Mechanism Design for Collaborative Normal Mean Estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study collaborative normal mean estimation, where m strategic agents collect i.i.d samples from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ at a cost. They all wish to estimate the mean μ. By sharing data with each other, agents can obtain better estimates while keeping the cost of data collection small. To facilitate this collaboration, we wish to design mechanisms that encourage agents to collect a sufficient amount of data and share it truthfully, so that they are all better off than working alone. In naive mechanisms, such as simply pooling and sharing all the data, an individual agent might find it beneficial to under-collect and/or fabricate data, which can lead to poor social outcomes. We design a novel mechanism that overcomes these challenges via two key techniques: first, when sharing the others' data with an agent, the mechanism corrupts this dataset proportional to how much the data reported by the agent differs from the others; second, we design minimax optimal estimators for the corrupted dataset. Our mechanism, which is incentive compatible and individually rational, achieves a social penalty (sum of all agents' estimation errors and data collection costs) that is at most a factor 2 of the global minimum. When applied to high dimensional (non-Gaussian) distributions with bounded variance, this mechanism retains these three properties, but with slightly weaker results. Finally, in two special cases where we restrict the strategy space of the agents, we design mechanisms that essentially achieve the global minimum.

1 Introduction

With the rise in popularity of machine learning, data is becoming an increasingly valuable resource for businesses, scientific organizations, and government institutions. However, data collection is often costly. For instance, to collect data, businesses may need to carry out market research, scientists may need to conduct experiments, and government institutions may need to perform surveys on public services. However, once data has been generated, it can be freely replicated and used by many organizations [19]. Hence, instead of simply collecting and learning from their own data, by sharing data with each other, organizations can mutually reduce their own data collection costs and improve the utility they derive from data [20]. In fact, there are already several platforms to facilitate data sharing among businesses [1, 38], scientific organizations [2, 3], and public institutions [15, 32].

However, simply pooling everyone's data and sharing with each other can lead to free-riding [22, 33]. If an agent (e.g an organization) sees that other agents are already contributing a large amount of data, then, the cost she incurs to collect her own dataset may not offset the marginal improvement in her *own* learned model due to diminishing returns of increasing dataset sizes (we describe this rigorously in §2). Hence, while she benefits from others' data, she has no incentive to collect and contribute data to the pool. A seemingly simple fix to this free-riding problem is to only return the datasets of the others if an agent submits a large enough dataset herself. However, this can be easily manipulated by a strategic agent who submits a large fabricated (fake) dataset without incurring any cost, receives the others' data, and then discards her fabricated dataset when learning. While the agent has benefited by this bad behavior, other agents who may use this fabricated dataset are worse off.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

In this work, we study these challenges in data sharing in one of the most foundational statistical problems, normal mean estimation, where the goal is to estimate the mean μ of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with known variance σ^2. We wish to design *mechanisms* for data sharing that satisfy the three fundamental desiderata of mechanism design; *incentive compatibility (IC):* agents have incentive to collect a sufficiently large amount of data and share it truthfully; *individual rationality (IR):* agents are better off participating in the mechanism than working on their own; and *efficiency:* the mechanism leads to outcomes with small estimation error and data collection costs for all agents.

Contributions: *(i)* In §2, we formalize collaborative normal mean estimation in the presence of strategic agents. *(ii)* In §3, we design an IC and IR mechanism for this problem to prevent free-riding and data fabrication and show that its social penalty, i.e sum of all agents' estimation errors and data collection costs, is at most twice that of the global minimum. *(iii)* In Appendix E, we study the same mechanism in high dimensional settings and relax the Gaussian assumption to distributions with bounded variance. We show that the mechanism retains its properties, with only a slight weakening of the IC and efficiency guarantees. *(iv)* In §4, we consider two special cases where we impose natural restrictions on the agents' strategy space. We show that it is possible to design mechanisms which essentially achieve the global minimum social penalty in both settings. Next, we will summarize our primary mechanism and the associated theorem in §3.

1.1 Summary of main results

Formalism: We assume that all agents have a fixed cost for collecting one sample, and define an agent's penalty (negative utility) as the sum of her estimation error and the cost she incurred to collect data. To make the problem well-defined, for the estimation error, we find it necessary to consider the *maximum risk*, i.e maximum expected error over all $\mu \in \mathbb{R}$. A mechanism asks agents to collect data, and then shares the data among the agents in an appropriate manner to achieve the three desiderata. An agent's strategy space consists of three components: how much data she wishes to collect, what she chooses to submit after collecting the data, and how she estimates the mean μ using the dataset she collected, the dataset she submitted, and the information she received from the mechanism.

Mechanism and theoretical result: In our mechanism, which we call C3D (Cross-Check and Corrupt based on Difference), each agent i collects a dataset X_i and submits a possibly fabricated or altered version Y_i to the mechanism. The mechanism then determines agent i's allocation in the following manner. It pools the data from the other agents and splits them into two subsets Z_i, Z'_i. Then, Z_i is returned as is, while Z'_i is corrupted by adding noise that is proportional to the difference between Y_i and Z_i. The mechanism also provides some side information to indicate the level of corruption. We show that this mechanism has a Nash equilibrium where all agents collect a sufficiently large amount of data, submit it truthfully, and use a carefully weighted average of the three datasets X_i, Z_i, and Z'_i as their estimate for μ. The weighting uses the side information. Below, we state an informal version of the main theoretical result of this paper, which summarizes the properties of our mechanism.

Theorem 1 (informal): *The above mechanism is incentive compatible, individually rational, and achieves a social penalty that is at most twice the globally minimum social penalty.*

Corruption is the first of two ingredients to achieving IC. If an agent collects less or fabricates, she risks looking different to the others, and will receive a dataset Z'_i of poorer quality. The second ingredient is the design of the weighted average estimator which is (minimax) optimal after corruption. Had the mechanism assumed that an agent will use any other sub-optimal estimator (e.g a simple average), then it will need to lower the amount of corruption to ensure IR and efficiency. However, a strategic agent who realizes that a better estimator (e.g weighted average), will reduce their estimation error, can leverage this insight to collect less data and decrease their overall penalty.

Proof techniques: The most challenging part of our analysis is to show IC, First, to show minimax optimality of our estimator, we construct a sequence of normal priors for μ and show that the minimum Bayes' risk converges to the maximum risk of the weighted average estimator. However, when compared to typical minimax proofs, we face more significant challenges. The first of these is that the combined dataset $X_i \cup Z_i \cup Z'_i$ is neither independent nor identically distributed as the corruption is data-dependent. The second is that the agent's submission Y_i also determines the degree of corruption, so we cannot look at the estimator in isolation when computing the minimum Bayes' risk; we should also consider the space of functions an agent may use to determine Y_i from X_i. The third is that the expressions for the minimum Bayes' risk do not have closed form solutions and require non-trivial algebraic manipulations. To complete the IC proof, we show that due to the

2

95 carefully chosen amount of corruption, the agent should collect a sufficient amount of data to avoid
96 excessive corruption, but not too much so as to increase her data collection costs.

1.2 Related Work

98 Mechanism design is one of the core areas of research in game theory [12, 17, 34]. Our work here is
99 more related to mechanism design without payments, which has seen applications in fair division [29],
100 matching markets [30], and kidney exchange [31] to name a few. There is a long history of work in
101 the intersection of machine learning and mechanism design, although the overwhelming majority
102 apply learning techniques when there is incomplete information about the mechanism or agent
103 preferences, (e.g [6, 8, 21, 27, 28]). On the flip side, some work have designed data marketplaces,
104 where customers may purchase data from contributors [4, 5, 18, 36]. These differ from our focus
105 where we wish to incentivize agents to collaborate without payments.

106 Due to the popularity of shared data platforms [1, 2, 15, 32] and federated learning [20], there has
107 been a recent interest in designing mechanisms for data sharing. Sim et al. [33] and Xu et al. [37]
108 study fairness in collaborative data sharing, where the goal is to reward agents according to the
109 amount of data they contribute. However, their mechanisms do not apply when strategic agents may
110 try to manipulate a mechanism. Blum et al. [9] and Karimireddy et al. [22] study collaboration in
111 federated learning. However, the strategy space of an agent is restricted to how much data they collect
112 and their mechanism rewards each agent according to the quantity of the data she submitted. The
113 above four works recognize that free-riding can be detrimental to data sharing, but assume that agents
114 will not fabricate data. As discussed above, if this assumption is not true, agents can easily manipulate
115 such mechanisms. Fraboni et al. [16] and Lin et al. [24] study federated learning settings where
116 free-riders may send in fabricated gradients without incurring the computational cost of computing
117 the gradients. However, their focus is on designing gradient descent algorithms that are robust to
118 such attacks and not on incentivizing agents to perform the gradient computations. Some work have
119 designed mechanisms for federated learning so as to elicit private information (such as data collection
120 costs), but their focus is not on preventing free-riding or fabrication [14, 25].

121 Our approach of using corruption to engender good behaviour draws inspiration from the robust
122 estimation literature, which design estimators that are robust to data from malicious agents [11, 13, 26].
123 However, to the best of our knowledge, the specific form of corruption and the subsequent design of
124 the minimax optimal estimator are new in this work, and require novel analysis techniques.

2 Problem Setup

126 We will now formally define our problem. We have m agents, who are each able to collect i.i.d
127 samples from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is known. They wish to estimate the mean μ
128 of this distribution. To collect one sample, the agent has to incur a cost c. We will assume that σ^2, c,
129 and m are public information. However, $\mu \in \mathbb{R}$ is unknown, and no agent has auxiliary information,
130 such as a prior, about μ. An agent wishes to minimize her estimation error, while simultaneously
131 keeping the cost of data collection low. While an agent may collect data on her own to manage this
132 trade-off, by sharing data with other agents, she can reduce costs while simultaneously improving her
133 estimate. We wish to design mechanisms to facilitate such sharing of data.

134 **Mechanism:** A mechanism receives a dataset from each agent, and in turn returns an *allocation* A_i
135 to each agent. An agent will use her allocation to estimate μ. This allocation could be, for instance, a
136 larger dataset obtained with other agents' datasets. The mechanism designer is free to choose a space
137 of allocations $\mathcal{A}$ to achieve the desired goals. Formally, we define a mechanism as a tuple $M = (\mathcal{A}, b)$
138 where $\mathcal{A}$ denotes the space of allocations, and b is a procedure to map the datasets collected from the
139 m agents to m allocations. Denoting the universal set by $\mathcal{U}$, we write the space of mechanisms $\mathcal{M}$ as

$$\mathcal{M} = \big\{ M = (\mathcal{A}, b) : \ \mathcal{A} \subset \mathcal{U}, \ \ b : (\bigcup_{n \geq 0} \mathbb{R}^n)^m \to \mathcal{A}^m \big\}. \tag{1}$$

140 As is customary, we will assume that the mechanism designer will publish the space of allocations $\mathcal{A}$
141 and the mapping b (the procedure used to obtain the allocations) ahead of time, so that agents can
142 determine their strategies. However, specific values computed/realized during the execution of the
143 mechanism are not revealed, unless the mechanism chooses to do so via the allocation A_i.

144 **Agents' strategy space:** Once the mechanism is published, the agent will choose a strategy. In our
145 setting, this will be the tuple (n_i, f_i, h_i), which determines how much data she wishes to collect, what
146 she chooses to submit, and how she wishes to estimate the mean μ. First, the agent samples n_i points

3

to collect her initial dataset $X_i = \{x_{i,j}\}_{j=1}^{n_i}$, where $x_{i,j} \sim \mathcal{N}(\mu, \sigma^2)$, incurring cn_i cost. She then submits $Y_i = \{y_{i,j}\}_j = f_i(X_i)$ to the mechanism. Here f_i is a function which maps the collected dataset to a possibly fabricated or falsified dataset of a potentially different size. In particular, this fabrication can depend on the data she has collected. Finally, the mechanism returns the agent's allocation A_i, and the agent computes an estimate $h_i(X_i, Y_i, A_i)$ for μ using her initial dataset X_i, the dataset she submitted Y_i, and the allocation she received A_i. We include Y_i as part of the estimate since an agent's submission may affect the allocation she receives. Consequently, agents could try to elicit additional information about μ via a carefully chosen Y_i. We can write the strategy space of an agent as $\mathcal{S} = \mathbb{N} \times \mathcal{F} \times \mathcal{H}$, where $\mathcal{F}$ is the space of functions mapping the dataset collected to the dataset submitted, and $\mathcal{H}$ is the space of all estimators using all the information she has. We have:

$$\mathcal{F} = \left\{ f : \bigcup_{n \geq 0} \mathbb{R}^n \to \bigcup_{n \geq 0} \mathbb{R}^n \right\}, \qquad \mathcal{H} = \left\{ h : \bigcup_{n \geq 0} \mathbb{R}^n \times \bigcup_{n \geq 0} \mathbb{R}^n \times \mathcal{A} \to \mathbb{R} \right\}. \quad (2)$$

One element of interest in $\mathcal{F}$ is the identity $\mathbf{I}$ which maps a dataset to itself. A mechanism designer would like an agent to use $f_i = \mathbf{I}$, i.e to submit only the data that she collected, so that other agents can benefit from her data. Going forward, when $s = \{s_i\}_i \in \mathcal{S}^m$ denotes the strategies of all agents, we will use $s_{-i} = \{s_j\}_{j \neq i}$ to denote the strategies of all agents except i.

