
\end{aligned}
$$

where $d_2(g)$ is the lower bound of $g''(\cdot)$

The first inequality follows if we rewrite $\sum_u \lambda_u x_u$ as $(\lambda_1+\lambda_2)\frac{\lambda_1 x_1+\lambda_2 x_2}{\lambda_1+\lambda_2} + \sum_{u>2} \lambda_u x_u$ and apply convexity.

Then we do several manipulations including taking $\lambda_1+\lambda_2$ outside. For continuous convex function $g(\cdot)$, we have $tg(x) + (1-t)g(y) - g(tx+(1-t)y) \geq \frac{1}{2} d_2(g) t(1-t)||x-y||^2$ according to [47], then we replace t by $\frac{\lambda_1}{\lambda_1+\lambda_2}$ and set $x = x_1, y = x_2$ and obtain the final result.

□

We can think of $\theta(u', v')$ and $\theta(u', w')$ as the two large coefficients (actually they are part of the coefficients). Then we need to find two "very different" entries that corresponding to those large coefficients. We pick two specific signals $s', t' \in \Sigma$ such that $\mathbf{q}_{s'}$ and $\mathbf{q}_{t'}$ are "very different" in position v' and w'. The reason we do this is that when we compute $\theta\mathbf{q}$, $\theta(u', v')$ and $\theta(u', w')$ are the two large entries which

correspond to the positions v' and w' in $\mathbf{q}$. Formally, we pick $s', t' \in \Sigma$ such that

$$\left\| \frac{q(v'|s')}{q(v'|t')} - \frac{q(w'|s')}{q(w'|t')} \right\| = \max_{s,t} \left\| \frac{q(v'|s)}{q(v'|t)} - \frac{q(w'|s)}{q(w'|t)} \right\|$$

Once we have chosen the two specific signals, since $Pr(s', t')(D^*(\mathbf{q}_{s'}, \mathbf{q}_{t'}) - D^*(\theta\mathbf{q}_{s'}, \theta\mathbf{q}_{t'}))$ is less than (A.31) based on the fact $D^*(\mathbf{q}_s, \mathbf{q}_t) - D^*(\theta\mathbf{q}_s, \theta\mathbf{q}_t) \geq 0$ for $s, t \neq s', t'$, we will give a lower bound of $Pr(s', t')(D^*(\mathbf{q}_{s'}, \mathbf{q}_{t'}) - D^*(\theta\mathbf{q}_{s'}, \theta\mathbf{q}_{t'}))$ which is also a lower bound of (A.31).

Let $f(x) = (\sqrt{x} - 1)^2$. For convenience, we will write the dot product of two vectors $\sum_v a(v)b(v)$ as $a(\cdot)b(\cdot)$. Now we give a explicit form of D^*:

$$Pr(s', t')(D^*(\mathbf{q}_{s'}, \mathbf{q}_{t'}) - D^*(\theta\mathbf{q}_{s'}, \theta\mathbf{q}_{t'})) \tag{A.32}$$

$$= Pr(s', t') \left(\sum_v q(v|s') f\left(\frac{q(v|t')}{q(v|s')} \right) - \sum_u \theta(u, \cdot) q(\cdot|s') f\left(\frac{1}{\theta(u, \cdot) q(\cdot|s')} \theta(u, \cdot) q(\cdot|t') \right) \right) \tag{A.33}$$

We take $\sum_u \theta(u, \cdot) q(\cdot|s')$ out and note that $\sum_u \theta(u, v) = 1$,
so $\sum_u \theta(u, \cdot) q(\cdot|s') \frac{1}{\theta(u, \cdot) q(\cdot|s')} \theta(u, v) = 1$, then we obtain (A.35) from (A.33).

$$\begin{aligned} \text{(A.33)} = & Pr(s', t') \sum_u \theta(u, \cdot) q(\cdot|s') * \\ & \left(\frac{1}{\theta(u, \cdot) q(\cdot|s')} \sum_v \theta(u, v) q(v|s') f\left(\frac{q(v|t')}{q(v|s')} \right) \right. && \text{(A.34)} \\ & \left. - f\left(\frac{1}{\theta(u, \cdot) q(\cdot|s')} \sum_v \theta(u, v) q(v|s') \frac{q(v|t')}{q(v|s')} \right) \right) && \text{(A.35)} \end{aligned}$$

Then we pick the special u' to obtain (A.37). For the part $\sum_{u \neq u'}$, since $f(\cdot)$ is a

convex function, we have

$$\frac{1}{\theta(u,\cdot)q(\cdot|s')}\sum_v \theta(u,v)q(v|s')f(\frac{q(v|t')}{q(v|s')}) \geq f\left(\frac{1}{\theta(u,\cdot)q(\cdot|s')}\sum_v \theta(u,v)q(v|s')\frac{q(v|t')}{q(v|s')}\right)$$

so (A.35) is greater than (A.37).

$$\begin{aligned}
\text{(A.35)} \geq & Pr(s',t')\theta(u',\cdot)q(\cdot|s')* \\
& \left(\frac{1}{\theta(u',\cdot)q(\cdot|s')}\sum_v \theta(u',v)q(v|s')f\left(\frac{q(v|t')}{q(v|s')}\right)\right. && \text{(A.36)} \\
& \left. - f\left(\frac{1}{\theta(u',\cdot)q(\cdot|s')}\sum_v \theta(u',v)q(v|s')\frac{q(v|t')}{q(v|s')}\right)\right) && \text{(A.37)}
\end{aligned}$$

Note that $\theta(u',v')$ and $\theta(u',w')$ are large, so in the convex function $f(\cdot)$, there are two large coefficients $\frac{1}{\theta(u',\cdot)q(\cdot|s')}\theta(u',v')q(v'|s')$ and $\frac{1}{\theta(u',\cdot)q(\cdot|s')}\theta(u',w')q(w'|s')$ which correspond to $\frac{q(v'|t')}{q(v'|s')}$ and $\frac{q(w'|t')}{q(w'|s')}$. Then based on our choice for s',t' and Claim 137, we have

$$\text{(A.37)} \geq Pr(s',t')\theta(u',\cdot)q(\cdot|s')\frac{c_4}{2}\left(\frac{(\theta(u',v')q(v'|s')) * (\theta(v',w')q(w'|s'))}{\theta(u',v')q(v'|s') + \theta(v',w')q(w'|s')}\left\|\frac{q(v'|t')}{q(v'|s')} - \frac{q(w'|t')}{q(w'|s')}\right\|^2\right) \quad \text{(A.38)}$$

$$\geq c_2(\tau c_1)^3 c_4 c_3 \quad \text{(A.39)}$$

The last inequality follows since $Pr(s',t') \geq c_2$, both $\theta(u',v')q(v'|s')$ and $\theta(v',w')q(w'|s')$ are greater than τc_1. Also note that:

$$\theta(u',\cdot)q(\cdot|s') \geq \theta(u',v')q(v'|s') + \theta(v',w')q(w'|s') \geq 2\tau c_1$$

and

$$\theta(u', v')q(v'|s') + \theta(v', w')q(w'|s') \leq 1.$$

□

Any *symmetric* equilibrium that has agent-welfare close to truth-telling must be close to a permutation equilibrium: We have already proved that no symmetric equilibrium pays more than truth-telling. For the symmetric equilibrium s^* such that

$$\mathit{ClassificationScore}(s^*) > \mathit{ClassificationScore}(\mathit{truthtelling}) - \gamma_1$$

we have

$$\begin{aligned}
&\mathit{TotalDivergence}(\mathit{truthtelling}) = \mathit{ClassificationScore}(\mathit{truthtelling}) \\
\leq &\mathit{ClassificationScore}(s^*) + \gamma_1 \\
\leq &\mathit{TotalDivergence}(s^*_{BP}) + \gamma_1 \leq \mathit{ClassificationScore}(\mathit{truthtelling}) + \gamma_1
\end{aligned}$$

Let $\gamma_1 = (\tau_1 c_1)^3 c_2 c_3 c_4$, then s^* is τ_1 close to a permutation equilibrium or there will be a contradiction based on Lemma 136. By manipulations, we will obtain our result.

A.1.2 Asymmetric equilibria

Theorem 138. *For any number of signals m, given any SNIFE prior, in $\mathcal{M}+(\alpha, \beta, PS(\cdot, \cdot))$ with $\frac{\alpha}{\beta} < \frac{1}{4m}$,*

1. *no equilibrium has agent welfare greater than $\gamma_2(n)$ more than that of truth-telling where n is the number of agents; and*

2. *any profile that pays within $\gamma_2(n)$ of truth-telling must be $\tau_2(n)$ close to a permutation strategy profile.*

where $\gamma_2(n) = O(\frac{m}{\sqrt{n}})$ and $\tau_2(n) = O(\sqrt[6]{\frac{m^2}{n}})$ (the constants we omit only depend on the first two moments of prior Q)[1].

Proof Outline for Theorem 138 We want to show that if the number of agents is sufficiently large, then no equilibrium can have a *ClassificationScore* that is much greater than truth-telling; any equilibrium that has *ClassificationScore* close to truth-telling must be close to a permutation equilibrium.

The proof is similar with the proof of our main theorem. At a high level, we will show when the number of agents is sufficiently large, any asymmetric equilibrium can be **symmetrized**. Then we follow the proof of our main theorem by using the **symmetrized** version of the asymmetric equilibrium.

We define *symmetrized* s_{BP} as a strategy where each agent plays $\bar{\theta}_n \mathbf{q}_\sigma$ given σ is his private signal where $\bar{\theta}_n$ is the average signal strategy of s_{BP} (also of s). We show the report profiles of *symmetrized* s_{BP} in the third picture of Figure A.1

Recall that in the proof of our main theorem, we show that

(1) $\boldsymbol{ClassificationScore}(s) \leq \boldsymbol{TotalDivergence}(s_{BP})$.[Lemma 131].

(2) $\boldsymbol{TotalDivergence}(truthtelling) \approx \boldsymbol{TotalDivergence}(s_{BP}) \Rightarrow \theta \approx \pi$. [Lemma 136]

Note that in part (2) is valid only if s is symmetric. However, if we replace θ by $\bar{\theta}_n$ where $\bar{\theta}_{\mathbf{n}}$ is the average signal strategy of s^*. We can rewrite part (2) as

(2) $\boldsymbol{TotalDivergence}(truthtelling) \approx \boldsymbol{TotalDivergence}(symmetrized\ s^*_{BP}) \Rightarrow \bar{\theta}_{\mathbf{n}} \approx \pi$ where $\bar{\theta}_{\mathbf{n}}$ is the average signal strategy of s^*. [Lemma 136]

[1]Actually $\gamma_2(n) = \frac{4\sqrt{2}m}{\sqrt{n}}$ and $\tau_2^6(n) = \frac{128*m^2}{nc_1^6(c_2c_3c_4)^2}$, $c_1 = \min_{s,t\in\Sigma} q(s|t)$, $c_2 = \min_{s,t\in\Sigma} Pr(s,t)$, $c_3 = \min_{u,v} \max_{s,t} ||\frac{q(u|s)}{q(u|t)} - \frac{q(v|s)}{q(v|t)}||^2$, $c_4 = \min_{s,t,u} f''(\frac{q(u|s)}{q(u|t)})$ where $f(x) = (\sqrt{x}-1)^2$.

It is valid for any equilibrium s.

In addition to the two key parts proved in the proof of our main theorem, we have to prove two more parts to prove the asymmetric case. The whole proof of Theorem 138 is illustrated in Figure A.1:

At a high level, the following two parts show that when the number of agents is sufficiently large, we can replace any asymmetric equilibrium by its **symmetrized** version since they are "close" to each other.

(3) ***TotalDivergence***$(s_{BP}) \approx$ ***TotalDivergence***$(symmetrized\ s_{BP})$ **when the number of agents is sufficiently large.** [Lemma 139] Intuitively, when n is large enough, θ_{-j} will be close to $\bar{\theta}_n$. We will use this observation to prove this part.

(4) ***ClassificationScore***$(s^*) \geq$ ***ClassificationScore***$(truthtelling)$
$\Rightarrow$ ***TotalDivergence***$(truthtelling) \approx$ ***TotalDivergence***$(symmetrized\ s^*_{BP})$ **when the number of agents is sufficiently large.**[Corollary 140] Here s^*_{BP} is the best prediction strategy of s^*. This part will also imply no equilibrium can have *ClassificationScore* that is much greater than truth-telling.

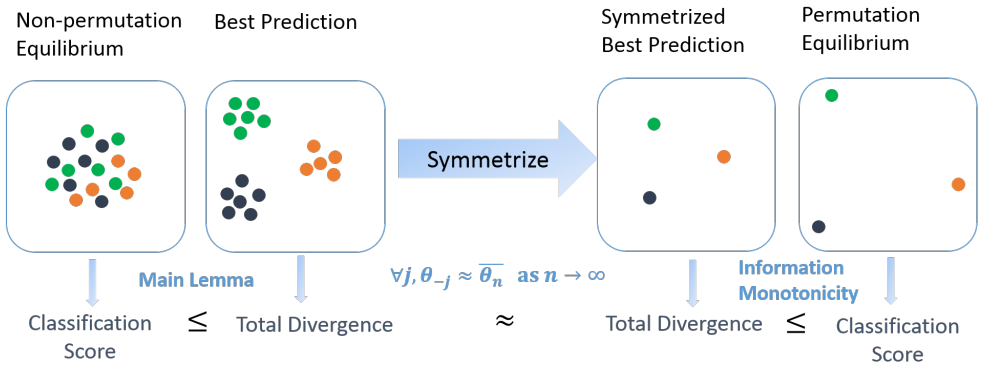

Figure A.1: Proof Outline for Theorem 138

A.1.3 Proof for Theorem 138

(3) *TotalDivergence*(s_{BP}) $\approx$ *TotalDivergence*($symmetrized\ s_{BP}$) when the number of agents is sufficiently large. We symmetrize s_{BP} which means we let each agent i report $\bar{\theta}_n \mathbf{q}_{\sigma_i}$ given σ_i is agent i's private signal and $\bar{\theta}_n$ is the average signal strategy of s_{BP} and show that the total divergence will not change much. Intuitively, this is because θ_{-i} are similar among agents when there are many agents. l

Lemma 139. *Given any SNIFE prior Q, for any $\epsilon > 0$, there exists $N_\epsilon = \frac{32*m^2}{\epsilon^2}$ such that if $n > N_\epsilon$, for any strategy $(\theta_1, \theta_2, ..., \theta_n)$, any two agents j, k,*

$$|D^*(\theta_{-j}\mathbf{q}_{\sigma_j}, \theta_{-k}\mathbf{q}_{\sigma_k}) - D^*(\bar{\theta}_n\mathbf{q}_{\sigma_j}, \bar{\theta}_n\mathbf{q}_{\sigma_k})| < \epsilon$$

Proof of Lemma 139. For convenience, let $s = \sigma_j, t = \sigma_k$

$$|D^*(\theta_{-j}\mathbf{q}_s, \theta_{-k}\mathbf{q}_t) - D^*(\bar{\theta}_n\mathbf{q}_s, \bar{\theta}_n\mathbf{q}_t)| \tag{A.40}$$

$$=|\sum_u \left(\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s} - \sqrt{\theta_{-k}(u,\cdot)\mathbf{q}_t}\right)^2 - \left(\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s} - \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_t}\right)^2| \tag{A.41}$$

$$=|\sum_u \left(\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s} - \sqrt{\theta_{-k}(u,\cdot)\mathbf{q}_t} - \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s} + \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_t}\right) *$$
$$\left(\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s} - \sqrt{\theta_{-k}(u,\cdot)\mathbf{q}_t} + \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s} - \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_t}\right)| \tag{A.42}$$

$$\leq 2 * m * \max_u \left(|\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s} - \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}| + |\sqrt{\theta_{-k}(u,\cdot)\mathbf{q}_t} - \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_t}|\right) \tag{A.43}$$

$$\leq 4 * m * \max_{u,s,j} |\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s} - \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}| \tag{A.44}$$

The first equality follows from the definition of Helinger-divergence.

The second equality is just formula for the difference of square.

To arrive at (A.43), $\sum_u |\left(\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s} - \sqrt{\theta_{-k}(u,\cdot)\mathbf{q}_t} + \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s} - \sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_t}\right)| \leq \sum_u 2 = 2m$ where the inequality follows from the fact $0 < D^* < 1$.

The last equality follows since both $|\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}-\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}|$ and $|\sqrt{\theta_{-k}(u,\cdot)\mathbf{q}_t}-\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_t}|$ are less than $\max_{u,s,j}|\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}-\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}|$.

Now we consider two cases for any u, s, j:

(1) $|\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}-\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}| \leq \frac{\epsilon}{4*m}$: It is clear the result in this Lemma follows.

(2) $|\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}-\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}| > \frac{\epsilon}{4*m}$: Notice that $(n-1)\theta_{-j} = n\bar{\theta}_n - \theta_j$, then we can see

$$\theta_{-j} = \bar{\theta}_n + \frac{1}{n}(\theta_{-j} - \theta_j)$$

$$4*m*|\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}-\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}| \tag{A.45}$$

$$= 4*m*|\frac{\theta_{-j}(u,\cdot)\mathbf{q}_s - \bar{\theta}_n(u,\cdot)\mathbf{q}_s}{\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}+\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}}| \tag{A.46}$$

$$= 4*m*\frac{\frac{1}{n}|(\theta_{-j}(u,\cdot)-\theta_j(u,\cdot))\mathbf{q}_s|}{\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}+\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}} \tag{A.47}$$

$$< 4*m*2*\frac{4*m}{\epsilon}\frac{1}{n} < \epsilon \tag{A.48}$$

when $n > N_\epsilon = \frac{32*m^2}{\epsilon^2}$

The first equality follows from the formula of the difference of squares.

The second equality follows from $\theta_{-j} = \bar{\theta}_n + \frac{1}{n}(\theta_{-j} - \theta_j)$.

If $|\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}-\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}| > \frac{\epsilon}{4*m}$, we have $|\sqrt{\theta_{-j}(u,\cdot)\mathbf{q}_s}+\sqrt{\bar{\theta}_n(u,\cdot)\mathbf{q}_s}| > \frac{\epsilon}{4*m}$ as well, the third line follows.

□

(4) *ClassificationScore*$(s^*) \geq$ *ClassificationScore*$(truthtelling)$

$\Rightarrow$ *TotalDivergence*$(truthtelling) \approx$ *TotalDivergence*$(symmetrized\ s^*_{BP})$ when the number of agents is sufficiently large. The following corollary is derived from Lemma 139. It will imply not only

TotalDivergence$(truthtelling) \approx$ *TotalDivergence*$(symmetrized\ s^*_{BP})$ but also any

equilibrium cannot have agent-welfare (*ClassificationScore*) that is much greater than truth-telling when the number of agents is sufficiently large.

Corollary 140. *Given any SNIFE prior Q, for any $\epsilon > 0$, if $n > N_\epsilon = \frac{128*m^2}{\epsilon^2}$, for any equilibrium s^* that has greater ClassificationScore than the truth-telling ClassificationScore minus $\epsilon/2$:*

$$\begin{aligned}
&\mathit{Classification}(\mathit{truthtelling}) \\
&< \mathit{Classification}(s^*) + \frac{\epsilon}{2} \\
&< \mathit{TotalDivergence}(\mathit{symmetrized}\ s^*_{BP}) + \epsilon \\
&\leq \mathit{Classification}(\mathit{truthtelling}) + \epsilon
\end{aligned}$$

Proof for Corollary 140.

$$\begin{aligned}
\mathit{TotalDivergence}(\mathit{truthtelling}) &= \mathit{ClassificationScore}(\mathit{truthtelling}) \\
&\leq \mathit{ClassificationScore}(s^*) + \frac{\epsilon}{2} \\
&\leq \mathit{TotalDivergence}(s^*_{BP}) + \frac{\epsilon}{2} \\
&< \mathit{TotalDivergence}(\mathit{symmetrized}\ s^*_{BP}) + \epsilon \\
&\leq \mathit{ClassificationScore}(\mathit{truthtelling}) + \epsilon
\end{aligned}$$

The first equality follows from Corollary 130.

The second inequality follows from the condition.

The third inequality follows from the main lemma.

The fourth inequality follows from Lemma 139.

The last inequality follows from information monotonicity since

$$\begin{aligned}
&\textit{ClassificationScore}(\textit{truthtelling}) - \textit{TotalDivergence}(\textit{symmetrized } s^*_{BP}) \\
&= \sum_{\substack{j \\ k \neq j}} \sum_{\sigma_j, \sigma_k} Pr(j,k) Pr(\sigma_j, \sigma_k) D^*(\mathbf{q}_{\sigma_j}, \mathbf{q}_{\sigma_k}) - \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k) Pr(\sigma_j, \sigma_k) D^*(\bar{\theta}_n \mathbf{q}_{\sigma_j}, \bar{\theta}_n \mathbf{q}_{\sigma_k}) \\
&= \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k) Pr(\sigma_j, \sigma_k) (D^*(\mathbf{q}_{\sigma_j}, \mathbf{q}_{\sigma_k}) - D^*(\bar{\theta}_n \mathbf{q}_{\sigma_j}, \bar{\theta}_n \mathbf{q}_{\sigma_k})) \geq 0
\end{aligned}$$

The second equality follows since if $j = k$, $D^*(\mathbf{q}_{\sigma_j}, \mathbf{q}_{\sigma_k}) = 0$

□

This corollary induces the following result:

No equilibrium can have agent-welfare that is much greater than truth-telling Let $\frac{\epsilon}{2} = \gamma_2$, we need $n \geq \frac{128 * m^2}{\epsilon^2}$ to obtain γ_2 tolerance based on Corollary 140. By manipulations, we obtain our result.