Agent penalty: The agent's *penalty* p_i (i.e negative utility) is the sum of her squared estimation error and the cost cn_i incurred to collect her dataset X_i of n_i points. The agent's penalty depends on the mechanism M and the strategies $s = \{s_j\}_j$ of all the agents. Making this explicit, p_i is defined as:

$$p_i(M, s) = \sup_{\mu \in \mathbb{R}} \mathbb{E}\left[(h_i(X_i, Y_i, A_i) - \mu)^2 \,\Big|\, \mu \right] + cn_i \quad (3)$$

The term inside the expectation is the squared difference between the agent's estimate and the true mean (conditioned on the true mean μ). The expectation is with respect to the randomness of all agents' data and possibly any randomness in the strategies employed by the agents and/or the mechanism. We consider the *maximum risk*, i.e supremum over $\mu \in \mathbb{R}$, since the true mean μ is unknown to the agent a priori, and their strategy should yield good estimates, and hence small penalty, over all possible values μ. To illustrate this further, note that when the value of true mean μ is μ', the optimal strategy for an agent will always be to not collect any data and choose the estimator $h_i(\cdot, \cdot, \cdot) = \mu'$ leading to 0 penalty. However, this strategy can be meaningfully realized by an agent only if she knew that $\mu = \mu'$ a priori which renders the problem meaningless[1]. Considering the maximum risk accounts for the fact that μ is unknown and makes the problem well-defined.

Recommended strategies: In addition to publishing the mechanism, the mechanism designer will recommend strategies $s^\star = \{s_i^\star\}_i \in \mathcal{S}^m$ for the agents so as to incentivize collaboration and induce optimal social outcomes.

Desiderata: We can now define the three desiderata for a mechanism:

1. *Incentive compatibility (IC):* A mechanism $M = (\mathcal{A}, b)$ is said to be IC at the recommended strategy profile $s^\star$ if, for each agent i, and for every other alternative strategy $s_i \in \mathcal{S}$ for that agent, we have $p_i(M, s^\star) \leq p_i(M, (s_i, s_{-i}^\star))$. That is, $s^\star$ is a Nash equilibrium so no agent has incentive to deviate if all other agents are following $s^\star$.

2. *Individual rationality (IR):* We say that a mechanism M is IR at $s^\star$ if no agent suffers from a higher penalty by participating in the mechanism than the lowest possible penalty she could achieve on her own. If an agent does not participate, she does not submit nor receive any data from the mechanism; she will simply choose how much data to collect and design the best possible estimator. Formally, we say that a mechanism M is IR if the following is true for each agent i:

$$p_i(M, s^\star) \leq \inf_{n_i \in \mathbb{N}, \, h_i \in \mathcal{H}} \left\{ \sup_{\mu \in \mathbb{R}} \mathbb{E}\left[(h_i(X_i, \varnothing, \varnothing) - \mu)^2 \,|\, \mu \right] + cn_i \right\}. \quad (4)$$

3. *Efficiency:* The *social penalty* $P(M, s)$ of a mechanism M when agents follow strategies s, is the sum of agent penalties (defined below). We define $\mathrm{PR}(M, s^\star)$ to be the ratio between the social penalty of a mechanism at the recommended strategies $s^\star$, and the lowest possible social penalty among all possible mechanisms and strategies (*without* IC or IR constraints). We have:

---

[1]This is akin to the reason why it is customary to study the maximum risk in frequentist statistics [23, 35]. An alternative approach is to take a Bayesian view, considering a prior on μ and using the Bayes' risk $\mathbb{E}_\mu[\mathbb{E}[(h_i(X_i, Y_i, A_i) - \mu)^2 | \mu]]$ instead of the maximum risk in p_i. While we have adopted a frequentist formalism here, our main proof ideas can be ported over to the Bayesian setting as well.

4

$$P(M, s) = \sum_{i \in [m]} p_i(M, s), \qquad \mathrm{PR}(M, s^\star) = \frac{P(M, s^\star)}{\inf\limits_{M' \in \mathcal{M}, \, s \in \mathcal{S}^m} P(M', s)} \tag{5}$$

Note that $\mathrm{PR} \geq 1$. We say that a mechanism is efficient if $\mathrm{PR}(M, s^\star) = 1$ and that it is approximately efficient if $\mathrm{PR}(M, s^\star)$ is bounded by some constant that does not depend on m. If $s_i^\star$ is a Nash equilibrium, then $\mathrm{PR}(M, s^\star)$ is an upper bound on the price of stability [7].

For what follows, we will discuss optimal strategies for agents working on her own and present a simple mechanism which minimizes the social penalty, but has a poor Nash equilibrium.

Optimal strategies for an agent working on her own: Recall that, given n samples $\{x_i\}_{i=1}^n$ from $\mathcal{N}(\mu, \sigma^2)$, the sample mean is a minimax optimal estimator [23]; i.e among all possible estimators h, the sample mean minimizes the maximum risk $\sup_{\mu \in \mathbb{R}} \mathbb{E}[(\mu - h(\{x_i\}_{i=1}^n, \varnothing, \varnothing))^2 \,|\, \mu]$ (note that the agent only has the dataset she collected). Moreover, its mean squared error is σ^2/n for all $\mu \in \mathbb{R}$. Hence, an agent acting on her own will choose the sample mean and collect $n_i = \sigma/\sqrt{c}$ samples so as to minimize their penalty; as long as the amount of data is less than $\sigma/\sqrt{c}$, an agent has incentive to collect more data since the cost of collecting one more point is offset by the marginal decrease in estimation error. This can be seen via the following simple calculation:

$$\inf_{\substack{n_i \in \mathbb{R} \\ h_i \in \mathcal{H}}} \left(\sup_\mu \mathbb{E}\left[(h_i(X_i, \varnothing, \varnothing) - \mu)^2 \,\Big|\, \mu \right] + cn_i \right) = \min_{n_i \in \mathbb{R}} \left(\frac{\sigma^2}{n_i} + cn_i \right) = 2\sigma\sqrt{c} \overset{\Delta}{=} p_{\min}^{\mathrm{IR}}. \tag{6}$$

Let $p_{\min}^{\mathrm{IR}} = 2\sigma\sqrt{c}$ denote the lowest achievable penalty by an agent working on her own. If all m agents work independently, then the total social penalty is $mp_{\min}^{\mathrm{IR}} = 2\sigma m\sqrt{c}$. Next, we will look at a simple mechanism and an associated set of strategies which achieve the global minimum penalty. This will show that it is possible for all agents to achieve a significantly lower penalty via collaboration.

A globally optimal mechanism *without* strategic considerations: The following simple mechanism $M_{\texttt{pool}}$, pools all the data from the other agents and gives it back to an agent. Precisely, it chooses the space of allocation $\mathcal{A} = \bigcup_{n \geq 0} \mathbb{R}^n$ to be datasets of arbitrary length, and sets agent i's allocation to be $A_i = \bigcup_{j \neq i} Y_i$. The recommended strategies $s^{\texttt{pool}} = \{(n_i^{\texttt{pool}}, f_i^{\texttt{pool}}, h_i^{\texttt{pool}})\}_i$ asks each agent to collect $n_i^{\texttt{pool}} = \sigma/\sqrt{cm}$ points[2], submit it as is $f_i^{\texttt{pool}} = \mathbf{I}$, and use the sample mean of all points as her estimate $h_i^{\texttt{pool}}(X_i, X_i, A_i) = \frac{1}{|X_i \cup A_i|} \sum_{z \in X_i \cup A_i} z$. It is straightforward to show that this minimizes the social penalty if all agents follow $s^{\texttt{pool}}$. After each agent has collected their datasets $\{X_i\}_i$, the social penalty is minimized if all agents have access to each other's datasets and they all use a minimax optimal estimator: this justifies using $M_{\texttt{pool}}$ with $f_i^{\texttt{pool}} = \mathbf{I}$ and setting $h_i^{\texttt{pool}}$ to be the sample mean. The following simple calculation justifies the choice of $\sum_i n_i^{\texttt{pool}}$:

$$\inf_{s \in \mathcal{S}^m} \sum_{i=1}^m \left(\sup_\mu \mathbb{E}\left[(h_i(X_i, f_i, A_i) - \mu)^2 \,\Big|\, \mu \right] + cn_i \right) = \min_{\{n_i\}_i} \left(\frac{m\sigma^2}{\sum_i n_i} + c \sum_i n_i \right) = 2\sigma\sqrt{mc}.$$

However, $s^{\texttt{pool}}$ is not a Nash equilibrium of this mechanism, as an agent will find it beneficial to free-ride. If all other agents are submitting $\sigma/\sqrt{cm}$ points, by collecting no points, an agent's penalty is $\sigma\sqrt{mc}/(m-1)$, as she does not incur any data collection cost. This is strictly smaller than $2\sigma\sqrt{c/m}$ when $m \geq 3$. In fact, it is not hard to show that $M_{\texttt{pool}}$ is at a Nash equilibrium only when the total amount of data is $\sigma/\sqrt{c}$; for additional points, the marginal reduction in the estimation error for an individual agent does not offset her data collection costs. The social penalty at these equilibria is $\sigma\sqrt{c}(m+1)$ which is significantly larger than the global minimum when there are many agents.

A seemingly simple way to fix this mechanism is to only return the datasets of the other agents if an agent submits at least $\sigma/\sqrt{cm}$ points. However, as we will see in §4.1, such a mechanism can also be manipulated by an agent who submits a fabricated dataset of $\sigma/\sqrt{cm}$ points without actually collecting any data and incurring any cost and then discarding the fabricated dataset when estimating. Next, we will present our mechanism for this problem which satisfies all three desiderata.

3   Method and Results

We have outlined our mechanism $M_{\texttt{C3D}}$, and its interaction with the agents in Algorithm 1 in the natural order of events. We will first describe it procedurally, and then motivate our design choices.

---

[2]To avoid rounding effects, henceforth we will treat $\sigma/\sqrt{cm}$, and $\sigma/\sqrt{c}$ as integers.

5

Algorithm 1 M_{C3D}

1: **Mechanism designer publishes:**
2:   The allocation space $\mathcal{A} = \bigcup_{n\geq0} \mathbb{R}^n \times \bigcup_{n\geq0} \mathbb{R}^n \times \mathbb{R}_+$, and the procedure in lines 6–15.
3: **Each agent i:**
4:   Choose strategy $s_i = (n_i, f_i, h_i)$.   # See (8) for recommended strategies.
5:   Sample n_i points $X_i = \{x_{i,j}\}_{j=1}^{n_i}$ and submit $Y_i = f_i(X_i)$ to the mechanism.
6: **Mechanism:**
7:   **For** each agent $i \in [m]$:   # can be done simultaneously for all agents
8:     $Y_{-i} \leftarrow \bigcup_{j\neq i} Y_j$.
9:     **If** $m \leq 4$:   # Simply pool and return all of the other agents' data to agent i.
10:       $A_i \leftarrow (Y_{-i}, \varnothing, 0)$. Return A_i to agent i.
11:     **Else:**
12:       $Z_i \leftarrow$ sample $\min\{|Y_{-i}|, \sigma/\sqrt{cm}\}$ points in Y_{-i} without replacement.
13:       $\eta_i^2 \leftarrow \alpha^2 \left(\frac{1}{|Y_i|} \sum_{y\in Y_i} y - \frac{1}{|Z_i|} \sum_{z\in Z_i} z \right)^2$ # See (7) for α.
14:       $Z_i' \leftarrow \{z + \epsilon_{z,i}, \text{ for all } z \in Y_{-i}\backslash Z_i \text{ where } \epsilon_{z,i} \sim \mathcal{N}(0, \eta_i^2)\}$
15:       $A_i \leftarrow (Z_i, Z_i', \eta_i^2)$. Return A_i to agent i.
16: **Each agent i:**
17:   Compute estimate $h_i(X_i, Y_i, A_i)$.   # See (8) for recommended estimator.

Our mechanism uses the following allocation space, $\mathcal{A} = \bigcup_{n\geq0} \mathbb{R}^n \times \bigcup_{n\geq0} \mathbb{R}^n \times \mathbb{R}_+$. An allocation $A_i = (Z_i, Z_i', \eta_i^2) \in \mathcal{A}$ consists of an uncorrupted dataset Z_i, a corrupted dataset Z_i', and the variance η_i^2 of the noise added to Z_i' for corruption. Once the mechanism and the allocation space are published, agent i chooses her strategy $s = (n_i, f_i, h_i)$. She collects a dataset $X_i = \{x_{i,j}\}_{j=1}^{n_i}$, where $x_{i,j} \sim \mathcal{N}(\mu, \sigma^2)$, and submits $Y_i = f_i(X_i)$ to the mechanism.