If the number of agents is sufficiently large, any equilibrium that has agent-welfare close to truth-telling must be close to permutation equilibrium: Let $\epsilon = (\tau_2 c_1)^3 c_2 c_3 c_4$, if $n > \frac{32 * m^2}{(\epsilon/2)^2}$, we have already proved that

$$\textit{TotalDivergence}(\textit{truthtelling}) - \textit{TotalDivergence}(\textit{symmetrized } s^*_{BP}) < \epsilon$$

based on Corollary 140. If s^* is not τ_2 close to a permutation equilibrium, we will have

$$\textit{TotalDivergence}(\textit{truthtelling}) - \textit{TotalDivergence}(\textit{symmetrized } s^*_{BP}) > (\tau_2 c_1)^3 c_2 c_3 c_4 = \epsilon$$

which is a contradiction based on Lemma 136. By manipulations, we obtain our result.

A.1.4 Proof for claims

Claim 72. Assume that the distribution over all agents' private signals is $\omega \in \Delta_\Sigma$, the distribution over all agents' reported signals will be $\bar{\theta}_n\omega$.

Proof for Claim 72. The probability of signal σ will be

$$\sum_i Pr(i) \sum_{\sigma'} \theta_i(\sigma, \sigma')\omega(\sigma') = \frac{1}{n} \sum_i \sum_{\sigma'} \theta_i(\sigma, \sigma')\omega(\sigma') = \sum_{\sigma'} \bar{\theta}_n(\sigma, \sigma')\omega(\sigma')$$

where $Pr(i)$ is the probability agent i is picked. For each agent i, we sum the probability agent i receives private signal σ' which is $\omega(\sigma')$ times the probability that he reports σ given he receives σ' which is $\theta_i(\sigma, \sigma')$ over all possible private signal σ'.

So the distribution of reported signals is $\bar{\theta}_n\omega$. □

Claim 74. For each agent i, if he receives private signal σ_i, agent i will believe that the expected likelihood of other agents' reported signals is $\theta_{-i}\mathbf{q}_{\sigma_i}$ where $\theta_{-i} = \frac{\sum_{j\neq i}\theta_j}{n-1}$.

Proof for Claim 74. For each agent i, given he receives private signal σ_i, he will believe the expected likelihood for other agents' private signals is $\mathbf{q}_{\sigma_i}$. Based on Claim 72, he will believe the expected likelihood for other agents' reported signals is the average signal strategy of other agents' signal strategies times $\mathbf{q}_{\sigma_i}$ which is $\theta_{-i}\mathbf{q}_{\sigma_i}$ where $\theta_{-i} = \frac{\sum_{j\neq i}\theta_j}{n-1}$. □

Claim 78. For any transition matrix $\theta_{m\times m}$ where the sum of every column is 1, θ is a permutation matrix iff for any row of θ, there at most one non-zero entry.

Proof for Claim 78. It is clear that any permutation matrix has exactly one non-zero entry, which is 1, in each row and each column. Thus we only need to prove the direction that if for any row of θ, there is at most one non-zero entry, θ must be a permutation matrix.

We first prove that there are exactly m non-zero entries in θ: if for any row of θ, there is at most one non-zero entry, we can see θ has at most m non-zero entries. θ

is a transition matrix where the sum of every column is 1,which implies that θ has at least m non-zero entries. Thus we proved there are exactly m non-zero entries in θ.

We have just shown that θ has exactly m non-zero entries. Since θ has at most one non-zero entry in each row, θ must have exactly one non-zero entry in each row. θ also has at least one non-zero entry in each column since it is a transition matrix, so θ must have exactly one non-zero entry 1 in each column. Thus θ has exactly one non-zero entry 1 in each row and each column which implies that θ is a permutation matrix. □

Claim 125. The *Disagreement Mechanism* has the same equilibria as the Divergence-based BTS.

Proof for Claim 125. The value of $score_C(r_j, r_k)$ does not depend on agent i's strategy. The term related to agent i's strategy contained in $score_{\mathcal{M}}$ is $payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i, \mathbf{r})$. This implies that agent i's marginal benefit from deviation in $\mathcal{M}+(\alpha, \beta, PS(\cdot,\cdot))$ is the same with its marginal benefit from the same deviation in $\mathcal{M}(\alpha, \beta, PS(\cdot,\cdot))$. □

Claim 126.

$$ClassificationScore = Diversity - Inconsistency$$

Proof for Claim 126. Based on the definition of *ClassificationScore*, we have

$$\sum_{\substack{i \\ j \neq i}} \sum_{\substack{k \neq i,j \\ \sigma_i, \sigma_j, \sigma_k}} Pr(i) Pr(\sigma_i) Pr(j,k) Pr(\sigma_j, \sigma_k | \sigma_i) * \tag{A.49}$$

$$\int_{\hat{\sigma}_j, \hat{\mathbf{p}}_j, \hat{\sigma}_k, \hat{\mathbf{p}}_k} Pr_{(\hat{\sigma}_j, \hat{\mathbf{p}}_j) \leftarrow s_j(\sigma_j)}(\hat{\sigma}_j, \hat{\mathbf{p}}_j) Pr_{(\hat{\sigma}_k, \hat{\mathbf{p}}_k) \leftarrow s_k(\sigma_k)}(\hat{\sigma}_k, \hat{\mathbf{p}}_k) score_C(r_j, r_k) \tag{A.50}$$

Now we begin our proof:

$$\sum_{\substack{i \\ j \neq i}} \sum_{\substack{k \neq i,j \\ \sigma_i, \sigma_j, \sigma_k}} Pr(i) Pr(\sigma_i) Pr(j,k) Pr(\sigma_j, \sigma_k | \sigma_i) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) score_C(r_j, r_k)$$

$$= \sum_{\substack{i \\ j \neq i}} \sum_{\substack{k \neq i,j \\ \sigma_j, \sigma_k}} \frac{1}{n(n-1)(n-2)} Pr(\sigma_j, \sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) score_C(r_j, r_k)$$

$$= \frac{1}{n(n-1)} \sum_{\substack{j \\ k \neq j}} \sum_{\sigma_j, \sigma_k} Pr(\sigma_j, \sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) score_C(r_j, r_k)$$

$$= \sum_{\substack{j \\ k \neq j}} \sum_{\sigma_j, \sigma_k} Pr(j,k) Pr(\sigma_j, \sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) score_C(r_j, r_k)$$

The first equality follows since fix j, k, $score_C(r_j, r_k)$ does not depend on i and we also have $\sum_{\sigma_i} Pr(\sigma_i) Pr(\sigma_j, \sigma_k | \sigma_i) = Pr(\sigma_j, \sigma_k)$.

The second equality follows since for any $(j,k), j \neq k$ pair, there are $n-2$ numbers that are neither j nor k which means (j,k) will repeat $n-2$ times since there are $n-2$ possible i.

By definition we can see *ClassificationScore* = *Diversity* − *Inconsistency*. □

Claim 127. Every permutation strategy profile has the same *ClassificationScore*, *Diversity*, and *Inconsistency* with truth-telling.

Proof for Claim 127. Any permutation strategy profile's report profiles can be seen as a relabeling to truth-telling's report profiles, which implies the claim. □

Claim 128. The average agent-welfare in our *Disagreement Mechanism* is *ClassificationScore*

Proof for Claim 128. We only need to prove $\sum_i score_{\mathcal{M}}(i, \mathbf{r}) = 0$.

$$\begin{aligned}\sum_i score_{\mathcal{M}}(i, \mathbf{r}) =& \sum_{i \in A} score_{\mathcal{M}}(i, \mathbf{r}) + \sum_{i \in B} score_{\mathcal{M}}(i, \mathbf{r}) \\ =& \sum_{i \in A} \left(payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i, \mathbf{r}) - \frac{1}{|A|} \sum_{i \in B} payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i, \mathbf{r}) \right) \\ &+ \sum_{i \in B} \left(payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i, \mathbf{r}) - \frac{1}{|B|} \sum_{i \in A} payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i, \mathbf{r}) \right) = 0\end{aligned}$$

□

Claim 129. For any strategy profile s,

$Diversity(s) = TotalDivergence(s) \Leftrightarrow Inconsistency(s) = 0$

Proof for Claim 129. Note that

$$\begin{aligned}& TotalDivergence(s) - Diversity(s) \\ =& \sum_{\substack{j \\ k \neq j}} \sum_{\sigma_j, \sigma_k} Pr(j, k) Pr(\sigma_j, \sigma_k) \int_{\hat{j}, \hat{k}} Pr(\hat{j}, \hat{k}) \delta(\hat{\sigma}_j = \hat{\sigma}_k) D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)\end{aligned}$$

while

$$Inconsistency(s) = \sum_{\substack{j \\ k \neq j}} \sum_{\sigma_j, \sigma_k} Pr(j, k) Pr(\sigma_j, \sigma_k) \int_{\hat{j}, \hat{k}} Pr(\hat{j}, \hat{k}) \delta(\hat{\sigma}_j = \hat{\sigma}_k) \sqrt{D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)}$$

Because each part in $TotalDivergence(s) - Diversity(s)$ is non-negative, $TotalDivergence(s) - Diversity(s) = 0$ will imply $Pr(\sigma_j, \sigma_k) \int_{\hat{j}, \hat{k}} Pr(\hat{j}, \hat{k}) \delta(\hat{\sigma}_j = \hat{\sigma}_k) D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) = 0$.

So we have $Pr(\hat{j}, \hat{k}) \delta(\hat{\sigma}_j = \hat{\sigma}_k) = 0$ or $D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) = 0$ which implies $Pr(\sigma_j, \sigma_k) \int_{\hat{j}, \hat{k}} Pr(\hat{j}, \hat{k}) \delta(\hat{\sigma}_j = \hat{\sigma}_k) \sqrt{D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)} = 0$. The proof for another direction is similar. □

A.1.5 Proof of Theorem 81

Theorem 81. *[55]*

For any $\alpha, \beta > 0$ and any strictly proper scoring rule PS, $\mathcal{M}(\alpha, \beta, PS)$ has truth-telling as a strict Bayesian-Nash equilibrium whenever the prior Q is informative and symmetric.

Proof. We must show that for every agent, if other agents tell the truth, then this agent can (strictly) maximize his expected payoff if and only if he chooses to tell the truth.