Our mechanism determines agent i's allocation as follows. Let Y_{-i} be the union of all datasets submitted by the other agents. If there are at most four agents, we simply return all of the other agents' data without corruption by setting $A_i \leftarrow (Y_{-i}, \varnothing, 0)$. If there are more agents, the mechanism first chooses a random subset of size $\min\{|Y_{-i}|, \sigma/\sqrt{cm}\}$ from Y_{-i}; denote this Z_i. In line 13, the mechanism individually adds Gaussian noise to the remaining points $Y_{-i}\backslash Z_i$ to obtain Z_i' (line 14). The variance η_i^2 of the noise depends on the difference between the sample means of the subset Z_i and the agent's submission Y_i. It is modulated by a value α, which is a function of c, m, and σ^2. Precisely, α is the smallest number larger than $\sqrt{\sigma}(cm)^{-1/4}$ which satisfies $G(\alpha) = 0$, where:

$$G(\alpha) := \left(\frac{m-4}{m-2} \frac{4\alpha^2}{\sigma/\sqrt{cm}} - 1 \right) \frac{4\alpha}{\sqrt{\sigma}(m/c)^{1/4}} - \left(4(m+1)\frac{\alpha^2}{\sigma\sqrt{m/c}} - 1 \right)\sqrt{2\pi}\exp\left(\frac{\sigma\sqrt{m/c}}{8\alpha^2}\right)\text{Erfc}\left(\frac{\sqrt{\sigma}(m/c)^{1/4}}{2\sqrt{2}\alpha}\right) \quad (7)$$

Finally, the mechanism returns the allocation $A_i = (Z_i, Z_i', \eta_i^2)$ to agent i and the agent estimates μ.

Recommended strategies: The recommended strategy $s_i^\star = (n_i^\star, f_i^\star, h_i^\star)$ for agent i is given in (8). The agent should collect $n_i^\star = \sigma/(m\sqrt{c})$ samples if there are at most four agents, and $n_i^\star = \sigma/\sqrt{cm}$ samples otherwise. She should submit it without fabrication or alteration $f_i = \mathbf{I}$, and then use a weighted average of the datasets (X_i, Z_i, Z_i') to estimate μ. The weighting is proportional to the inverse variance of the data. For X_i and Z_i this is simply σ^2, but for Z_i', the variance is $\sigma^2 + \eta_i^2$ since the mechanism adds Gaussian noise with variance η_i^2. We have:

$$n_i^\star = \begin{cases} \frac{\sigma}{m\sqrt{c}} & \text{if } m \leq 4 \\ \frac{\sigma}{\sqrt{cm}} & \text{if } m > 4 \end{cases}, \quad f_i^\star = \mathbf{I}, \quad h_i^\star(X_i, Y_i, (Z_i, Z_i', \eta_i^2)) = \frac{\frac{1}{\sigma^2}\sum_{u\in X_i\cup Z_i} u + \frac{1}{\sigma^2+\eta_i^2}\sum_{u\in Z_i'} u}{\frac{1}{\sigma^2}|X_i \cup Z_i'| + \frac{1}{\sigma^2+\eta_i^2}|Z_i'|}$$
$$(8)$$

Design choices: Next, we will describe our design choices and highlight some key challenges. When $m \leq 4$, it is straightforward to show that the mechanism satisfies all our desired properties (see beginning of §3.1), so we will focus on the case $m > 4$. First, recall that the mechanism needs to incentivize agents to collect a sufficient amount of samples. However, simply counting the number of samples can be easily manipulated by an agent who simply submits a fabricated dataset of a large number of points. Instead, Algorithm 1 attempts to infer the quality of the data submitted by the agents using how well an agent's submission Y_i approximates μ. Ideally, we would set the variance η_i^2 of this corruption to be proportional to the difference $(\frac{1}{|Y_i|}\sum_{y\in Y_i} y - \mu)^2$, so that the

6

more data she submits, the less the variance of Z_i', which in turn yields a more accurate estimate for μ. However, since μ is unknown, we use a subset Z_i obtained from other agents' data as a proxy for μ, and set η_i^2 proportional to $\left(\frac{1}{|Y_i|} \sum_{y \in Y_i} y - \frac{1}{|Z_i|} \sum_{z \in Z_i} z \right)^2$. If all agents are following $s^\star$, then $|Y_i| = |Z_i| = \sigma/\sqrt{cm} = n_i^\star$; it is sufficient to use only $n_i^\star$ points for validating Y_i since both $\frac{1}{|Y_i|} \sum_{y \in Y_i} y$ and $\frac{1}{|Z_i|} \sum_{z \in Z_i} z$ will have the same order of error in approximating μ.

The second main challenge is the design of the recommended estimator $h_i^\star$. In §3.1 we show how splitting Y_{-i} into a clean and corrupted parts Z_i, Z_i' allows us to design a minimax optimal estimator. A minimax optimal estimator is crucial to achieving IC. To explain this, say that the mechanism assumes that agents will use a sub-optimal estimator, e.g sample mean of $X_i \cup Z_i \cup Z_i'$. Then, to account for the larger estimation error, it will need to choose a lower level of corruption η_i^2 to minimize the social penalty. However, a smart agent who uses the weighted average (or any other better estimator) will realize that she can achieve a lower maximum risk with a better estimator instead of more data. She can leverage this insight to collect less data and lower her overall penalty.

This concludes the description of our mechanism. The following theorem, which is the main theoretical result of this paper, states that M_{C3D} achieves the three desiderata outlined in §2.

Theorem 1. *Let $m > 1$, α be as defined in (7), and $s_i^\star$ be as defined in (8). Then, the following statements are true about the mechanism M_{C3D} in Algorithm 1. (i) The strategy profile $s^\star$ is a Nash equilibrium. (ii) The mechanism is individually rational at $s^\star$. (iii) The mechanism is approximately efficient, with $\mathrm{PR}(M_{\text{C3D}}, s^\star) \leq 2$.*

The mechanism is IC as, provided that others are following $s_i^\star$, there is no reason for any one agent to deviate. Moreover, we achieve low social penalty at $s_i^\star$. Before we proceed, the expression for α in (7) warrants explanation. If we treat α is a variable, we find that different choices of α can lead to other Nash equilibria with corresponding bounds on PR. This specific choice of α leads to a Nash equilibrium where agents collect $\sigma/\sqrt{cm}$ points, and a small bound on PR. Throughout this manuscript, we will treat α as the specific value obtained by solving (7), and *not* as a variable.

High dimensional non-Gaussian distributions: In Appendix E, we study M_{C3D} when applied to d–dimensional distributions. In Theorem 7, we show that under bounded variance assumptions, $s^\star$ is an ε_m-approximate Nash equilibrium and that $\mathrm{PR}(M_{\text{C3D}}, s^\star) \leq 2 + \varepsilon_m$ where ε_m vanishes with m.

3.1 Proof sketch of Theorem 1

When $m \leq 4$: First, consider the (easy) case $m \leq 4$. At $s_i^\star$, the total amount of data collected is $\sigma/\sqrt{c}$ (see $n_i^\star$ in (8)), and as there is no corrupted dataset, $h_i^\star$ simply reduces to the sample mean of $X_i \cup Y_{-i}$. The mechanism is IR since an agent's penalty will be $\sigma\sqrt{c}(1 + 1/m)$ which is smaller than $p_{\min}^{\mathrm{IR}}$ (6). It is approximately efficient since the social penalty is $\sigma\sqrt{c}(m + 1)$ which is at most twice the global minimum $2\sigma\sqrt{mc}$ when $m \leq 4$. Finally, IC is guaranteed by the same argument used in (6); as long as the total amount of data is less than $\sigma/\sqrt{c}$, the cost of collecting one more point is offset by the marginal decrease in the estimation error. Moreover, as A_i does not depend on f_i under these conditions, there is no incentive to fabricate or falsify data.

When $m > 4$: We will divide this proof into four parts. We first show that $G(\alpha) = 0$ in line (6) has a solution α larger than $\sqrt{n_i^\star} = \sqrt{\sigma}(cm)^{-1/4}$. This will also be useful when analyzing the efficiency.

1. Equation (7) has a solution. We derive an asymptotic expansion of $\mathrm{Erfc}(\cdot)$ using integration by parts to analyze the solution to (7). When $m \geq 5$, we show that $G\left(\sqrt{n_i^\star}\right) \times G\left(\sqrt{n_i^\star}(1 + 8/\sqrt{m})\right) < 0$. By continuity of G, there exists $\alpha_m \in \left(\sqrt{n_i^\star}, \sqrt{n_i^\star}(1 + 8/\sqrt{m})\right)$ s.t. $G(\alpha_m) = 0$. For m large enough such that the residual in the asymptotic expansion is negligible, we show $\alpha_m \in \left(\sqrt{n_i^\star}, \sqrt{n_i^\star}(1 + \log m/m)\right)$ via an identical technique.

2. The strategies $s^\star$ in (8) is a Nash equilibrium: We show this via the following two steps. First (**2.1**), We show that fixing any n_i, the maximum risk and thus the penalty p_i is minimized when agent i submits the raw data and uses the weighted average as specified in (8), i.e for all n_i,

$$p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)) \leq p_i(M_{\text{C3D}}, ((n_i, f_i, h_i), s_{-i}^\star)), \quad \forall (n_i, f_i, h_i) \in \mathbb{N} \times \mathcal{F} \times \mathcal{H}. \quad (9)$$

Second (**2.2**), we show that p_i is minimized when agent i collects $n_i^\star$ samples under $(f_i^\star, h_i^\star)$, i.e.

$$p_i(M_{\text{C3D}}, ((n_i^\star, f_i^\star, h_i^\star), s_{-i}^\star)) \leq p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)), \quad \forall n_i \in \mathbb{N}. \quad (10)$$

7

2.1: Proof of (9). As the data collection cost does not change for fixed n_i, it is sufficient to show that $(f_i^\star, h_i^\star)$ minimizes the maximum risk. The following is a well-known recipe for proving minimax optimality of an estimator [23]: design a sequence of priors $\{\Lambda_\ell\}_\ell$, compute the minimum Bayes' risk $\{R_\ell\}_\ell$ for any estimator, and then show that R_ℓ converges to the maximum risk of the proposed estimator as $\ell \to \infty$. To apply this recipe, we use a sequence of normal priors $\Lambda_\ell = \mathcal{N}(0, \ell^2)$ for μ.

Before we proceed, we need to handle two issues. The first of these concerns the posterior for μ when conditioned on (X_i, Z_i, Z_i'). Since the corruption terms $\epsilon_{z,i}$ added to Z_i' depend on X_i and Z_i, this dataset is not independent. Moreover, as the variance η_i^2 is the difference between two normal random variables, Z_i' is not normal. Despite these, we are able to show that the posterior $\mu|(X_i, Z_i, Z_i')$ is normal. The second challenge is that the submission f_i also affects the estimation error as it determines the amount of noise η_i^2. We handle this by viewing $\mathcal{F} \times \mathcal{H}$ as a rich class of estimators and derive the optimal Bayes' estimator $(f_{i,\ell}^B, h_{i,\ell}^B) \in \mathcal{F} \times \mathcal{H}$ under the prior Λ_ℓ. We then show that the minimum Bayes' risk converges to the maximum risk when using $(f_i^\star, h_i^\star)$.

Next, under the prior $\Lambda_\ell = \mathcal{N}(0, \ell^2)$, we can minimize the Bayes' risk with respect to $h_i \in \mathcal{H}$ by setting $h_{i,\ell}^B$ to be the posterior mean. Then, the minimum Bayes' risk R_ℓ can be written as,

$$R_\ell = \inf_{f_i \in \mathcal{F}} \mathbb{E}\left[\left(|Z_i'|\left(\sigma^2 + \alpha^2\left(\frac{1}{|Y_i|}\sum_{y \in Y_i} y - \frac{1}{|Z_i|}\sum_{z \in Z_i} z\right)^2\right)^{-1} + \frac{|X_i| + |Z_i|}{\sigma^2} + \frac{1}{\ell^2}\right)^{-1}\right]$$

Note that $Y_i = f_i(X_i)$ depends on f_i. Via the Hardy-Littlewood inequality [10], we can show that the above quantity is minimized when $f_{i,\ell}^B$ is chosen to be a shrunk version of the agent's initial dataset X_i, i.e $f_{i,\ell}^B(X_i) = \left\{\left(1 + \sigma^2/(|X|\ell^2)\right)^{-1} x, \ \forall x \in X_i\right\}$. This gives us an expression for the minimum Bayes' risk R_ℓ under prior Λ_ℓ. To conclude the proof, we note that the minimum Bayes' risk under any prior is a lower bound on the maximum risk, and show that R_ℓ approaches the maximum risk of $(f_i^\star, h_i^\star)$ from below. Hence, $(f_i^\star, h_i^\star)$ is minimax optimal for any n_i. (Above, it is worth noting that $f_{i,\ell}^B \to f_i^\star = \mathbf{I}$ as $\ell \to \infty$. In the Appendix, we also find that $h_{i,\ell}^B \to h_i^\star$.)