Assume that all agents other than i are telling the truth. The probability that agent i is matched with agent j is $Pr(j) = \frac{1}{n-1}$. The expected payoff for agent i is:

$$\mathbb{E}[payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i, \mathbf{r})|\sigma_i] \tag{A.51}$$

$$= \sum_{j \neq i} (Pr(j)\mathbb{E}[\alpha score_P(r_i, r_j) + \beta score_I(r_i, r_j)|\sigma_i]) \tag{A.52}$$

$$= \sum_{j \neq i} \frac{1}{n-1}[\alpha PS(\mathbb{E}(\hat{\sigma}_j|\sigma_i), \hat{\mathbf{p}}_i) + \beta(-Pr(\hat{\sigma}_j = \hat{\sigma}_i|\sigma_i)\mathbb{E}[(PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_j) - PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_i))|\sigma_i, \hat{\sigma}_j = \hat{\sigma}_i]] \tag{A.53}$$

$$= \sum_{j \neq i} \frac{1}{n-1}[\alpha PS(\mathbb{E}(\sigma_j|\sigma_i), \hat{\mathbf{p}}_i) + \beta(Pr(\sigma_j = \hat{\sigma}_i|\sigma_i)(PS(\mathbf{q}_{\hat{\sigma}_i}, \hat{\mathbf{p}}_i) - PS(\mathbf{q}_{\hat{\sigma}_i}, \mathbf{q}_{\hat{\sigma}_i})))] \tag{A.54}$$

$$= \alpha PS(\mathbb{E}(\sum_{j \neq i} \frac{1}{n-1}\sigma_j|\sigma_i), \hat{\mathbf{p}}_i) + \sum_{j \neq i} \frac{\beta}{n-1}[(Pr(\sigma_j = \hat{\sigma}_i|\sigma_i)(PS(\mathbf{q}_{\hat{\sigma}_i}, \hat{\mathbf{p}}_i) - PS(\mathbf{q}_{\hat{\sigma}_i}, \mathbf{q}_{\hat{\sigma}_i})))] \tag{A.55}$$

$$= \alpha PS(\theta_{-i}\mathbf{q}_{\sigma_i}, \hat{\mathbf{p}}_i) + \sum_{j \neq i} \frac{\beta}{n-1}[(Pr(\sigma_j = \hat{\sigma}_i|\sigma_i)(PS(\mathbf{q}_{\hat{\sigma}_i}, \hat{\mathbf{p}}_i) - PS(\mathbf{q}_{\hat{\sigma}_i}, \mathbf{q}_{\hat{\sigma}_i})))] \tag{A.56}$$

$$= \alpha PS(\mathbf{q}_{\sigma_i}, \hat{\mathbf{p}}_i) + \sum_{j \neq i} \frac{\beta}{n-1}[(Pr(\sigma_j = \hat{\sigma}_i|\sigma_i)(PS(\mathbf{q}_{\hat{\sigma}_i}, \hat{\mathbf{p}}_i) - PS(\mathbf{q}_{\hat{\sigma}_i}, \mathbf{q}_{\hat{\sigma}_i})))] \tag{A.57}$$

From (A.52) to (A.53): When $\hat{\sigma}_i \neq \hat{\sigma}_j$, the information score is 0, so we only need

to consider the case $\hat{\sigma}_i = \hat{\sigma}_j$.

From (A.53) to (A.54): All agents other than i tell the truth, so $\hat{\sigma}_j = \sigma_j$ and

$$\hat{\mathbf{p}}_j = \mathbf{q}_{\sigma_j} = \mathbf{q}_{\hat{\sigma}_j} = \mathbf{q}_{\hat{\sigma}_i}.$$

From (A.54) to (A.55): The proper scoring rule is linear for the first entry.

From (A.55) to (A.56): Based on Claim 74, $E(\sum_{j \neq i} \frac{1}{n-1} \sigma_j | \sigma_i) = \theta_{-i} \mathbf{q}_{\sigma_i}$.

From (A.56) to (A.57): Note that for any $j \neq i$, agent j tells the truth so $\theta_{-i} = I$.

First, if agent i plays truthfully, then $\hat{\sigma}_i = \sigma_i, \hat{\mathbf{p}}_i = \mathbf{q}_{\sigma_i}$, and we will have $\mathbb{E}(payment(i, \mathcal{M}) | \sigma_i) = \alpha PS(\mathbf{q}_{\sigma_i}, \mathbf{q}_{\sigma_i})$ because $PS(\mathbf{q}_{\hat{\sigma}_i}, \hat{\mathbf{p}}_i) - PS(\mathbf{q}_{\hat{\sigma}_i}, \mathbf{q}_{\hat{\sigma}_i}) = 0$.

Now show that to receive a payment this high, agent i must play truthfully. Assume that
$\mathbb{E}(payment(i, \mathcal{M}) | \sigma_i) \geq \alpha PS(\mathbf{q}_{\sigma_i}, \mathbf{q}_{\sigma_i})$. First, the second term of Equation (A.57) is non-positive based on the property of proper scoring rule. Then we must have that $PS(\mathbf{q}_{\sigma_i}, \hat{\mathbf{p}}_i) \geq PS(\mathbf{q}_{\sigma_i}, \mathbf{q}_{\sigma_i})$, but because PS is a strictly proper scoring rule, this happens only if $\hat{\mathbf{p}}_i = \mathbf{q}_{\sigma_i}$. But this implies that the second term of Equations (A.57) equals 0, and this requires that $PS(\mathbf{q}_{\hat{\sigma}_i}, \hat{\mathbf{p}}_i) = PS(\mathbf{q}_{\hat{\sigma}_i}, \mathbf{q}_{\hat{\sigma}_i})$.

However, by the properties of strictly proper scoring rules, this means $\mathbf{q}_{\hat{\sigma}_i} = \hat{\mathbf{p}}_i$. However, we already showed that $\hat{\mathbf{p}}_i = \mathbf{p}_i = \mathbf{q}_{\sigma_i}$. Putting this together we see that $\mathbf{q}_{\hat{\sigma}_i} = \mathbf{q}_{\sigma_i}$. Based on the informative prior assumption, this implies that $\hat{\sigma}_i = \sigma_i$.

So we proved that for any agent i, when other agents tell the truth, agent i can obtain the best expected payoff if and only if he tells the truth which means truth-telling is a strict Bayesian-Nash equilibrium in the truthful mechanism. □

A.2 Proof of the truthfulness of BTS

Proof of Theorem 85 (i) [51]. When everyone else tells the truth, for every i, agent i will report truthful $\mathbf{p}_i$ to maximize her expected prediction score based on the

properties of log scoring rule. Thus, $\hat{\mathbf{p}}_i = \mathbf{p}_i$ for every i.

For the expected information score, we want to calculate the optimal σ agent i should report to maximize her expected information score when everyone else tells the truth, given that she receives $\Psi_i = \sigma_i$).

$$
\begin{aligned}
&\arg\max_{\sigma} \mathbb{E}_{\Psi_j, W|\Psi_i=\sigma_i} \log(\frac{Pr[\Psi_i = \sigma|W]}{Pr[\Psi_i = \sigma|\Psi_j]}) \\
&= \arg\max_{\sigma} \mathbb{E}_{\Psi_j, W|\Psi_i=\sigma_i} \log(\frac{Pr[\Psi_i = \sigma|W, \Psi_j]}{Pr[\Psi_i = \sigma|\Psi_j]}) && \text{(Conditional independence)} \\
&= \arg\max_{\sigma} \mathbb{E}_{\Psi_j, W|\Psi_i=\sigma_i} \log(\frac{Pr[\Psi_i = \sigma, W|\Psi_j]}{Pr[\Psi_i = \sigma|\Psi_j]Pr[W|\Psi_j]}) \\
&= \arg\max_{\sigma} \mathbb{E}_{\Psi_j, W|\Psi_i=\sigma_i} \log(\frac{Pr[W|\Psi_i = \sigma, \Psi_j]}{Pr[W|\Psi_j]}) \\
&= \arg\max_{\sigma} \mathbb{E}_{\Psi_j} L(\Pr[\boldsymbol{W}|\Psi_i = \sigma_i, \Psi_j], \Pr[\boldsymbol{W}|\Psi_i = \sigma, \Psi_j]) \\
&\qquad\qquad \text{(we can add } \log \Pr[W|\Psi_j] \text{ which is independent of } \sigma) \\
&= \sigma_i && (\arg\max_{\mathbf{q}} L(\mathbf{p}, \mathbf{q}) = \mathbf{p})
\end{aligned}
$$

Therefore, in BTS, for every i, agent i's best response is $(\sigma_i, \mathbf{p}_i)$ when everyone else tells the truth. BTS is truthful.

□

A.3 Expertise elicitation

Fact 12 (Information monotonicity of proper scoring rules)**.** Given any strictly proper scoring rule PS,

$$\mathbb{E}_{X,Y,Z} PS(Y, \Pr[\boldsymbol{Y}|X, Z]) \geq \mathbb{E}_{X,Y} PS(Y, \Pr[\boldsymbol{Y}|X]).$$

The equality holds if and only if $\Pr[\boldsymbol{Y}|X = x, Z = z] = \Pr[\boldsymbol{Y}|X = x]$ for all (x, z) where $\Pr[X = x, Z = z] > 0$.

Proof.

$$\begin{aligned}
\mathbb{E}_{X,Y} PS(Y, \Pr[\boldsymbol{Y}|X]) &= \sum_{x,y} \Pr[X=x, Y=y] PS(Y=y, \Pr[\boldsymbol{Y}|X=x]) \\
&= \sum_{x,y,z} \Pr[X=x, Y=y, Z=z] PS(Y=y, \Pr[\boldsymbol{Y}|X=x]) \\
&= \sum_{x,z} \Pr[X=x, Z=z] \\
&\quad * \sum_{y} \Pr[Y=y|X=x, Z=z] PS(Y=y, \Pr[\boldsymbol{Y}|X=x]) \\
&= \sum_{x,z} \Pr[X=x, Z=z] PS(\Pr[\boldsymbol{Y}|X=x, Z=z], \Pr[\boldsymbol{Y}|X=x]) \\
&\leq \sum_{x,z} \Pr[X=x, Z=z] PS(\Pr[\boldsymbol{Y}|X=x, Z=z], \Pr[\boldsymbol{Y}|X=x, Z=z]) \\
&\qquad\qquad (PS \text{ is strictly proper}) \\
&= \mathbb{E}_{X,Y,Z} PS(Y, \Pr[\boldsymbol{Y}|X, Z])
\end{aligned}$$

The equality holds if and only if $\Pr[\boldsymbol{Y}|X=x, Z=z] = \Pr[\boldsymbol{Y}|X=x]$ for all (x, z) where $\Pr[X=x, Z=z] > 0$ since PS is *striclty* proper.

□

Theorem 96. *With Assumption 88, 89, 90, Multi-HMIM($\{\alpha_m\}_m$) is truthful; moreover, when $\{\alpha_m\}_m$ are potent for Multi-HMIM($\{\alpha_m\}_m$), Multi-HMIM($\{\alpha_m\}_m$) is potent and truthful.*

Proof. Since we assume all tasks are a priori similar, without loss of generality, we can assume every agent uses the same (possibly mixed) report and (possibly mixed) effort strategy for all tasks.

Truthful We divide the proof into two parts. For each agent i, given that she believes other agents report honestly (may not report all signals they have), we will show (1) conditioning on agent i playing pure effort strategy, she should maximize her

payment as well as the utility by playing truthful strategy; (2) it's better for agent i to play pure effort strategy—performing the same method all the time—than mixed effort strategy.