2.2: Proof of (10). We can now write $p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)) = R_\infty + cn_i$, where R_∞ is the maximum risk of $(f_i^\star, h_i^\star)$ (and equivalently, the limit of the minimum Bayes' risk):

$$R_\infty := \mathbb{E}_{x \sim \mathcal{N}(0,1)}\left[\left((m-2)n_i^\star\left(\sigma^2 + \alpha^2(\sigma^2/n_i + \sigma^2/n_i^\star)x^2\right)^{-1} + (n_i + n_i^\star)\sigma^{-2}\right)^{-1}\right]$$

The term inside the expectation is convex in n_i for each fixed x. As expectation preserves convexity, we can conclude that p_i is a convex function of n_i. The choice of α in (7) ensures that the derivative is 0 at $n^\star$ which implies that $n^\star$ is a minimum of this function.

3. M_{C3D} is individually rational at $s^\star$: This is a direct consequence of step 2 as we can show that an agent 'working on her own' is a valid strategy in M_{C3D}.

4. M_{C3D} is approximately efficient at $s^\star$: By observing that the global minimum penalty is $2\sigma\sqrt{cm}$, we use a series of algebraic manipulations to show $\text{PR}(M_{\text{C3D}}, s^\star) = \frac{1}{2}\left(\frac{10\alpha^2/n_i^\star - 1}{4(m+1)\alpha^2/(mn_i^\star) - 1} + 1\right)$. As $\alpha > \sqrt{n_i^\star}$, some simple algebra leads to $\text{PR}(M_{\text{C3D}}, s^\star) < 2$. (The above expression holds for the specific value of α in (7), and should not be viewed as a function of α.)

4 Special Cases: Restricting the Agents' Strategy Space

In this section, we study two special cases motivated by some natural use cases, where we restrict the agents' strategy space. In addition to providing better guarantees on the efficiency, this will also help us better illustrate the challenges in our original setting.

4.1 Agents cannot fabricate or falsify data

First, we study a setting where agents are not allowed to fabricate data or falsify data. Specifically, in (2), $\mathcal{F}$ is restricted to functions which map a dataset to any subset. This is applicable when there are regulations preventing such behavior (e.g government institutions, hospitals)

Mechanism: The discussion at the end of §2 motivates the following modification to the pooling mechanism. We set the allocation space to be $\mathcal{A} = \bigcup_{n \geq 0} \mathbb{R}^n$, i.e the space of all datasets. If an

8

agent i submits at least $\sigma/\sqrt{cm}$ points, then given her all the other agents' datasets, i.e $A_i = \cup_{j \neq i} Y_j$; otherwise, set $A_i = \varnothing$. The recommended strategy $s_i^\star = (n_i^\star, f_i^\star, h_i^\star)$ of each agent is to collect $\sigma/\sqrt{cm}$ points, submit it as is $f_i^\star = \mathbf{I}$, and then use the sample mean of $Z_i \cup A_i$ to estimate μ. The theorem below, whose proof is straightforward, states the main properties of this mechanism.

Theorem 2. *The following statements about the mechanism and strategy profile $s^\star$ in the paragraph above are true when $\mathcal{F}$ is restricted to functions which map a dataset to any subset: (i) $s^\star$ is a Nash equilibrium. (ii) The mechanism is individually rational at $s^\star$. (iii) At $s^\star$, the mechanism is efficient.*

It is not hard to see that this mechanism can be easily manipulated by the agent if there are no restrictions on $\mathcal{F}$. As the mechanism only checks for the amount of data submitted, the agent can submit a fabricated dataset of $\sigma/\sqrt{cm}$ points, and then discard this dataset when computing the estimate, which results in detrimental free-riding. Any naive modification to check for the quality of the data can also be sidestepped by agents who sample only a few points, and use that to fabricate a larger dataset (e.g by sampling a large number of points from a Gaussian fitted to the small sample).

4.2 Agents accept an estimated value from the mechanism

Our next setting is motivated by use cases where the mechanism may directly deploy the estimated value for μ in some downstream application for the agent, i.e the agents are forced to use this value. This is motivated by federated learning, where agents collect and send data to a server (mechanism), which deploys a model (estimate) directly on the agent's device [9, 22]. This requires modifying the agent's strategy space to $\mathcal{S} = \mathbb{N} \times \mathcal{F}$. Now, an agent can only choose (n_i, f_i), how much data she wishes to collect, and how to fabricate or falsify the dataset. A mechanism is defined as a procedure $b : \left(\bigcup_{n \geq 0} \mathbb{R}^n\right)^m \to \mathbb{R}^m$, which maps m datasets to m estimated mean values.

Algorithm 3 (see Appendix D) outlines a family of mechanisms parametrized by $\epsilon > 0$ for this setting. As we will see shortly, with parameter ϵ, the mechanism can achieve a PR of $(1+\epsilon)$. This mechanism computes agent i's estimate for μ as follows. First, let Y_{-i} be the union of all datasets submitted by the other agents. Similar to Algorithm 1, the algorithm individually adds Gaussian to each Y_{-i} to obtain Z_i (line 10). Unlike before, this noise is added to the entire dataset and the variance η_i^2 of this noise depends on the difference between the sample means of the agent's submission Y_i and all of the other agents' submissions Y_{-i}. It also depends on two ϵ-dependent parameters defined in line 6. Finally, the mechanism deploys the sample mean of $Y_i \cup Z_i$ as the estimate for μ. The recommended strategies $s_i^\star = (n_i^\star, f_i^\star)$ for the agents is to simply collect $n_i^\star = \sigma/\sqrt{cm}$ points and submit it as is $f_i^\star = \mathbf{I}$. The following theorem states the main properties of the mechanism.

Theorem 3. *Let $\epsilon > 0$. The following statements about Algorithm 3 and the strategy profile $s^\star$ given in the paragraph above are true: (i) $s^\star$ is a Nash equilibrium. (ii) The mechanism is individually rational at $s^\star$. (iii) At $s^\star$, the mechanism is approximately efficient with $\mathrm{PR}(M, s^\star) \leq 1 + \epsilon$.*

The above theorem states that it is possible to obtain a social penalty that is arbitrarily close to the global minimum under the given restriction of the strategy space. However, this mechanism is not IC if agents are allowed to design their own estimator. For instance, if the mechanism returns $A_i = Z_i$ (line 10), then using a weighted average of the data in X_i and Z_i yields a lower estimation error than simple average used by the mechanism (see Appendix D). An agent can leverage this insight to collect and submit less data and obtain a lower overall penalty at the expense of other agents.

5 Conclusion

We studied collaborative normal mean estimation in the presence of strategic agents. Naive mechanisms which only look at the quantity of the dataset submitted, can be manipulated by agents who under-collect and/or fabricate data, leaving all agents worse off. To address this issue, when sharing the others' data with an agent, our mechanism M_{C3D} corrupts this dataset proportional to how much the data reported by the agent differs from the other agents. We design minimax optimal estimators for this corrupted dataset to achieve a socially desirable Nash equilibrium.

Future directions: We believe that designing mechanisms for other collaborative learning settings may require relaxing the *exact* IC guarantees to make the analysis tractable. For many learning problems, it is difficult to design exactly optimal estimators, and it is common to settle for rate-optimal (i.e up to constants) estimators [23]. For instance, even simply relaxing to high dimensional distributions with bounded variance, M_{C3D} can only provide an approximate IC guarantee.

9

References

[1] Ads Data Hub. `https://developers.google.com/ads-data-hub/guides/intro`. Accessed: 2022-05-10.

[2] PubChem. `https://pubchem.ncbi.nlm.nih.gov/`. Accessed: 2022-05-10.

[3] Data Coordinating Center Burton Robert 67 Jensen Mark A 53 Kahn Ari 53 Pihl Todd 53 Pot David 53 Wan Yunhu 53 and Tissue Source Site Levine Douglas A 68. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[4] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726, 2019.

[5] Anish Agarwal, Munther Dahleh, Thibaut Horel, and Maryann Rui. Towards data auctions with externalities. *arXiv preprint arXiv:2003.08345*, 2020.

[6] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning Prices for Repeated Auctions with Strategic Buyers. In *Advances in Neural Information Processing Systems*, pages 1169–1177, 2013.

[7] Elliot Anshelevich, Anirban Dasgupta, Jon Kleinberg, Éva Tardos, Tom Wexler, and Tim Roughgarden. The price of stability for network design with fair cost allocation. *SIAM Journal on Computing*, 38(4):1602–1623, 2008.

[8] Susan Athey and Ilya Segal. An Efficient Dynamic Mechanism. *Econometrica*, 81(6):2463–2485, 2013.

[9] Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, and Han Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *International Conference on Machine Learning*, pages 1005–1014. PMLR, 2021.

[10] Almut Burchard. A short course on rearrangement inequalities. *Lecture notes, IMDEA Winter School, Madrid*, 2009.

[11] Yiding Chen, Xuezhou Zhang, Kaiqing Zhang, Mengdi Wang, and Xiaojin Zhu. Byzantine-robust online and offline distributed reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3230–3269. PMLR, 2023.

[12] Edward H Clarke. Multipart Pricing of Public Goods. *Public Choice*, 1971.

[13] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*.

[14] Ningning Ding, Zhixuan Fang, and Jianwei Huang. Incentive mechanism design for federated learning with multi-dimensional private information. In *2020 18th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, pages 1–8. IEEE, 2020.

[15] Mona Flores, Ittai Dayan, Holger Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Abidin, Andrew Liu, Anthony Costa, Bradford Wood, et al. Federated learning used for predicting outcomes in sars-cov-2 patients. *Research Square*, 2021.

[16] Yann Fraboni, Richard Vidal, and Marco Lorenzi. Free-rider attacks on model aggregation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1846–1854. PMLR, 2021.

[17] Theodore Groves. Efficient Collective Choice when Compensation is Possible. *The Review of Economic Studies*, 1979.

[18] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

10

[19] Charles I Jones and Christopher Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–58, 2020.

[20] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[21] Sham M Kakade, Ilan Lobel, and Hamid Nazerzadeh. An Optimal Dynamic Mechanism for Multi-armed Bandit Processes. *arXiv preprint arXiv:1001.4598*, 2010.

[22] Sai Praneeth Karimireddy, Wenshuo Guo, and Michael I Jordan. Mechanisms that incentivize data sharing in federated learning. *arXiv preprint arXiv:2207.04557*, 2022.

[23] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[24] Jierui Lin, Min Du, and Jian Liu. Free-riders in federated learning: Attacks and defenses. *arXiv preprint arXiv:1911.12560*, 2019.

[25] Yuan Liu, Mengmeng Tian, Yuxin Chen, Zehui Xiong, Cyril Leung, and Chunyan Miao. A contract theory based incentive mechanism for federated learning. In *Federated and Transfer Learning*, pages 117–137. Springer, 2022.

[26] Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. 2021.

[27] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian Incentive-compatible Bandit Exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582, 2015.

[28] Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. Dynamic Cost-per-action Mechanisms and Applications to Online Advertising. In *Proceedings of the 17th International Conference on World Wide Web*, pages 179–188, 2008.

[29] Ariel D Procaccia. Cake Cutting: Not just Child's Play. *Communications of the ACM*, 56(7): 78–87, 2013.

[30] Alvin E Roth. On the Allocation of Residents to Rural Hospitals: A General Property of Two-sided Matching Markets. *Econometrica: Journal of the Econometric Society*, pages 425–427, 1986.

[31] Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney Exchange. *The Quarterly journal of economics*, 119(2):457–488, 2004.

[32] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 92–104. Springer, 2019.

[33] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *International conference on machine learning*, pages 8927–8936. PMLR, 2020.

[34] William Vickrey. Counterspeculation, Auctions, and Competitive Sealed Tenders. *The Journal of Finance*, 1961.

[35] Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, 1939.

[36] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive*, pages 153–167, 2020.

11

[37] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. *Advances in Neural Information Processing Systems*, 34:16104–16117, 2021.

[38] Wenting Zheng, Raluca Ada Popa, Joseph E Gonzalez, and Ion Stoica. Helen: Maliciously secure coopetitive learning for linear models. In *2019 IEEE symposium on security and privacy (SP)*, pages 724–738. IEEE, 2019.

12

Note to reviewers: In the main text of the supplementary submission, we fixed 3 typos in equations and rephrased a potentially confusing sentence in line 272 after the deadline. The original submission can still be evaluated without any significant hindrance in understanding the paper.

A Proof of Theorem 1

In this section, we prove Theorem 1. This section is organized as follows. First, in §A.1, we consider the case where $m \leq 4$. In the remainder of this section, we will assume $m \geq 5$. First, in §A.2, we will show that (7) can be solved for α and state some properties about the solution. Then, in §A.3, we will prove the incentive compatibility result, in §A.4 we will prove individual rationality, and in §A.5, we will prove the result on efficiency.