Part (1). We want to show that for each agent i who always perform method m_i, given other agents honestly report their methods and signals, for each $m \preceq m_i$, she should honestly report her real signal ψ_i^m to maximize her expected information score in m's level, that is,

$$\mathbb{E}[2\alpha_m Corr(\hat{\boldsymbol{\psi}}_i^m; \hat{\boldsymbol{\psi}}_{-i}^m | \{\hat{\boldsymbol{\psi}}_{-i}^{m'}\}_{m' \prec m}].$$

Since we assume other agents report honestly and we have assumed that the signals agents receive for every method are homogeneous, we replace $\hat{\boldsymbol{\psi}}_{-i}^m, \hat{\boldsymbol{\psi}}_{-i}^{m'}$ by $\boldsymbol{\psi}_{-i}^m, \boldsymbol{\psi}_{-i}^{m'}$.

When we run algorithm 1 to calculate $Corr(\hat{\boldsymbol{\psi}}_i^m; \hat{\boldsymbol{\psi}}_{-i}^m | \{\hat{\boldsymbol{\psi}}_{-i}^{m'}\}_{m' \prec m})$, in the situation the algorithm does not return "success"—situation 0—her information score in m's level is 0 regardless of agent i reports for method m's output. In the situation the algorithm returns "success", either it runs $Corr(\hat{\boldsymbol{\psi}}_i^m; \hat{\boldsymbol{\psi}}_{-i}^m)$ and returns "success"—situation 1—or it runs $Corr(\hat{\boldsymbol{\psi}}_i^m(D); \hat{\boldsymbol{\psi}}_{-i}^m(D))$ and returns "success"—situation 2.

For each task, each m, fixing agents' choices for whether to provide a signal or $\emptyset$, the situation which the algorithm runs in is fixed as well. We only need to consider each situation separately.

Claim 141. Given that other agents report honestly, for each agent i who always perform m_i, for all $m \preceq m_i$, when agent i honestly reports method m's output, her expected information score in m's level per each reward task is

$$\alpha_m MI^{tvd}(\Psi_i^m; \Psi_{-i}^m)$$

in situation 1;

$$\alpha_m MI^{tvd}(\Psi_i^m; \Psi_{-i}^m | \{\Psi_{-i}^{m'}\}_{m' \prec m})$$

in situation 2.

Claim 142. Given that other agents report honestly, for each agent i, when agent i reports method m's output as $\hat{\psi}_i^m$, her expected information score in m's level per each reward task is $\leq$

$$\alpha_m MI^{tvd}(\hat{\Psi}_i^m; \Psi_{-i}^m)$$

in situation 1;

$$\alpha_m MI^{tvd}(\hat{\Psi}_i^m; \Psi_{-i}^m | \{\Psi_{-i}^{m'}\}_{m' \prec m})$$

in situation 2. The equality holds if $\hat{\Psi}_i^m$ is positively correlated with Ψ_{-i}^m (conditioning on $\{\Psi_{-i}^{m'}\}_{m' \prec m}$).

Once we show the above two claims. Since

$$\begin{aligned}
&MI^{tvd}(\hat{\Psi}_i^m; \Psi_{-i}^m | \{\Psi_{-i}^{m'}\}_{m' \prec m}) \\
&= MI^{tvd}(f_m(\{\Psi_i^{m'}\}_{m' \preceq m_i}); \Psi_{-i}^m | \{\Psi_{-i}^{m'}\}_{m' \prec m}) \\
&\qquad\qquad \text{(Agent } i \text{ uses the report strategy } f_m \text{ to report } m\text{'s output)} \\
&\leq MI^{tvd}(\{\Psi_i^{m'}\}_{m' \preceq m_i}; \Psi_{-i}^m | \{\Psi_{-i}^{m'}\}_{m' \prec m}) \qquad \text{(Information Monotonicity of } MI^f\text{)} \\
&= MI^{tvd}(\Psi_i^m; \Psi_{-i}^m | \{\Psi_{-i}^{m'}\}_{m' \prec m}) \qquad \text{(Assumption 90)}
\end{aligned}$$

and similarly $MI^{tvd}(\hat{\Psi}_i^m; \Psi_{-i}^m) \leq MI^{tvd}(\Psi_i^m; \Psi_{-i}^m)$. Part (1) follows immediately.

Part (2). This part is implied by the complexity of MI^{tvd}. We give a formal proof here. We consider situation 1 here. For any $0 \leq \lambda \leq 1$, any two methods m_1, m_2, if agent i perform method m_1 with probability λ, method m_2 with probability $1 - \lambda$, for every m, agent i's utility in m's level is less than

$$
\begin{aligned}
&MI^{tvd}(\hat{\Psi}_i^m; \Psi_{-i}^m) - (\lambda h_i(m_1) + (1-\lambda)h_i(m_2)) \\
&\leq \max_{f_m} MI^{tvd}(f_m(\text{her received signals}); \Psi_{-i}^m) - (\lambda h_i(m_1) + (1-\lambda)h_i(m_2)) \\
&= MI^{tvd}(f_m^*(\text{her received signals}); \Psi_{-i}^m) - (\lambda h_i(m_1) + (1-\lambda)h_i(m_2)) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (f_m^* \text{ is the optimal report strategy.}) \\
&\leq \lambda(MI^{tvd}(f_m^*(\{\Psi_i^{m'}\}_{m' \preceq m_1}); \Psi_{-i}^m) - h_i(m_1)) + (1-\lambda)(MI^{tvd}(f_m^*(\{\Psi_i^{m'}\}_{m' \preceq m_2}); \Psi_{-i}^m) - h_i(m_2)) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{Convexity of } MI^f) \\
&\leq \max\{MI^{tvd}(f_m^*(\{\Psi_i^{m'}\}_{m' \preceq m_1}); \Psi_{-i}^m) - h_i(m_1), MI^{tvd}(f_m^*(\{\Psi_i^{m'}\}_{m' \preceq m_2}); \Psi_{-i}^m) - h_i(m_2)\}
\end{aligned}
$$

in situation 1. Without loss of generality, we assume

$$
MI^{tvd}(f_m^*(\{\Psi_i^{m'}\}_{m' \preceq m_1}); \Psi_{-i}^m) - h_i(m_1) \geq MI^{tvd}(f_m^*(\{\Psi_i^{m'}\}_{m' \preceq m_2}); \Psi_{-i}^m) - h_i(m_2).
$$

Then

$$
\begin{aligned}
&MI^{tvd}(\hat{\Psi}_i^m; \Psi_{-i}^m) - (\lambda h_i(m_1) + (1-\lambda)h_i(m_2)) \\
&\leq MI^{tvd}(f_m^*(\{\Psi_i^{m'}\}_{m' \preceq m_1}); \Psi_{-i}^m) - h_i(m_1) \\
&\leq \max_{f_m} MI^{tvd}(f_m(\{\Psi_i^{m'}\}_{m' \preceq m_1}); \Psi_{-i}^m) - h_i(m_1)
\end{aligned}
$$

The analysis for situation 2 is similar. With the positively correlated guess assumption (Assumption 92) and Claim 142, we know $\max_{f_m} MI^{tvd}(f_m(\{\Psi_i^{m'}\}_{m' \preceq m_1}); \Psi_{-i}^m)$ can be obtained by agent i in Multi-HMIM by always performing m_1 and playing a proper report strategy. Thus, agent i cannot obtain better utility by playing mixed effort strategy.

Potent We can follow the proof of truthful property and additionally show that when the coefficients are potent , for each agent i, when she believes others agents play prudent strategy, agent i should pick the effort strategy defined by the prudent strategy as her optimal effort strategy. When the coefficients are potent , based on the definition of potent coefficients, for each agent i, when she believe other agents play prudent strategy, for each task she finished, there must exists another agent who finished the same task with her, using the method that is higher or equal to her. Thus, agent i's all tasks are reward tasks for her, and algorithm 1 will always run into situation 2 since the mechanism always has access to all levels of information. With the positively correlated guess assumption (Assumption 92) and Claim 142, agent i's optimal utility is proportional to

$$\sum_{m\in M}\max_{f_m:\Pi_{\ell\preceq m_i}\Sigma_\ell\mapsto\Sigma_m}\alpha_m MI^{tvd}(f_m(\{\Psi_i^\ell\}_{\ell\preceq m_i});\Psi_{-i}^m|\{\Psi_{-i}^{m'}\}_{m'\prec m})-h_i(m_i)$$

by always performing method m_i. Thus, agent i's optimal effort strategy should be the effort strategy defined by the prudent strategy, given that she believes other agents play prudent strategy. □

Theorem 100. *With Assumption 88, Learning based multi-HMIM is dominant truthful.*

Moreover, with Assumption 97, when the rule $\mathcal{RULE}$ is potent , Learning based multi-HMIM is potent , dominant truthful and will output the hierarchical information structure as well as the maximal level(s) answer vector given that agents play prudent strategy.

Proof for Theorem 100. Since we assume all tasks are a priori similar, without loss of generality, we can assume every agent use the same report and effort strategy for all tasks.

In order to show the dominant truthful property, we will show for each agent, fixing any other agents' strategies, (1) conditioning on using pure effort strategy, she can maximize her payment as well as the utility by reporting her received information honestly; (2) pure effort strategy has higher utility than mixed effort strategy.

Part (1). Even if the mechanism clusters incorrectly, part (1) still follows directly from the information monotonicity property of f-mutual information MI^f.

Part (2). The proof here is the same with the part (2) proof in Theorem 96. We give a formal proof here.

For any $0 \leq \lambda \leq 1$, any two methods m_1, m_2, if agent i perform method m_1 with probability λ, method m_2 with probability $1-\lambda$, agent i's utility in m's level is

$$
\begin{aligned}
& MI^f(\text{her reported signals}; \hat{\Psi}_{-i}^m | \{\hat{\Psi}_{-i}^{m'}\}_{m' \prec m, m' \in M_{-i}}) - (\lambda h_i(m_1) + (1-\lambda) h_i(m_2)) \\
\leq & MI^f(\text{her received signals}; ; \hat{\Psi}_{-i}^m | \{\hat{\Psi}_{-i}^{m'}\}_{m' \prec m, m' \in M_{-i}}) - (\lambda h_i(m_1) + (1-\lambda) h_i(m_2)) \\
\leq & \lambda (MI^f(\{\Psi_i^{m'}\}_{m' \preceq m_1}; \hat{\Psi}_{-i}^m | \{\hat{\Psi}_{-i}^{m'}\}_{m' \prec m, m' \in M_{-i}}) - h_i(m_1)) \qquad \text{(convexity of } MI^f\text{)} \\
& + (1-\lambda)(MI^f(\{\Psi_i^{m'}\}_{m' \preceq m_2}; \hat{\Psi}_{-i}^m | \{\hat{\Psi}_{-i}^{m'}\}_{m' \prec m, m' \in M_{-i}}) - h_i(m_2)) \\
\leq & \max\{MI^f(\{\Psi_i^{m'}\}_{m' \preceq m_1}; \hat{\Psi}_{-i}^m | \{\hat{\Psi}_{-i}^{m'}\}_{m' \prec m, m' \in M_{-i}}) - h_i(m_1), \\
& MI^f(\{\Psi_i^{m'}\}_{m' \preceq m_2}; \hat{\Psi}_{-i}^m | \{\hat{\Psi}_{-i}^{m'}\}_{m' \prec m, m' \in M_{-i}}) - h_i(m_2)\}
\end{aligned}
$$

Thus, each agent i cannot obtain higher utility by playing a mixed effort strategy.