A.1 When $m \leq 4$

First, consider the (easy) case $m \leq 4$. At $s_i^\star$, the total amount of data collected is $\sigma/\sqrt{c}$ as each agent will be collecting $n_i^\star = \frac{\sigma}{m\sqrt{c}}$ (see (8)). As there is no corrupted dataset, $h_i^\star$ simply reduces to the sample mean of $X_i \cup Y_{-i}$. The individual rationality property follows from the following simple calculation:

$$p_i(M_{\text{C3D}}, s^\star) = \left(1 + \frac{1}{m}\right)\sqrt{c}\sigma < 2\sqrt{c}\sigma = p_{\min}^{\text{IR}}.$$

Similarly, the bound on the ratio between the penalties can also be obtained via the following calculation:

$$\text{PR} = \frac{m\left(1 + \frac{1}{m}\right)\sqrt{c}\sigma}{2\sigma\sqrt{cm}} < \sqrt{m} \leq 2.$$

Finally, to show IC, consider agent i and assume that all other agents have followed the recommended strategies, i.e collected $\sigma/(m\sqrt{c})$. Then, the agent will have an *uncorrupted* dataset $Y_{-i} = \bigcup_{j \neq i} X_j$ of $n_{-i}^\star = (m-1)\sigma/(m\sqrt{c})$ points with no corruption. Regardless of what she chooses to submit, the best estimator she could use with the union of this dataset Y_{-i} and the data she collects X_i and will be the sample mean as it is minimax optimal. The number of points that minimizes her penalty is,

$$\operatorname*{argmin}_{n_i} \left(\sup_\mu \mathbb{E}\left[(h_i(X_i, Y_i, Y_{-i}) - \mu)^2 \,\Big|\, \mu \right] \;+\; cn_i \right) = \operatorname*{argmin}_{n_i \in \mathbb{R}} \left(\frac{\sigma^2}{n_i + n_{-i}^\star} + cn_i \right) = \frac{\sigma}{m\sqrt{c}}$$

Finally, as A_i does not depend on f_i under these conditions, there is no incentive to fabricate or falsify data, i.e choosing anything other than $f^\star = \mathbf{I}$ does not lower her utility.

In the remainder of this section, will study the harder case, $m \geq 4$.

A.2 Existence of a solution to (7) and some of its properties

In this section, we show that $G\left(\frac{\sigma^{1/2}}{(cm)^{1/4}}\right) < 0$ and $G\left(\left(1 + \frac{C_m}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right) > 0$, where $C_m = 20$ when $m \leq 20$ and $C_m = 5$ when $m > 20$. This means equation $G(\alpha) = 0$ has solution in $\left(\frac{\sigma^{1/2}}{(cm)^{1/4}}, \left(1 + \frac{C_m}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right)$.

First, in Lemma 12, we derive an asymptotic expansion of the Gaussian complementary error function, and construct lower and upper bounds for $G(\alpha)$ that are easier to work with. We have restated these lower (Erfc_{LB}) and upper (Erfc_{UB}) bounds below.

$$\text{Erfc}_{\text{UB}}(x) := \frac{1}{\sqrt{\pi}}\left(\frac{\exp(-x^2)}{x} - \frac{\exp(-x^2)}{2x^3} + \frac{3\exp(-x^2)}{4x^5} \right) \tag{11}$$

$$\text{Erfc}_{\text{LB}}(x) := \frac{1}{\sqrt{\pi}}\left(\frac{\exp(-x^2)}{x} - \frac{\exp(-x^2)}{2x^3} \right) \tag{12}$$

We can now use this to derive the following lower (G_{LB}) and upper (G_{UB}) bounds on G. Here, we have used the fact that $4(m+1)\frac{\alpha^2}{\sigma\sqrt{m/c}} - 1 > 0$ when $\alpha \geq (\sigma/\sqrt{cm})^{1/2}$. We have:

$$G_{\text{LB}}(\alpha) := \left(\frac{m-4}{m-2}\frac{4\alpha^2}{\sigma/\sqrt{cm}} - 1 \right)\frac{4\alpha}{\sqrt{\sigma}(m/c)^{1/4}}$$

13
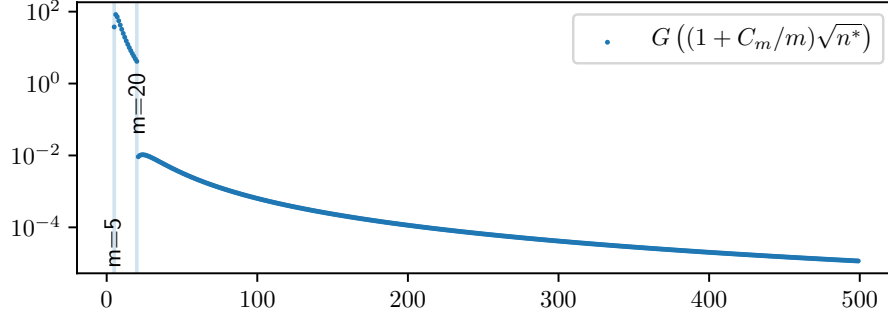
Figure 1: Plot for $G\left(\left(1 + \frac{C_m}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right)$. See `G_em_plot.py`. The discontinuity at $m = 20$ is due to the different values for C_m when $m \leq 20$ and when $m > 20$.

$$G_{\text{UB}}(\alpha) := \begin{array}{l} -\left(4(m+1)\frac{\alpha^2}{\sigma\sqrt{m/c}} - 1\right)\sqrt{2\pi}\exp\left(\frac{\sigma\sqrt{m/c}}{8\alpha^2}\right)\text{Erfc}_{\text{UB}}\left(\frac{\sqrt{\sigma}(m/c)^{1/4}}{2\sqrt{2}\alpha}\right), \\[2ex] \left(\frac{m-4}{m-2}\frac{4\alpha^2}{\sigma/\sqrt{cm}} - 1\right)\frac{4\alpha}{\sqrt{\sigma}(m/c)^{1/4}} \\[2ex] -\left(4(m+1)\frac{\alpha^2}{\sigma\sqrt{m/c}} - 1\right)\sqrt{2\pi}\exp\left(\frac{\sigma\sqrt{m/c}}{8\alpha^2}\right)\text{Erfc}_{\text{LB}}\left(\frac{\sqrt{\sigma}(m/c)^{1/4}}{2\sqrt{2}\alpha}\right). \end{array}$$

By first, substituting $\sigma/\sqrt{cm}$ for α in the expressions for G_{UB} and Erfc_{UB}, and then via a sequence of algebraic manipulations, we can verify that

$$G\left(\frac{\sigma^{1/2}}{(cm)^{1/4}}\right) \leq G_{\text{UB}}\left(\frac{\sigma^{1/2}}{(cm)^{1/4}}\right)$$

$$= \frac{4\left(\frac{4(m-4)}{m-2} - 1\right)\left(\frac{\sigma}{\sqrt{cm}}\right)^{1/2}}{\sqrt{\sigma}\left(\frac{m}{c}\right)^{1/4}} - \sqrt{2}\left(\frac{4(m+1)}{\sqrt{\frac{m}{c}}\sqrt{cm}} - 1\right)\left(\frac{2\sqrt{2}\left(\frac{\sigma}{\sqrt{cm}}\right)^{1/2}}{\sqrt{\sigma}\left(\frac{m}{c}\right)^{1/4}} - \frac{8\sqrt{2}\left(\frac{\sigma}{\sqrt{cm}}\right)^{3/2}}{\sigma^{3/2}\left(\frac{m}{c}\right)^{3/4}}\right)$$

$$= -\frac{128}{(m-2)m^{5/2}} < 0.$$

Next, we will show that $G\left(\left(1 + \frac{C_m}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right) > 0$ by studying the lower bound G_{LB}. For $m \in [5, 500]$, we can verify individually that $G\left(\left(1 + \frac{C_m}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right) > 0$ (See Figure 1). For $m > 500$, we have:

$$G\left(\left(1 + \frac{C_m}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right) = G\left(\left(1 + \frac{5}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right) \geq G_{\text{LB}}\left(\left(1 + \frac{5}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right)$$

$$= \frac{4\left(\frac{4\left(\frac{5}{m}+1\right)^2(m-4)}{m-2} - 1\right)\left(\frac{5}{m}+1\right)\left(\frac{\sigma}{\sqrt{cm}}\right)^{1/2}}{\sqrt{\sigma}\left(\frac{m}{c}\right)^{1/4}}$$

$$- \sqrt{2}\left(\frac{4\left(\frac{5}{m}+1\right)^2(m+1)}{\sqrt{\frac{m}{c}}\sqrt{cm}} - 1\right)\left(\frac{2\sqrt{2}\left(\frac{5}{m}+1\right)\left(\frac{\sigma}{\sqrt{cm}}\right)^{1/2}}{\sqrt{\sigma}\left(\frac{m}{c}\right)^{1/4}} - \frac{8\sqrt{2}\left(\frac{5}{m}+1\right)^3\left(\frac{\sigma}{\sqrt{cm}}\right)^{3/2}}{\sigma^{3/2}\left(\frac{m}{c}\right)^{3/4}}\right)$$

$$+ \frac{96\sqrt{2}\left(\frac{5}{m}+1\right)^5\left(\frac{\sigma}{\sqrt{cm}}\right)^{5/2}}{\sigma^{5/2}\left(\frac{m}{c}\right)^{5/4}}\right)$$

$$= \frac{64(m+5)^3\left(m^6 - 191m^5 - 1566m^4 - 3920m^3 + 2100m^2 + 19500m + 15000\right)}{(m-2)m^{21/2}}.$$

14

544 When $m > 500$,
$$m^6 - 191m^5 - 1566m^4 - 3920m^3 = m^3(m^3 - 191m^2 - 1566m - 3920)$$
$$> m^3((200 + 200 + 100)m^2 - 191m^2 - 1566m - 3920)$$
$$> m^3(200m^2 + 10^5 m + 2.5 \times 10^7 - 191m^2 - 1566m - 3920) > 0.$$

545 Combining the results from the two previous displays, we have, $G\left(\left(1 + \frac{C_m}{m}\right)\frac{\sigma^{1/2}}{(cm)^{1/4}}\right) > 0$ which
546 completes the proof for this section.

A.3 Algorithm 1 is incentive compatible

548 In this section, we will prove the following lemma which states that $s_i^\star$, as defined in (8) is a Nash
549 equilibrium in M_{C3D}.

550 **Lemma 4** (NIC). *The recommended strategies $s^\star = \{(n_i^\star, f_i^\star, h_i^\star)\}_i$ as defined in (8) in mechanism*
551 M_{C3D} *(Algorithm 1) satisfies:*
$$p_i(M_{\text{C3D}}, s^\star) \le p_i(M_{\text{C3D}}, (s_i, s_{-i}^\star))$$
552 *for all $i \in [m]$ and $s_i \in \mathbb{N} \times \mathcal{F} \times \mathcal{H}$.*

553 The Proof of Lemma 4 relies on the following two lemmas:

554 **Lemma 5** (Optimal Estimation and Submission). *For all $i \in [m]$ and $(n_i, f_i, h_i) \in \mathbb{N} \times \mathcal{F} \times \mathcal{H}$.*
$$p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)) \le p_i(M_{\text{C3D}}, ((n_i, f_i, h_i), s_{-i}^\star)).$$

555 See the Proof of Lemma 5 in §A.3.1

556 **Lemma 6** (Optimal Sample Size). *For all $i \in [m]$ and $n_i \in \mathbb{N}$.*
$$p_i(M_{\text{C3D}}, ((n_i^\star, f_i^\star, h_i^\star), s_{-i}^\star)) \le p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)).$$

557 See the Proof of Lemma 6 in §A.3.2

558 *Proof of Lemma 4.* By Lemma 5 and 6, we have, for all $i \in [m]$ and $s_i' = (n_i, f_i, h_i) \in \mathbb{N} \times \mathcal{F} \times \mathcal{H}$,
$$p_i(M_{\text{C3D}}, s^\star) = p_i(M_{\text{C3D}}, ((n_i^\star, f_i^\star, h_i^\star), s_{-i}^\star)) \le p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star))$$
$$\le p_i(M_{\text{C3D}}, ((n_i, f_i, h_i), s_{-i}^\star)) = p_i(M_{\text{C3D}}, (s_i', s_{-i}^\star))$$
559 $\square$

A.3.1 Proof of Lemma 5

561 In this section, we will prove Lemma 5, which, intuitively states that, regardless of the amount
562 of data collected, agent i should submit the data as is ($f_i^\star = \mathbf{I}$) and use the weighted average
563 estimator in (8) to estimate μ. We will do so via the following three step procedure, inspired by
564 well–known techniques for proving minimax optimality of estimators (e.g see Theorem 1.12, Chapter
565 5 of Lehmann and Casella [23]).

    1. First, we construct a sequence of prior distributions $\{\Lambda_\ell\}_{\ell \ge 1}$ for μ and calculate the sequence
567   of Bayesian risks under the prior distributions:
$$R_\ell := \inf_{f_i \in \mathcal{A}, h_i \in \mathcal{H}} \mathbb{E}_{\mu \sim \Lambda_\ell}\left[\mathbb{E}\left[(h_i(X_i, f_i(X_i), A_i) - \mu)^2 \,\middle|\, \mu\right]\right], \quad \ell \ge 1.$$

568   2. Then, we will show that $\lim_{\ell \to \infty} R_\ell = \sup_\mu \mathbb{E}\left[(h_i^\star(X_i, f_i^\star(X_i), A_i) - \mu)^2 \,\middle|\, \mu\right]$.
569   3. Finally, as the Bayesian risk is a lower bound on maximum risk, we will conclude that
570       $(f_i^\star, h_i^\star)$ is minimax optimal.

571 Without loss of generality, we focus only on the deterministic f_i and h_i. If either of them are
572 stochastic, we can condition on the external source of randomness and treat them as deterministic
573 functions. Our proof holds for any realization of this external source of randomness, and hence it will
574 hold in expectation as well. Similarly, Z_i is randomly chosen in Algorithm 1. In the following, we
575 condition on this randomness and the entire proof will carry through.

576 Note that $Y_i = f_i(X_i)$. We will use both of them interchangeably in the subsequent proof.

15

577 **Step 1 (Bounding the Bayes' risk under the sequence of priors):** We will use a sequence of
578 normal priors $\Lambda_\ell := \mathcal{N}(0, \ell^2)$ for all $\ell \geq 1$. To bound the Bayes' risk under these priors, we will first
579 note that for a fixed $f_i \in \mathcal{F}$,

$$x|\mu \sim \mathcal{N}(\mu, \sigma^2) \quad \forall x \in X_i \cup Z_i; \tag{13}$$

$$x|\mu, \eta_i^2 \sim \mathcal{N}(\mu, \sigma^2 + \eta_i^2) \quad \forall x \in Z_i'. \tag{14}$$

580 Here, recall that η_i^2 is a function of Y_i and Z_i. Because both $Y_i = f_i(X_i)$ and η_i^2 are deterministic
581 functions of X_i, Z_i when f_i is fixed, the posterior distribution for μ conditioned on (X_i, Y_i, A_i) can
582 be calculated as follows:

$$p(\mu|X_i, Y_i, A_i) = p\big(\mu|X_i, Y_i, Z_i, Z_i', \eta_i^2\big) = p(\mu|X_i, Z_i, Z_i')$$

$$\propto p(\mu, X_i, Z_i, Z_i') = p(Z_i'|X_i, Z_i, \mu)p(X_i, Z_i|\mu)p(\mu) = p(Z_i'|X_i, Z_i, \mu)p(X_i|\mu)p(Z_i|\mu)p(\mu)$$

$$\propto \exp\left(-\frac{1}{2(\sigma^2 + \eta_i^2)}\sum_{x \in Z_i'}(x - \mu)^2\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{x \in X_i \cup Z_i}(x - \mu)^2\right)\exp\left(-\frac{\mu^2}{2\ell^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2} + \frac{1}{\ell^2}\right)\mu^2\right)\exp\left(\frac{1}{2}2\left(\frac{\sum_{x \in Z_i'} x}{\sigma^2 + \eta_i^2} + \frac{\sum_{x \in X_i \cup Z_i} x}{\sigma^2}\right)\mu\right)$$

$$= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_\ell^2}\mu^2 - 2\frac{\mu_\ell}{\sigma_\ell^2}\mu\right)\right) \propto \exp\left(-\frac{1}{2\sigma_\ell^2}(\mu - \mu_\ell)^2\right),$$

583 where

$$\mu_\ell = \frac{\frac{\sum_{x \in Z_i'} x}{\sigma^2 + \eta_i^2} + \frac{\sum_{x \in X_i \cup Z_i} x}{\sigma^2}}{\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2} + \frac{1}{\ell^2}}, \quad \text{and} \quad \sigma_\ell^2 = \frac{1}{\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2} + \frac{1}{\ell^2}}. \tag{15}$$

584 We can therefore conclude that (despite the non i.i.d nature of the data), the posterior for μ is Gaussian
585 with mean and variance as shown above. We have:

$$\mu|X_i, Y_i, A_i \sim \mathcal{N}(\mu_\ell, \sigma_\ell^2).$$

586 Next, following standard steps (See Corollary 1.2 in Chapter 4 of [23]), we know that
587 $\mathbb{E}_\mu\Big[(h_i(X_i, Y_i, A_i) - \mu)^2|X_i, Y_i, A_i\Big]$ is minimized when $h_i(X_i, Y_i, A_i) = \mathbb{E}_\mu[\mu|X_i, Y_i, A_i] = \mu_\ell$.
588 This shows that for any $f_i \in h_i$, the optimal h_i is simply the posterior mean of μ under the prior Λ_ℓ
589 conditioned on $(X_i, f_i(X_i), A_i)$. We can rewrite the minimum averaged risk over $\mathcal{H}$ by switching
590 the order of expectation:

$$\inf_{h_i \in \mathcal{H}} \mathbb{E}_{\mu \sim \Lambda_\ell}\Big[\mathbb{E}\Big[(h_i(X_i, Y_i, A_i) - \mu)^2|\mu\Big]\Big]$$

$$= \inf_{h_i \in \mathcal{H}} \mathbb{E}_{X_i, Z_i, Z_i'}\Big[\mathbb{E}_\mu\Big[(h_i(X_i, Y_i, A_i) - \mu)^2|X_i, Z_i, Z_i'\Big]\Big]$$

$$= \mathbb{E}_{X_i, Z_i, Z_i'}\Big[\mathbb{E}_\mu\Big[(\mu_\ell - \mu)^2|X_i, Z_i, Z_i'\Big]\Big] = \mathbb{E}_{X_i, Z_i, Z_i'}\big[\sigma_\ell^2\big]$$

$$= \mathbb{E}_{X_i, Z_i}\left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2} + \frac{1}{\ell^2}}\right], \tag{16}$$

591 the expectation in the last step involves only X_i, Z_i because σ_ℓ^2 depends only on X_i, Z_i and $|Z_i'|$, but
592 not the instantiation of Z_i'.

593 Next, we will show that (16) is minimized for the following choice of f_i which shrinks each points in
594 X_i by an amount that depends on the prior Λ_ℓ's variance ℓ^2:

$$f_i(X_i) = \left\{\frac{|X_i|/\sigma^2}{|X_i|/\sigma^2 + 1/\ell^2} x \text{ , for each } x \in X_i\right\}. \tag{17}$$

595 **Remark 1.** *An interesting observation (albeit not critical to the proof) here is that f_i in (17) converges*
596 *pointwise to $f_i^\star$, i.e. **I**, as $\ell \to \infty$. This shows that the optimal submission function under the prior*
597 *converges to $f_i^\star$. We can make a similar observation about the posterior mean in (15), where μ_ℓ*
598 *converges to $h_i^\star$ as $\ell \to \infty$.*

16

To prove (17), we first define the following quantities.

$$\widehat{\mu}(X_i) := \frac{1}{|X_i|} \sum_{x \in X_i} x, \quad \widehat{\mu}(Y_i) := \frac{1}{|Y_i|} \sum_{x \in Y_i} x, \quad \widehat{\mu}(Z_i) := \frac{1}{|Z_i|} \sum_{s \in Z_i} x.$$

We will also find it useful to express η_i^2 as follows. Here α is as defined in (7). We have:

$$\eta_i^2 = \alpha^2 (\widehat{\mu}(Y_i) - \widehat{\mu}(Z_i))^2$$

The following calculations show that, conditioned on X_i, $\widehat{\mu}(Z_i) - \mu$ and $\mu - \frac{|X_i|/\sigma^2}{|X_i|/\sigma^2 + 1/\ell^2} \widehat{\mu}(X_i)$ are independent Gaussian random variables[3]:

$$p(\widehat{\mu}(Z_i) - \mu, \mu | X_i) \propto p(\widehat{\mu}(Z_i) - \mu, \mu, X_i)$$
$$= p(\widehat{\mu}(Z_i) - \mu, X_i | \mu) p(\mu) = p(\widehat{\mu}(Z_i) - \mu | \mu) p(X_i | \mu) p(\mu)$$
$$\propto \exp\left(-\frac{1}{2} \frac{|Z_i|}{\sigma^2} (\widehat{\mu}(Z_i) - \mu)^2 \right) \exp\left(-\frac{1}{2\sigma^2} \sum_{x \in X_i} (x - \mu)^2 \right) \exp\left(-\frac{1}{2\ell^2} \mu^2 \right)$$
$$\propto \underbrace{\exp\left(-\frac{1}{2} \frac{|Z_i|}{\sigma^2} (\widehat{\mu}(Z_i) - \mu)^2 \right)}_{\propto p(\widehat{\mu}(Z_i) - \mu | X_i)} \underbrace{\exp\left(-\frac{1}{2} \left(\frac{|X_i|}{\sigma^2} + \frac{1}{\ell^2} \right) \left(\mu - \frac{|X_i|/\sigma^2}{|X_i|/\sigma^2 + 1/\ell^2} \widehat{\mu}(X_i) \right)^2 \right)}_{\propto p\left(\mu - \frac{|X_i|/\sigma^2}{|X_i|/\sigma^2 + 1/\ell^2} \widehat{\mu}(X_i) | X_i \right)}$$

Thus conditioning on X_i, we can write

$$\begin{pmatrix} \widehat{\mu}(Z_i) - \mu \\ \mu - \frac{|X_i|/\sigma^2}{|X_i|/\sigma^2 + 1/\ell^2} \widehat{\mu}(X_i) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{|Z_i|} & 0 \\ 0 & \frac{1}{|X_i|/\sigma^2 + 1/\ell^2} \end{pmatrix} \right).$$

which leads us to

$$\widehat{\mu}(Z_i) - \frac{|X_i|/\sigma^2}{|X_i|/\sigma^2 + 1/\ell^2} \widehat{\mu}(X_i) \bigg| X_i \sim \mathcal{N}\left(0, \underbrace{\frac{\sigma^2}{|Z_i|} + \frac{1}{|X_i|/\sigma^2 + 1/\ell^2}}_{=:\tilde{\sigma}_\ell^2} \right) \quad (18)$$

Next, we will rewrite the squared difference in η_i^2 as follows:

$$\frac{\eta_i^2}{\alpha^2} = (\widehat{\mu}(Y_i) - \widehat{\mu}(Z_i))^2$$
$$= \left(\underbrace{\widehat{\mu}(Z_i) - \frac{|X_i|/\sigma^2}{|X_i|/\sigma^2 + 1/\ell^2} \widehat{\mu}(X_i)}_{=\tilde{\sigma}_\ell e} + \underbrace{\left(\frac{|X_i|/\sigma^2}{|X_i|/\sigma^2 + 1/\ell^2} \widehat{\mu}(X_i) - \widehat{\mu}(Y_i) \right)}_{=:\phi(X_i, f_i)} \right)^2.$$

Here, we observe that the first part of the RHS above is equal to $\tilde{\sigma}_\ell$, where e is a normal noise $e | X_i \sim \mathcal{N}(0, 1)$ and $\tilde{\sigma}_\ell$ is as defined in (18). For brevity, we will denote the second part of the RHS as $\phi(X_i, f_i)$, which intuitively characterizes the difference between X_i and Y_i. Importantly, $\phi(X_i, f_i) = 0$ when f_i is chosen to be (17).

Using e and ϕ, we can rewrite (16) using conditional expectation:

$$\mathbb{E}_{X_i, Z_i} \left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2} + \frac{1}{\ell^2}} \right] = \mathbb{E}_{X_i} \left[\mathbb{E}_{Z_i | X_i} \left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2} + \frac{1}{\ell^2}} \right] \right]$$
$$= \mathbb{E}_{X_i} \left[\mathbb{E}_{e | X_i} \left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \alpha^2 (\tilde{\sigma}_\ell e + \phi(X_i, f_i))^2} + \frac{|X_i| + |Z_i|}{\sigma^2} + \frac{1}{\ell^2}} \right] \right]$$

---

[3]This is akin to the observation that given $u, v \sim \mathcal{N}(0, 1)$, then $u - v$ and $u + v$ are independent.