It remains to show the potent property. When the rule is potent , for each agent i, when she believes other agents play prudent strategy, the mechanism must have access to all levels of honest answer vectors due to the definition of prudent strategy and potent rule. With Assumption 97, the mechanism can correctly learn the whole hierarchical information structure without agent i's report and use coefficients

$\boldsymbol{\alpha}(\mathcal{RULE})$. Thus, her optimal payment for performing method m_i will be

$$\sum_{m \in M} \alpha_m MI^f(\{\Psi_i^\ell\}_{\ell \preceq m_i}; \Psi_{-i}^m | \{\Psi_{-i}^{m'}\}_{m' \prec m})$$

due to the information monotonicity of MI^f. In this case, her optimal strategy is her prudent strategy. Therefore, learning based Multi-HMIM is potent and will output the correct hierarchical information structure as well as the maximal level(s) answer vector(s) when agents play prudent strategy.

□

Theorem 106. *With Assumption 101, single-HMIM is strictly truthful; moreover, when the coefficients is potent for single-HMIM, single-HMIM is potent and strictly truthful.*

for Theorem 106. For each agent i, her highest information score is 0. When she believes all other agents honestly report their signals and predictions, she can obtain her highest prediction score via providing her truthful prediction based on the property of the strictly proper scoring rule. While during the same time, she can obtain 0 (the highest) information score according to the common prior assumption. If agent i tell lies about her predictions, in expectation she will receive strictly lower prediction score since PS is strictly proper. If she honestly provides her predictions but lie for the signals, then she will be punished for her information score with positive probability. Therefore, when agent i believes everyone else tells the truth, honestly reporting her truthful signals and predictions strictly maximize her payment.

It remains to show the potent property. In Single-HMIM, when the coefficients are potent , for each agent i, when she believes other agents play prudent strategy, for each m, there must exist a reference agent for agent i who reports method m's

output. Thus agent i's optimal expected payment by performing method m_i is

$$\sum_{m \in M} \alpha_m \mathbb{E}_{Q_m}[PS(\sigma^m, p^m_{m_i})]$$

since her optimal information score is always 0. In this case, agent i's optimal strategy is prudent for her. Therefore, Single-HMIM is potent . □

Claim 141. Given that other agents report honestly, for each agent i who always perform m_i, for all $m \preceq m_i$, when agent i honestly reports method m's output, her expected information score in m's level per each reward task is

$$\alpha_m MI^{tvd}(\Psi^m_i; \Psi^m_{-i})$$

in situation 1;

$$\alpha_m MI^{tvd}(\Psi^m_i; \Psi^m_{-i} | \{\Psi^{m'}_{-i}\}_{m' \prec m})$$

in situation 2.

Proof for Claim 141. We first show

$$\mathbb{E}[Corr(\boldsymbol{\psi}^m_i; \boldsymbol{\Psi}^m_{-i})] = \frac{1}{2} MI^{tvd}(\Psi^m_i; \Psi^m_{-i}).$$

$$\frac{1}{2}MI^{tvd}(\Psi_i^m;\Psi_{-i}^m) = \frac{1}{2}\sum_{\sigma,\sigma'} |\Pr[\Psi_i^m=\sigma,\Psi_{-i}^m=\sigma'] - \Pr[\Psi_i^m=\sigma]\Pr[\Psi_{-i}^m=\sigma']|$$

(Definition of MI^{tvd})

$$= \frac{1}{2}\sum_{\sigma,\sigma'} \mathbb{1}(\sigma=\sigma')\left(\Pr[\Psi_i^m=\sigma,\Psi_{-i}^m=\sigma'] - \Pr[\Psi_i^m=\sigma]\Pr[\Psi_{-i}^m=\sigma']\right)$$

$$+ \mathbb{1}(\sigma\neq\sigma')\left(\Pr[\Psi_i^m=\sigma]\Pr[\Psi_{-i}^m=\sigma'] - \Pr[\Psi_i^m=\sigma,\Psi_{-i}^m=\sigma']\right)$$

(Assumption 89)

$$= \sum_{\sigma}\left(\Pr[\Psi_i^m=\sigma,\Psi_{-i}^m=\sigma] - \Pr[\Psi_i^m=\sigma]\Pr[\Psi_{-i}^m=\sigma]\right)$$

(Combining like terms, $\Pr[E]-\Pr[\neg E] = 2\Pr[E]-1$)

$$= \mathbb{E}[Corr(\boldsymbol{\psi}_i^m;\boldsymbol{\psi}_{-i}^m)] \quad \text{(see Algorithm 1)}$$

To show

$$\mathbb{E}[Corr(\boldsymbol{\psi}_i^m;\boldsymbol{\psi}_{-i}^m|\{\boldsymbol{\psi}_{-i}^{m'}\}_{m'\prec m})] = \frac{1}{2}MI^{tvd}(\Psi_i^m;\Psi_{-i}^m|\{\Psi_{-i}^{m'}\}_{m'\prec m}),$$

we only need to replace every $\Pr[\cdot]$ in the above equations by $\Pr[\cdot|\{\Psi_{-i}^{m'}\}_{m'\prec m} = \{\sigma^{m'}\}_{m'\prec m}]$ with putting $\sum_{\{\sigma^{m'}\}_{m'\prec m}}\Pr[\{\Psi_{-i}^{m'}\}_{m'\prec m}=\{\sigma^{m'}\}_{m'\prec m}]$ ahead. Note that assumption 89 can be applied to this case as well. □

Claim 142. Given that other agents report honestly, for each agent i, when agent i reports method m's output as $\hat{\psi}_i^m$, her expected information score in m's level per each reward task is $\leq$

$$\alpha_m MI^{tvd}(\hat{\Psi}_i^m;\Psi_{-i}^m)$$

in situation 1;

$$\alpha_m MI^{tvd}(\hat{\Psi}_i^m;\Psi_{-i}^m|\{\Psi_{-i}^{m'}\}_{m'\prec m})$$

in situation 2. The equality holds if $\hat{\Psi}_i^m$ is positively correlated with Ψ_{-i}^m.

Proof for Claim 142. The proof is similar with the proof of Claim 141. We only need to replace Ψ_i^m by $\hat{\Psi}_i^m$ and change the second equation to greater than, that is,

$$\begin{aligned}
&\frac{1}{2}\sum_{\sigma,\sigma'} |\Pr[\Psi_i^m = \sigma, \Psi_{-i}^m = \sigma'] - \Pr[\Psi_i^m = \sigma]\Pr[\Psi_{-i}^m = \sigma']| \\
&\geq \frac{1}{2}\sum_{\sigma,\sigma'} \mathbb{1}(\sigma = \sigma')\left(\Pr[\Psi_i^m = \sigma, \Psi_{-i}^m = \sigma'] - \Pr[\Psi_i^m = \sigma]\Pr[\Psi_{-i}^m = \sigma']\right) \\
&+ \mathbb{1}(\sigma \neq \sigma')\left(\Pr[\Psi_i^m = \sigma]\Pr[\Psi_{-i}^m = \sigma'] - \Pr[\Psi_i^m = \sigma, \Psi_{-i}^m = \sigma']\right). \quad (\textstyle\sum |x| \geq \sum x)
\end{aligned}$$

Note that the equality holds if $\hat{\Psi}_i^m$ is positively correlated with Ψ_{-i}^m. Follow the same proof of Claim 141, we finish the proof. □

A.4 Forecast elicitation and an information aggregation problem: co-training

Claim 110. When random variables X_A, X_B are independent conditioning on Y,

$$\begin{aligned}
K(X_A = x_A, X_B = x_B) &= \sum_y \Pr[Y = y] K(X_A = x_A, Y = y) K(X_B = x_B, Y = y) \\
&= \sum_y \Pr[Y = y | X_A = x_A] K(X_B = x_B, Y = y) \\
&= \sum_y \frac{\Pr[Y = y | X_A = x_A]\Pr[Y = y | X_B = x_B]}{\Pr[Y = y]}.
\end{aligned}$$

Proof.

$$
\begin{aligned}
K(X_A = x_A, X_B = x_B) =& \frac{\Pr[X_A = x_A, X_B = x_B]}{\Pr[X_A = x_A]\Pr[X_B = x_B]} \\
=& \frac{\sum_y \Pr[Y = y]\Pr[X_A = x_A, X_B = x_B | Y = y]}{\Pr[X_A = x_A]\Pr[X_B = x_B]} \\
=& \frac{\sum_y \Pr[Y = y]\Pr[X_A = x_A | Y = y]\Pr[X_B = x_B | Y = y]}{\Pr[X_A = x_A]\Pr[X_B = x_B]} \\
& \qquad \text{(Conditional independence)} \\
=& \sum_y \Pr[Y = y] K(X_A = x_A, Y = y) K(X_B = x_B, Y = y) \\
& \qquad \text{(PMI=posterior/prior)} \\
=& \sum_y \Pr[Y = y | X_A = x_A] K(X_B = x_B, Y = y) \\
=& \sum_y \frac{\Pr[Y = y | X_A = x_A]\Pr[Y = y | X_B = x_B]}{\Pr[Y = y]}.
\end{aligned}
$$

□

Theorem 119. *Given the prior distribution over the Y, with the conditional independence assumption, with a priori similar and random order assumption, when $\max\{|\mathcal{L}_A|, |\mathcal{L}_B|\} \geq 2$ and the prior is stable and well-defined, when the convex function f is differentiable and f' is invertible, $MCG(f)$ is focal.*

When both Alice and Bob are honest, each of them's expected payment in $MCG(f)$ is

$$MI^f(X_A; X_B).$$

Proof. Given that Alice's strategy is s_A and Bob's strategy is s_B, with the a priori similar and random order assumption, we represent agents' report as the output (possibly being random) of their strategy operating on the private information.

We start to show $MCG(f)$ is strictly truthful. Given that Alice is honest, based on Lemma 120, Bob will maximize his expected payment if and only if $\forall \ell_1, \ell_2$,

$$R(\mathbf{p}_{x_A{}^{\ell_1}}, s_B(x_B^{\ell_2})) = f'(K(x_A^{\ell_1}, x_B^{\ell_2})).$$

Note that in $MCG(f)$,

$$R(\mathbf{p}_{x_A{}^{\ell_1}}, s_B(x_B^{\ell_2})) = f'(\sum_y \frac{\mathbf{p}_{x_A{}^{\ell_1}}(y) s_B(x_B^{\ell_2})(y)}{\Pr[Y=y]})$$

Since the prior is stable, the above equation is satisfied for all possible $x_A{}^{\ell_1}$ if and only if Bob tells the truth—reporting $\mathbf{p}_{x_B{}^{\ell_2}}$. Therefore, $MCG(f)$ is strictly truthful.