17

$$= \mathbb{E}_{X_i}\left[\int_{-\infty}^{\infty} \underbrace{\frac{1}{\frac{|Z_i'|}{\sigma^2+\alpha^2\tilde\sigma_\ell^2(e+\phi(X_i,f_i)/\tilde\sigma_\ell)^2} + \frac{|X_i|+|Z_i|}{\sigma^2} + \frac{1}{\ell^2}}}_{=:F_1(e+\phi(X_i,f_i)/\tilde\sigma_\ell)} \underbrace{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{e^2}{2}\right)}_{=:F_2(e)} de\right], \quad (19)$$

where we use the fact that $e|X_i \sim \mathcal{N}(0,1)$ in the last step. To proceed, we will consider the inner expectation in the RHS above. For any fixed X_i, $F_1(\cdot)$ (as marked on the RHS) is an even function that monotonically increases on $[0,\infty)$ bounded by $\frac{\sigma}{|X_i|+|Z_i|}$ and $F_2(\cdot)$ (as marked on the RHS) is an even function that monotonically decreases on $[0,\infty)$. That means, for any $a \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} F_1(e-a)F_2(e)de \leq \int_{-\infty}^{\infty}\frac{\sigma}{|X_i|+|Z_i|}F_2(e)de = \frac{\sigma}{|X_i|+|Z_i|} < \infty.$$

By a corollary of the Hardy-Littlewood inequality in Lemma 9, we have

$$\int_{-\infty}^{\infty} F_1(e+\phi(X_i,f_i)/\tilde\sigma_\ell)F_2(e)de \geq \int_{-\infty}^{\infty} F_1(e)F_2(e)de, \quad (20)$$

the equality is achieved when $\phi(X_i,f_i)/\tilde\sigma_\ell = 0$. In particular, the equality holds when f_i is chosen as specified in (17).

Now, to complete Step 1, we combine (16), (19) and (20) to obtain

$$\inf_{h_i\in\mathcal{H}} \mathbb{E}_{\mu\sim\Lambda_\ell}\left[\mathbb{E}\left[(h_i(X_i,Y_i,A_i)-\mu)^2|\mu\right]\right] = \mathbb{E}_{X_i}\left[\int_{-\infty}^{\infty} F_1(e+\phi(X_i,f_i)/\tilde\sigma_\ell)F_2(e)de\right]$$

$$\geq \mathbb{E}_{X_i}\left[\int_{-\infty}^{\infty} F_1(e)F_2(e)de\right] = \int_{-\infty}^{\infty} F_1(e)F_2(e)de, \quad (21)$$

where the last step is because conditioning on each realization of X_i, the term inside the expectation is a constant. Using (21), we can write the Bayes risk R_ℓ under any prior Λ_ℓ as:

$$R_\ell := \inf_{f_i\in\mathcal{A},h_i\in\mathcal{H}} \mathbb{E}_{\mu\sim\Lambda_\ell}\left[\mathbb{E}\left[(h_i(X_i,Y_i,A_i)-\mu)^2\big|\mu\right]\right] = \int_{-\infty}^{\infty} F_1(e)F_2(e)de$$

$$= \mathbb{E}_{e\sim\mathcal{N}(0,1)}\left[\frac{1}{\frac{|Z_i'|}{\sigma^2+\alpha^2\tilde\sigma_\ell^2 e^2} + \frac{|X_i|+|Z_i|}{\sigma^2} + \frac{1}{\ell^2}}\right]$$

Because the term inside the expectation is bounded by $\frac{\sigma^2}{|X_i|+|Z_i|}$ and $\lim_{\ell\to\infty}\tilde\sigma_\ell^2 = \frac{\sigma^2}{|Z_i|} + \frac{\sigma^2}{|X_i|}$, we can use dominated convergence theorem to show that:

$$R_\infty := \lim_{\ell\to\infty} R_\ell = \mathbb{E}_{e\sim\mathcal{N}(0,1)}\left[\frac{1}{\frac{|Z_i'|}{\sigma^2+\alpha^2\left(\frac{\sigma^2}{|Z_i|}+\frac{\sigma^2}{|X_i|}\right)e^2} + \frac{|X_i|+|Z_i|}{\sigma^2}}\right] \quad (22)$$

Step 2: Maximum risk of $(f_i^\star, h_i^\star)$**:** Next, we will compute the maximum risk of the $(f_i^\star, h_i^\star)$ (see (8)) and show that it is equal to the RHS of (22). First note that we can write,

$$\begin{pmatrix}\hat\mu(X_i)-\mu \\ \hat\mu(Z_i)-\mu\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\frac{\sigma^2}{|X_i|} & 0 \\ 0 & \frac{\sigma^2}{|Z_i|}\end{pmatrix}\right).$$

By a linear transformation of this Gaussian vector, we obtain

$$\begin{pmatrix}\frac{|X_i|}{\sigma^2}(\hat\mu(X_i)-\mu)+\frac{|Z_i|}{\sigma^2}(\hat\mu(Z_i)-\mu) \\ \hat\mu(X_i)-\hat\mu(Z_i)\end{pmatrix} = \begin{pmatrix}\frac{|X_i|}{\sigma^2} & \frac{|Z_i|}{\sigma^2} \\ 1 & -1\end{pmatrix}\begin{pmatrix}\hat\mu(X_i)-\mu \\ \hat\mu(Z_i)-\mu\end{pmatrix}$$

$$\sim \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\frac{|X_i|+|Z_i|}{\sigma^2} & 0 \\ 0 & \frac{\sigma^2}{|X_i|}+\frac{\sigma^2}{|Z_i|}\end{pmatrix}\right),$$

18

626 which means $\frac{|X_i|}{\sigma^2}(\widehat{\mu}(X_i) - \mu) + \frac{|Z_i|}{\sigma^2}(\widehat{\mu}(Z_i) - \mu)$ and $\frac{\eta_i}{\alpha} = \widehat{\mu}(X_i) - \widehat{\mu}(Z_i)$ are independent Gaussian
627 random variables. Therefore, the the maximum risk of $(f_i^\star, h_i^\star)$ is:

$$\sup_{\mu} \mathbb{E}\big[(h_i^\star(X_i, Y_i, A_i) - \mu)^2|\mu\big] = \sup_{\mu} \mathbb{E}_{\eta_i}\left[\mathbb{E}\left[\left(\frac{\frac{\sum_{x \in Z_i'} x}{\sigma^2 + \eta_i^2} + \frac{|X_i|}{\sigma^2}\widehat{\mu}(X_i) + \frac{|Z_i|}{\sigma^2}\widehat{\mu}(Z_i)}{\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2}} - \mu\right)^2 \bigg| \eta_i\right]\right]$$

$$= \sup_{\mu} \mathbb{E}_{\eta_i}\left[\mathbb{E}\left[\left(\frac{\frac{\sum_{x \in Z_i'}(x - \mu)}{\sigma^2 + \eta_i^2} + \frac{|X_i|}{\sigma^2}(\widehat{\mu}(X_i) - \mu) + \frac{|Z_i|}{\sigma^2}(\widehat{\mu}(Z_i) - \mu)}{\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2}}\right)^2 \bigg| \eta_i\right]\right]$$

$$= \sup_{\mu} \mathbb{E}_{\eta_i}\left[\frac{\mathbb{E}\left[\left(\frac{\sum_{x \in Z_i'}(x - \mu)}{\sigma^2 + \eta_i^2} + \frac{|X_i|}{\sigma^2}(\widehat{\mu}(X_i) - \mu) + \frac{|Z_i|}{\sigma^2}(\widehat{\mu}(Z_i) - \mu)\right)^2 \bigg| \eta_i\right]}{\left(\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2}\right)^2}\right]$$

$$= \sup_{\mu} \mathbb{E}_{\eta_i}\left[\frac{1}{\left(\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2}\right)^2}\left(\frac{|Z_i'|(\sigma^2 + \eta_i^2)}{(\sigma^2 + \eta_i^2)^2} + \frac{|X_i| + |Z_i|}{\sigma^2}\right)\right]$$

$$= \mathbb{E}_{\eta_i}\left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \eta_i^2} + \frac{|X_i| + |Z_i|}{\sigma^2}}\right] = \mathbb{E}\left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \alpha^2(\widehat{\mu}(Z_i) - \widehat{\mu}(X_i))^2} + \frac{|X_i| + |Z_i|}{\sigma^2}}\right]$$

628 Because $\widehat{\mu}(Z_i) - \widehat{\mu}(X_i) \sim \mathcal{N}\left(0, \frac{\sigma^2}{|X_i|} + \frac{\sigma^2}{|Z_i|}\right)$, we can further write the maximum risk as:

$$\sup_{\mu} \mathbb{E}\big[(h_i^\star(X_i, Y_i, A_i) - \mu)^2|\mu\big] = \mathbb{E}_{e \sim \mathcal{N}(0,1)}\left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \alpha^2\left(\frac{\sigma^2}{|Z_i|} + \frac{\sigma^2}{|X_i|}\right)e^2} + \frac{|X_i| + |Z_i|}{\sigma^2}}\right] = R_\infty$$

629 Here, we have observed that the final expression in the above equation is exactly the same as the
630 Bayes' risk in the limit in (22) from Step 1.

631 **Step 3: Minimax optimality of $(f_i^\star, h_i^\star)$:** As the maximum is larger than the average, we can write,
632 for any prior Λ_ℓ, and any $(f_i, h_i) \in \mathcal{F} \times \mathcal{H}$,

$$\sup_{\mu} \mathbb{E}\big[(h_i(X_i, f_i(X_i), A_i) - \mu)^2|\mu\big] \geq \mathbb{E}_{\Lambda_\ell}\big[\mathbb{E}\big[(h_i(X_i, f_i(X_i), A_i) - \mu)^2|\mu\big]\big] \geq R_\ell.$$

633 As this is true for all ℓ, by taking the limit we have, for all $(f_i, h_i) \in \mathcal{F} \times \mathcal{H}$,

$$\sup_{\mu} \mathbb{E}\big[(h_i(X_i, f_i(X_i), A_i) - \mu)^2|\mu\big] \geq R_\infty = \sup_{\mu} \mathbb{E}\big[(h_i^\star(X_i, f_i^\star(X_i), A_i) - \mu)^2|\mu\big].$$

634 That is, the recommended $(f_i^\star, h_i^\star)$ has a smaller maximum risk than all other $(f_i, h_i) \in \mathcal{F} \times \mathcal{H}$. This
635 establishes that for any n_i,

$$p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)) = \inf_{f_i \in \mathcal{A}} \inf_{h_i \in \mathcal{H}} p_i(M_{\text{C3D}}, ((n_i, f_i, h_i), s_{-i}^\star)).$$

636 ### A.3.2 Proof of Lemma 6

637 In the previous section, we showed that for any n_i, the optimal (f_i, h_i) were $(f_i^\star, h_i^\star)$ as given in (8).
638 Now, we show that for the given $(f_i^\star, h_i^\star)$, the optimal number of samples is $n_i^\star = \sigma/\sqrt{cm}$. For this,
639 we will show that p_i is a convex function of n_i and then show that its gradient is 0 at $n_i^\star$.

640 First, noting that

$$\widehat{\mu}(Z_i) - \widehat{\mu}(X_i) \sim \mathcal{N}\left(0, \frac{\sigma^2}{|X_i|} + \frac{\sigma^2}{|Z_i|}\right),$$

19

we can rewrite the penalty term as:

$$p(n_i) := p_i\big(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)\big) = \mathbb{E}\left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \alpha^2(\widehat{\mu}(Z_i) - \widehat{\mu}(X_i))^2} + \frac{|X_i| + |Z_i|}{\sigma^2}}\right] + cn_i$$

$$= \mathbb{E}_{x \sim \mathcal{N}(0,1)}\left[\frac{1}{\frac{|Z_i'|}{\sigma^2 + \alpha^2\left(\frac{\sigma^2}{|X_i|} + \frac{\sigma^2}{|Z_i|}\right)x^2} + \frac{|X_i| + |Z_i|}{\sigma^2}}\right] + cn_i$$

$$= \mathbb{E}_{x \sim \mathcal{N}(0,1)}\Bigg[\underbrace{\frac{1}{\frac{(m-2)n_i^\star}{\sigma^2 + \alpha^2\left(\frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_i^\star}\right)x^2} + \frac{n_i + n_i^\star}{\sigma^2}}}_{=:l(n_i, x; \alpha)}\Bigg] + cn_i \tag{23}$$

Convexity of penalty function: To show that $p(n_i)$ is convex in n_i, let us consider $l(n_i, x; \alpha)$. Fixing α and x, we have

$$\frac{\partial}{\partial n_i} l(n_i, x; \alpha) = -\sigma^2 \frac{1 + \frac{(m-2)n_i^\star}{\left(1 + \alpha^2\left(\frac{1}{n_i} + \frac{1}{n_i^\star}\right)x^2\right)^2} \frac{\alpha^2 x^2}{n_i^2}}{\left(\frac{(m-2)n_i^\star}{1 + \alpha^2\left(\frac{1}{n_i} + \frac{1}{n_i^\star}\right)x^2} + n_i + n_i^\star\right)^2} = -\sigma^2 \frac{1 + \frac{(m-2)n_i^\star \alpha^2 x^2}{\left(n_i + \alpha^2\left(1 + \frac{n_i}{n_i^\star}\right)x^2\right)^2}}{\left(\frac{(m-2)n_i^\star}{1 + \alpha^2\left(\frac{1}{n_i} + \frac{1}{n_i^\star}\right)x^2} + n_i + n_i^\star\right)^2}$$

As $\frac{\partial}{\partial n_i} l(n_i, x; \alpha)$ is an increasing function of n_i, we have that $l(n_i, x; \alpha)$ is a convex function in n_i. As expectation preserves convexity (see Lemma 10), $p(n_i)$ is a convex function.