It remains to show $MCG(f)$ pays truth-telling the most and strictly better than any other non-permutation strategy profile. When agents maximize the expected payment,

$$R(s_A(x_A{}^{\ell_1}), s_B(x_B^{\ell_2})) = f'(K(x_A^{\ell_1}, x_B^{\ell_2})).$$

Recall that we defined

$$R(s_A(x_A{}^{\ell_1}), s_B(x_B^{\ell_2})) = f'(\sum_y \frac{s_A(x_A{}^{\ell_1})(y) s_B(x_B^{\ell_2})(y)}{\Pr[Y=y]}).$$

Thus, when f' is invertible, we have

$$\sum_y \frac{s_A(x_A{}^{\ell_1})(y) s_B(x_B^{\ell_2})(y)}{\Pr[Y=y]} = K(x_A^{\ell_1}, x_B^{\ell_2})$$

for any $x_A^{\ell_1}, x_B^{\ell_2}$. This is exactly system (6.1).

With the conditional independence assumption, when agents tell the truth, the

above system will be satisfied. Therefore, agents can maximize their expected payment via truth-telling. Moreover, when the prior is well-defined, if the prior $\Pr[Y]$ is a uniform distribution, then any permutation strategy profile can solve the above system and as well as maximize agents' expected payment. Even if the prior $\Pr[Y]$ is not a uniform distribution, although not all permutation strategy profiles solve the above system, still any solution of the above system must correspond to a permutation strategy profile, given the prior is well-defined. Therefore, when agents maximize their expected payment, their strategy profile must be a permutation strategy profile or truth-telling, which implies $MCG(f)$ is focal.

□

Lemma 120. *With the conditional independence assumption, the expected total payment is maximized over Alice and Bob's strategies if and only if* $\forall \ell_1 \in \mathcal{L}_A, \ell_2 \in \mathcal{L}_B$, *for any* $(x_A^{\ell_1}, x_B^{\ell_2}) \in \Sigma_A \times \Sigma_B$,

$$R(\hat{\mathbf{p}}^{\ell_1}_{x_A{}^{\ell_1}}, \hat{\mathbf{p}}^{\ell_2}_{x_B{}^{\ell_2}}) = f'(K(x_A^{\ell_1}, x_B^{\ell_2})).$$

The maximum is

$$MI^f(X_A; X_B).$$

Proof. Without loss of generality, it is sufficient to analyze Alice's strategy and report. With the a priori similar and random order assumption, $\hat{\mathbf{p}}^{\ell_1}_{x_A{}^{\ell_1}}$ can be represented as $s_A(x_A{}^{\ell_1})$ since the index of the task ℓ_1 is meaningless to Alice when all tasks appear in a random order, independently drawn for each agent. The strategy can be seen as a random predictor. Thus, we can use the same proof of Lemma 114 to prove Lemma 120. □

APPENDIX B

Mutual information calculations

We show the calculations for the mutual information table.

For the length signal, since agents has no uncertainty for this signal, the mutual information between agent i's length signal and agent $j \neq i$'s length signal will be the entropy of length signal. Recall that we have assumed an essay has long length with probability 0.5. Thus,

$$MI(length; length) = 0.5 * \log(0.5) + 0.5 * \log(0.5) = 0.6931$$

Since an essay's length is independent with its writing and quality, we have the mutual information between the length signal and writing signal, quality, writing conditioning length, quality conditioning writing and length are all zero.

$\Pr[\Psi_i^{m_w} = ☺, \Psi_j^{m_w} = ☺] = 0.5 * 0.9 * 0.9 + 0.5 * 0.1 * 0.1 = 0.41$

$\Pr[\Psi_i^{m_w} = ☺, \Psi_j^{m_w} = ☹] = \Pr[\Psi_i^{m_w} = ☹; \Psi_j^{m_w} = ☺] = 0.5*0.9*0.1+0.5*0.1*0.9 = 0.09$

$\Pr[\Psi_i^{m_w} = ☹, \Psi_j^{m_w} = ☹] = 0.5 * 0.1 * 0.1 + 0.5 * 0.9 * 0.9 = 0.41$

We can put the above joint distribution over $(\Psi_i^{m_w}; \Psi_j^{m_w})$ to the formula $MI(X;Y) = \sum_{x,y} \Pr[X = x, Y = y] \log \frac{\Pr[X=x,Y=y]}{\Pr[X=x]\Pr[Y=y]}$ and obtain

$$MI(length, writing; writing)$$

$$= MI(writing; writing)$$

$$= MI(\Psi_i^{m_w}; \Psi_j^{m_w}) = 0.2218$$

Note that MI(writing;writing) is not the entropy of the writing signal since it is the mutual information between different agents' writing signals.

Similarly, we can calculate the joint distribution over $(\Psi_i^{m_q}, \Psi_i^{m_w}, \Psi_j^{m_q}, \Psi_j^{m_w})$ and set ☹ = 0 and ☺ = 1:

$$\Pr[\Psi_i^{m_q} = a, \Psi_i^{m_w} = b, \Psi_j^{m_q} = c, \Psi_j^{m_w} = d]$$

$$= 0.4 * 0.3^a * 0.7^{1-a} * 0.1^b * 0.9^{1-b} * 0.3^c * 0.7^{1-c} * 0.1^d * 0.9^{1-d}$$

(when the essay has bad quality, bad writing:)

$$+ 0.1 * 0.3^a * 0.7^{1-a} * 0.9^b * 0.1^{1-b} * 0.3^c * 0.7^{1-c} * 0.9^d * 0.1^{1-d}$$

(when the essay has bad quality, good writing:)

$$+ 0.1 * 0.7^a * 0.3^{1-a} * 0.1^b * 0.9^{1-b} * 0.7^c * 0.3^{1-c} * 0.1^d * 0.9^{1-d}$$

(when the essay has good quality, bad writing:)

$$+ 0.4 * 0.7^a * 0.3^{1-a} * 0.9^b * 0.1^{1-b} * 0.7^c * 0.3^{1-c} * 0.9^d * 0.1^{1-d}$$

(when the essay has good quality, good writing:)

The fact that the length signal is independent with writing and quality will ease the calculation a lot since we can ignore the length signal if it only shows in one side when we calculate the mutual information. Moreover, since the length signal has no uncertainty, length—length will be a value without uncertainty and can be ignored in the calculation of mutual information.

Aided by the calculator, we can obtain

$$
\begin{aligned}
& MI(length, writing; quality) \\
=& MI(writing; quality) \\
=& MI(\Psi_i^{m_w}; \Psi_j^{m_q}) = 0.0185
\end{aligned}
$$

$$
\begin{aligned}
& MI(length, writing; writing|length) \\
=& MI(writing; writing) = 0.2218;
\end{aligned}
$$

$$
\begin{aligned}
& MI(length, writing, quality; writing) \\
=& MI(quality, writing; writing) \\
=& MI(\Psi_i^{m_w}, \Psi_i^{m_q}; \Psi_j^{m_w}) = 0.2259
\end{aligned}
$$

$$
\begin{aligned}
& MI(length, writing; quality|writing, length) \\
=& MI(writing; quality|writing) \\
=& MI(writing, quality; writing) - MI(writing; writing) = 0.2259 - 0.2218 = 0.0041
\end{aligned}
$$

$$
\begin{aligned}
& MI(length, writing, quality; quality) \\
=& MI(quality, writing; quality) \\
=& MI(\Psi_i^{m_w}, \Psi_i^{m_q}; \Psi_j^{m_q}) = 0.0267
\end{aligned}
$$

$$
\begin{aligned}
& MI(length, writing, quality; writing|length) \\
=& MI(quality, writing; writing) = 0.2259
\end{aligned}
$$

$$
\begin{aligned}
& MI(length, writing, quality; quality|writing, length) \\
=& MI(writing, quality; quality|writing) \\
=& MI(writing, quality; quality, writing) - MI(writing, quality; writing) \\
=& 0.2374 - 0.2259 = 0.0115
\end{aligned}
$$

$$
\begin{aligned}
& MI(length, writing, quality; length, writing) \\
=& MI(length, writing, quality; length) + MI(length, writing, quality; writing|length) \\
=& MI(length; length) + MI(length, writing, quality; writing|length) \\
=& 0.6931 + 0.2259 = 0.9190
\end{aligned}
$$

$$
\begin{aligned}
& MI(length, writing, quality; length, writing, quality) \\
=& MI(length, writing, quality; length) + MI(length, writing, quality; writing|length) \\
& + MI(length, writing, quality; quality|writing, length) \\
=& 0.6931 + 0.2259 + 0.0115 = 0.9305
\end{aligned}
$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Arpit Agarwal and Shivani Agarwal. "On consistent surrogate risk minimization and property elicitation". In: *Conference on Learning Theory*. 2015, pp. 4–22.

[2] Arpit Agarwal et al. "Peer Prediction with Heterogeneous Users". In: *Proceedings of the 2017 ACM Conference on Economics and Computation*. EC '17. Cambridge, Massachusetts, USA: ACM, 2017, pp. 81–98. ISBN: 978-1-4503-4527-9. DOI: 10.1145/3033274.3085127. URL: http://doi.acm.org/10.1145/3033274.3085127.

[3] Syed Mumtaz Ali and Samuel D Silvey. "A general class of coefficients of divergence of one distribution from another". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1966), pp. 131–142.

[4] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. Vol. 191. American Mathematical Soc., 2007.

[5] S-I Amari and A Cichocki. "Information geometry of divergence functions". In: *Bulletin of the Polish Academy of Sciences: Technical Sciences* 58.1 (2010), pp. 183–195.

[6] Dana Angluin and Philip Laird. "Learning from noisy examples". In: *Machine Learning* 2.4 (1988), pp. 343–370.

[7] Suzanna Becker. "Mutual information maximization: models of cortical self-organization". In: *Network: Computation in neural systems* 7.1 (1996), pp. 7–31.

[8] Anthony J Bell and Terrence J Sejnowski. "An information-maximization approach to blind separation and blind deconvolution". In: *Neural computation* 7.6 (1995), pp. 1129–1159.

[9] Avrim Blum and Tom Mitchell. "Combining labeled and unlabeled data with co-training". In: *Proceedings of the eleventh annual conference on Computational learning theory*. ACM. 1998, pp. 92–100.

[10] Lev M Bregman. "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming". In: *USSR computational mathematics and mathematical physics* 7.3 (1967), pp. 200–217.

[11] Yang Cai, Constantinos Daskalakis, and Christos H Papadimitriou. "Optimum statistical estimation with strategic data sources". In: *arXiv preprint arXiv:1408.2539* (2014).

[12] J-F Cardoso. "Infomax and maximum likelihood for blind source separation". In: *IEEE Signal processing letters* 4.4 (1997), pp. 112–114.

[13] Jesús Cid-Sueiro. "Proper losses for learning from partial labels". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1565–1573.

[14] Michael Collins and Yoram Singer. "Unsupervised models for named entity classification". In: *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. 1999.

[15] Thomas M Cover and Joy A Thomas. "Elements of information theory 2nd edition". In: (2006).

[16] Imre Csiszár, Paul C Shields, et al. "Information theory and statistics: A tutorial". In: *Foundations and Trends® in Communications and Information Theory* 1.4 (2004), pp. 417–528.

[17] Nilesh Dalvi et al. "Aggregating crowdsourced binary ratings". In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 285–294.

[18] Anirban Dasgupta and Arpita Ghosh. "Crowdsourced judgement elicitation with endogenous proficiency". In: *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2013, pp. 319–330.

[19] Sanjoy Dasgupta, Michael L Littman, and David A McAllester. "PAC generalization bounds for co-training". In: *Advances in neural information processing systems*. 2002, pp. 375–382.

[20] A. P. Dawid and A. M. Skene. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 20–28. ISSN: 00359254, 14679876. URL: `http://www.jstor.org/stable/2346806`.

[21] Boi Faltings et al. "Incentives to counter bias in human computation". In: *Second AAAI Conference on Human Computation and Crowdsourcing*. 2014.

[22] A. Gao, J. R. Wright, and K. Leyton-Brown. "Incentivizing Evaluation via Limited Access to Ground Truth: Peer-Prediction Makes Things Worse". In: *ArXiv e-prints* (June 2016). arXiv: `1606.07042 [cs.GT]`.

[23] Arpita Ghosh, Satyen Kale, and Preston McAfee. "Who moderates the moderators?: crowdsourcing abuse detection in user-generated content". In: *Proceedings of the 12th ACM conference on Electronic commerce*. ACM. 2011, pp. 167–176.

[24] Tilmann Gneiting and Adrian E Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378.

[25] Sharad Goel, Daniel M. Reeves, and David M. Pennock. "Collective revelation: A mechanism for self-verified, weighted, and truthful predictions". In: *Proceedings of the 10th ACM conference on Electronic commerce (EC 2009)*. Stanford, California, USA, 2009. ISBN: 978-1-60558-458-4.

[26] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[27] Harold V Henderson and Shayle R Searle. "The vec-permutation matrix, the vec operator and Kronecker products: A review". In: *Linear and multilinear algebra* 9.4 (1981), pp. 271–288.

[28] Radu Jurca and Boi Faltings. "Collusion-resistant, incentive-compatible feedback payments". In: *Proceedings of the 8th ACM conference on Electronic commerce*. ACM. 2007, pp. 200–209.

[29] Radu Jurca and Boi Faltings. "Mechanisms for making crowds truthful". In: *J. Artif. Int. Res.* 34.1 (Mar. 2009).

[30] Sham M Kakade and Dean P Foster. "Multi-view regression via canonical correlation analysis". In: *International Conference on Computational Learning Theory*. Springer. 2007, pp. 82–96.

[31] Vijay Kamble et al. "Truth Serums for Massively Crowdsourced Evaluation Tasks". In: *arXiv preprint arXiv:1507.07045* (2015).

[32] David R Karger, Sewoong Oh, and Devavrat Shah. "Budget-optimal task allocation for reliable crowdsourcing systems". In: *Operations Research* 62.1 (2014), pp. 1–24.

[33] Roni Khardon and Gabriel Wachman. "Noise tolerant variants of the perceptron algorithm". In: *Journal of Machine Learning Research* 8.Feb (2007), pp. 227–248.

[34] Y. Kong and G. Schoenebeck. "A Framework For Designing Information Elicitation Mechanisms That Reward Truth-telling". In: *ArXiv e-prints* (May 2016). arXiv: `1605.01021 [cs.GT]`.

[35] Y. Kong and G. Schoenebeck. "Equilibrium Selection in Information Elicitation without Verification via Information Monotonicity". In: *ArXiv e-prints* (Mar. 2016). arXiv: 1603.07751 [cs.GT].

[36] Y. Kong, G. Schoenebeck, and K. Ligett. "Putting Peer Prediction Under the Micro(economic)scope and Making Truth-telling Focal". In: *ArXiv e-prints* (Mar. 2016). arXiv: 1603.07319 [cs.GT].

[37] Yuqing Kong and Grant Schoenebeck. "Eliciting expertise without verification". Unpublished Manuscript. 2017.

[38] Yuqing Kong and Grant Schoenebeck. "Equilibrium selection in information elicitation without verification via information monotonicity". In: *arXiv preprint arXiv:1603.07751* (2016).

[39] Yingming Li, Ming Yang, and Zhongfei Zhang. "Multi-view representation learning: A survey from shallow methods to deep methods". In: *arXiv preprint arXiv:1610.01206* (2016).

[40] Friedrich Liese and Igor Vajda. "On divergences and informations in statistics and information theory". In: *IEEE Transactions on Information Theory* 52.10 (2006), pp. 4394–4412.

[41] Yang Liu and Yiling Chen. "Machine-Learning Aided Peer Prediction". In: *Proceedings of the 2017 ACM Conference on Economics and Computation*. EC '17. Cambridge, Massachusetts, USA: ACM, 2017, pp. 63–80. ISBN: 978-1-4503-4527-9. DOI: 10.1145/3033274.3085126. URL: http://doi.acm.org/10.1145/3033274.3085126.

[42] Debmalya Mandal et al. "Peer Prediction with Heterogeneous Tasks". In: *arXiv preprint arXiv:1612.00928* (2016).

[43] David McAllester. "Information Theoretic Co-Training". In: *CoRR* abs/1802.07572 (2018). arXiv: 1802.07572. URL: http://arxiv.org/abs/1802.07572.

[44] John McCoy and Drazen Prelec. "A statistical model for aggregating judgments by incorporating peer predictions". In: *arXiv preprint arXiv:1703.04778* (2017).

[45] N. Miller, P. Resnick, and R. Zeckhauser. "Eliciting informative feedback: The peer-prediction method". In: *Management Science* (2005), pp. 1359–1373.

[46] Nagarajan Natarajan et al. "Learning with noisy labels". In: *Advances in neural information processing systems*. 2013, pp. 1196–1204.

[47] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.

[48] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. "Estimating divergence functionals and the likelihood ratio by convex risk minimization". In: *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5847–5861.

[49] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. "On surrogate loss functions and f-divergences". In: *The Annals of Statistics* (2009), pp. 876–904.

[50] Hanchuan Peng, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on pattern analysis and machine intelligence* 27.8 (2005), pp. 1226–1238.

[51] D. Prelec. "A Bayesian Truth Serum for subjective data". In: *Science* 306.5695 (2004), pp. 462–466.

[52] Dražen Prelec, H Sebastian Seung, and John McCoy. "A solution to the single-question crowd wisdom problem". In: *Nature* 541.7638 (2017), pp. 532–535.

[53] Goran Radanovic and Boi Faltings. "A robust bayesian truth serum for non-binary signals". In: *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013*. EPFL-CONF-197486. 2013, pp. 833–839.

[54] Goran Radanovic and Boi Faltings. "Incentive schemes for participatory sensing". In: *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2015, pp. 1081–1089.

[55] Goran Radanovic and Boi Faltings. "Incentives for truthful information elicitation of continuous signals". In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014.

[56] Alexander J Ratner et al. "Data programming: Creating large training sets, quickly". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3567–3575.

[57] Vikas C Raykar et al. "Learning from crowds". In: *Journal of Machine Learning Research* 11.Apr (2010), pp. 1297–1322.

[58] Blake Riley. "Mechanisms for Making Accurate Decisions in Biased Crowds". In: (2015).

[59] Blake Riley. "Minimum truth serums with optional predictions". In: *Proceedings of the 4th Workshop on Social Computing and User Generated Content (SC14)*. 2014.

[60] R Tyrrell Rockafellar et al. "Extension of Fenchel'duality theorem for convex functions". In: *Duke mathematical journal* 33.1 (1966), pp. 81–89.

[61] William E Roth. "On direct product matrices". In: *Bulletin of the American Mathematical Society* 40.6 (1934), pp. 461–468.

[62] Clayton Scott, Gilles Blanchard, and Gregory Handy. "Classification with asymmetric label noise: Consistency and maximal denoising". In: *Conference On Learning Theory*. 2013, pp. 489–511.

[63] Victor Shnayder et al. "Informed Truthfulness in Multi-Task Peer Prediction". In: *Proceedings of the 2016 ACM Conference on Economics and Computation*. EC '16. Maastricht, The Netherlands: ACM, 2016, pp. 179–196. ISBN: 978-1-4503-3936-0. DOI: 10.1145/2940716.2940790. URL: http://doi.acm.org/10.1145/2940716.2940790.

[64] Sainbayar Sukhbaatar and Rob Fergus. "Learning from noisy labels with deep neural networks". In: *arXiv preprint arXiv:1406.2080* 2.3 (2014), p. 4.

[65] Peter Welinder et al. "The Multidimensional Wisdom of Crowds." In: *NIPS*. Vol. 23. 2010, pp. 2424–2432.

[66] Robert L Winkler. "Scoring rules and the evaluation of probability assessors". In: *Journal of the American Statistical Association* 64.327 (1969), pp. 1073–1078.

[67] J. Witkowski and D. Parkes. "A robust Bayesian Truth Serum for small populations". In: *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*. 2012.

[68] Jens Witkowski. "Robust Peer Prediction Mechanisms". PhD thesis. Department of Computer Science, Albert-Ludwigs-Universität Freiburg, May 2014.

[69] Jens Witkowski, Bernhard Nebel, and David C Parkes. "Robust Peer Prediction Mechanisms". PhD thesis. Ph. D. Dissertation, Albert-Ludwigs-Universitat Freiburg: Institut fur Informatik, 2014.

[70] Jens Witkowski and David C Parkes. "Learning the prior in minimal peer prediction". In: *Proceedings of the 3rd Workshop on Social Computing and User Generated Content at the ACM Conference on Electronic Commerce*. Citeseer. 2013, p. 14.

[71] Jens Witkowski and David C Parkes. "Peer prediction without a common prior". In: *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM. 2012, pp. 964–981.

[72] Jens Witkowski et al. "Proper Proxy Scoring Rules." In: *AAAI*. 2017, pp. 743–749.

[73] Chang Xu, Dacheng Tao, and Chao Xu. "A survey on multi-view learning". In: *arXiv preprint arXiv:1304.5634* (2013).

[74] Peter Zhang and Yiling Chen. "Elicitability and knowledge-free elicitation with peer prediction". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2014, pp. 245–252.

[75] Yuchen Zhang et al. "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing". In: *Advances in neural information processing systems*. 2014, pp. 1260–1268.

[76] Denny Zhou et al. "Learning from the wisdom of crowds by minimax entropy". In: *Advances in neural information processing systems*. 2012, pp. 2195–2203.