Penalty is minimized when $n_i = n_i^\star$. Lemma 13 provides an expression for the derivative of $p(n_i)$ (obtained purely via algebraic manipulations). Using this, we have

$$p'(n_i^\star) = -\frac{\sigma^2}{64\frac{\alpha^2}{m-2}\frac{\alpha}{\sqrt{mn_i^\star}} mn_i^\star}\left(\frac{4\alpha}{\sqrt{mn_i^\star}}\left(\frac{4\alpha^2 m}{(m-2)n_i^\star} - 1\right)\right.$$

$$\left. - \exp\left(\frac{mn_i^\star}{8\alpha^2}\right)\left(\frac{4\alpha^2}{mn_i^\star}(m+1) - 1\right)\sqrt{2\pi}\,\text{Erfc}\left(\frac{1}{2\sqrt{2}\sqrt{\frac{\alpha^2}{mn_i^\star}}}\right)\right)$$

$$+ c \hspace{4cm} \text{(By Lemma 13)}$$

$$= -\frac{\sigma^2}{64\frac{\alpha^2}{m-2}\frac{\alpha}{\sqrt{mn_i^\star}} mn_i^\star}\left(\frac{4\alpha}{\sqrt{mn_i^\star}}\left(\frac{4\alpha^2(m-4)}{(m-2)n_i^\star} - 1\right)\right.$$

$$\left. - \exp\left(\frac{mn_i^\star}{8\alpha^2}\right)\left(\frac{4\alpha^2}{mn_i^\star}(m+1) - 1\right)\sqrt{2\pi}\,\text{Erfc}\left(\frac{1}{2\sqrt{2}\sqrt{\frac{\alpha^2}{mn_i^\star}}}\right)\right)$$

$$= G(\alpha) = 0.$$

Here, the second step uses the fact that $n_i^\star = \frac{\sigma}{\sqrt{cm}}$. Finally, we have observed that the expression is equal to $G(\alpha)$ as defined in (7) which is 0 by our choice of α. Since $p'(n_i^\star) = 0$ and $p(\cdot)$ is convex, we can conclude that $p(n_i)$ is minimized when $n_i = n_i^\star$. Therefore,

$$p_i\big(M_{\text{C3D}}, ((n_i^\star, f_i^\star, h_i^\star), s_{-i}^\star)\big) \le p_i\big(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)\big).$$

20

651 **A.4 Algorithm 1 is individually rational**

As outlined in the main text, the IC property implies IR since 'working on her own' is a valid strategy in the mechanism. Precisely, if an agent collects any number of points n_i, chooses not to submit anything $f_i(\cdot) = \varnothing$, and then uses the sample average of the points she collected $h_i(X_i, \varnothing, A_i) = |X_i|^{-1} \sum_{x \in X_i} x$, then $(n_i, f_i, h_i) \in \mathcal{S}$.

Below, we will prove this more formally and also show that the agent's penalty is strictly smaller when participating. For any fixed n_i, without participating in the mechanism, the smallest penalty the agent can achieve is by using empirical mean estimation and the penalty is:

$$\frac{\sigma^2}{n_i} + cn_i$$

When participating, the agent gets an additional $n_i^\star$ number of clean data along with some noisy data, provided that all other agents are following $s_{-i}^\star$. By using the empirical mean over the clean data, the penalty is:

$$\frac{\sigma^2}{n_i + n_i^\star} + cn_i < \frac{\sigma^2}{n_i} + cn_i$$

Now, since the weighted average estimator in $s_i^\star$ is minimax optimal, the agent gets even smaller maximum risk and hence smaller penalty. In other words, for any n_i,

$$p_i(M_{\text{C3D}}, s^\star) \leq p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)) \leq \frac{\sigma^2}{n_i + n_i^\star} + cn_i < \frac{\sigma^2}{n_i} + cn_i$$

By minimizing the RHS with respect to n_i, we get $p_i(M_{\text{C3D}}, s^\star) < p_{\min}^{\text{IR}}$. Thus Algorithm 1 is IR.

A.5 Algorithm 1 is approximately efficient

In this section, we will bound the penalty ratio PR for M_{C3D} at the strategy profiles $s_i^\star$.

First, noting that $G(\alpha) = 0$ (see (7)), we can rearrange the terms in the equation to obtain:

$$\exp\left(\frac{mn_i^\star}{8\alpha^2}\right) \text{Erfc}\left(\frac{1}{2\sqrt{2}\sqrt{\frac{\alpha^2}{mn_i^\star}}}\right) = \frac{1}{\sqrt{2\pi}} \frac{\frac{4\alpha}{\sqrt{mn_i^\star}}\left(\frac{4\alpha^2(m-4)}{(m-2)n_i^\star} - 1\right)}{\frac{4\alpha^2}{mn_i^\star}(m+1) - 1} \tag{24}$$

Next, we will use the expression for $p(n_i) = p_i(M_{\text{C3D}}, (s_{-i}^\star, (n_i, f_i^\star, h_i^\star)))$ in Lemma 13 and the equation in (24) to simplify $p(n_i^\star)$ as follows:

$$p(n_i^\star) = \frac{\sqrt{\frac{\alpha^2}{mn_i^\star}}\sigma^2\left(2m\sqrt{2\pi}\sqrt{\frac{\alpha^2}{mn_i^\star}} - \exp\left(\frac{mn_i^\star}{8\alpha^2}\right)(m-2)\pi \,\text{Erfc}\left(\frac{1}{2\sqrt{2}\sqrt{\frac{\alpha^2}{mn_i^\star}}}\right)\right)}{4\sqrt{2\pi}\alpha^2} + cn_i^\star$$

(By Lemma 13)

$$= \frac{\sqrt{\frac{\alpha^2}{mn_i^\star}}\sigma^2\left(2m\sqrt{2\pi}\sqrt{\frac{\alpha^2}{mn_i^\star}} - (m-2)\pi\frac{1}{\sqrt{2\pi}}\frac{\frac{4\alpha}{\sqrt{mn_i^\star}}\left(\frac{4\alpha^2(m-4)}{(m-2)n_i^\star} - 1\right)}{\frac{4\alpha^2}{mn_i^\star}(m+1) - 1}\right)}{4\sqrt{2\pi}\alpha^2} + cn_i^\star \quad \text{(By (24))}$$

$$= \frac{\sigma^2\left(m - (m-2)\frac{\frac{4\alpha^2(m-4)}{(m-2)n_i^\star} - 1}{\frac{4\alpha^2}{mn_i^\star}(m+1) - 1}\right)}{2mn_i^\star} + cn_i^\star$$

$$= \frac{\sigma^2}{2mn_i^\star}\frac{\frac{4\alpha^2}{n_i^\star}(m+1) - m - \frac{4\alpha^2}{n_i^\star}(m-4) + (m-2)}{\frac{4\alpha^2}{n_i^\star}\frac{m+1}{m} - 1} + cn_i^\star$$

$$= \frac{\sigma^2}{2mn_i^\star}\frac{\frac{20\alpha^2}{n_i^\star} - 2}{\frac{4\alpha^2}{n_i^\star}\frac{m+1}{m} - 1} + cn_i^\star = \frac{\sigma^2}{mn_i^\star}\frac{\frac{10\alpha^2}{n_i^\star} - 1}{\frac{4\alpha^2}{n_i^\star}\frac{m+1}{m} - 1} + cn_i^\star$$

21

Algorithm 2 M_{PCS}

1: **Mechanism designer publishes:**
2:     The allocation space $\mathcal{A} = \bigcup_{n \geq 0} \mathbb{R}^n$, and the procedure in lines 6 –11.
3: **Each agent i:**
4:     Choose strategy $s_i = (n_i, f_i, h_i)$.
5:     Sample n_i points $X_i = \{x_{i,j}\}_{j=1}^{n_i}$ and submit $Y_i = f_i(X_i)$ to the mechanism.
6: **Mechanism:**
7:     For each agent $i \in [m]$:   # can be done simultaneously for all agents
8:         $A_i \leftarrow \bigcup_{j \neq i} Y_j$ if $|Y_i| \geq \sigma/\sqrt{cm}$, $A_i \leftarrow \varnothing$ otherwise.
9:         Return A_i to each agent.
10: **Each agent i:**
11:     Compute estimate $h_i(X_i, Y_i, A_i)$.

$$= \sigma \sqrt{\frac{c}{m}} \left(\frac{\frac{10\alpha^2}{n_i^\star} - 1}{\frac{4\alpha^2}{n_i^\star} \frac{m+1}{m} - 1} + 1 \right)$$

From our conclusion in §A.2, we have $\alpha^2 > \frac{\sigma}{\sqrt{cm}} = n_i^\star$, i.e. $\frac{\alpha^2}{n_i^\star} > 1$. Therefore, we have:

$$\text{PR}(M_{\text{C3D}}, s^\star) = \frac{mp(n_i^\star)}{2\sigma\sqrt{cm}} = \frac{1}{2}\left(\frac{\frac{10\alpha^2}{n_i^\star} - 1}{\frac{4\alpha^2}{n_i^\star}\frac{m+1}{m} - 1} + 1 \right)$$

$$< \frac{1}{2}\left(\frac{\frac{10\alpha^2}{n_i^\star} - 1 + \frac{10\alpha^2}{n_i^\star}\frac{1}{m} + \left(\frac{2\alpha^2}{n_i^\star}\frac{m+1}{m} - 2\right)}{\frac{4\alpha^2}{n_i^\star}\frac{m+1}{m} - 1} + 1 \right) = 2.$$

B Proof of Theorem 2

We will use M_{PCS} to denote the mechanism in §4.1, as it *pools* the datsets, but *checks* for the *size* of the dataset submitted by each agent. For clarity, we have stated M_{PCS} algorithmically in Algorithm 2. We will also re-state the recommended strategies $s_i^\star = \{(n_i^\star, f_i^\star, h_i^\star)\}_i$ below:

$$n_i^\star = \frac{\sigma}{\sqrt{cm}}, \qquad f_i^\star = \mathbf{I}, \qquad h_i^\star(X_i, Y_i, A_i) = \frac{1}{|X_i \cup A_i|} \sum_{u \in X_i \cup A_i} u \qquad (25)$$

Throughout this section, $s_i^\star$ will refer to (25) (and not (8)).

We will first prove that $s_i^\star$ is a Nash equilibrium. Because the sample mean achieves minimax error for Normal mean estimation [23], we immediately have, for all $(n_i, f_i, h_i) \in \mathcal{S}$.

$$p_i(M_{\text{PCS}}, ((n_i, f_i, h_i^\star), s_{-i}^\star)) \leq p_i(M_{\text{PCS}}, ((n_i, f_i, h_i), s_{-i}^\star)).$$

Because the agent can only submit the raw dataset or a subset, and the agent's allocation only depends on the size of the dataset, the size of the dataset she receives can always be maximized by submittng the whole data set she collects, i.e. chooses $f_i = \mathbf{I}$. Therefore, we have for all $(n_i, f_i, h_i) \in \mathcal{S}$,

$$p_i(M_{\text{PCS}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)) \leq p_i(M_{\text{PCS}}, ((n_i, f_i, h_i^\star), s_{-i}^\star)) \leq p_i(M_{\text{PCS}}, ((n_i, f_i, h_i), s_{-i}^\star)).$$

Finally, we can use the fact that the maximum risk of the sample mean estimator using n points is σ^2/n to show that the penalty is minimized when $n_i = n_i^\star = \sigma/\sqrt{cm}$. In particular, we have that if $n_i < \sigma/\sqrt{cm}$,

$$p_i(M_{\text{PCS}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)) = \frac{\sigma^2}{n_i} + cn_i > 2\sigma\sqrt{c}.$$

And if $n_i \geq \sigma/\sqrt{cm}$,

$$p_i(M_{\text{PCS}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star)) = \frac{\sigma^2}{n_i + (m-1)\sigma/\sqrt{cm}} + cn_i \geq 2\sigma\sqrt{\frac{c}{m}}$$

22

685 Because $2\sigma\sqrt{c} \geq 2\sigma\sqrt{c/m}$, $p_i(M_{\text{PCS}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star))$ is minimized when $n_i = \sigma/\sqrt{cm}$. We
686 thus conclude that $s^\star$ is a Nash equilibrium. That is, for all $(n_i, f_i, h_i) \in \mathbb{N} \times \mathcal{F} \times \mathcal{H}$

$$p_i(M_{\text{PCS}}, s^\star) \leq p_i(M_{\text{PCS}}, ((n_i, f_i, h_i), s_{-i}^\star)).$$

687 Next, the IR and efficiency properties follow trivially from the fact that $p_i(M_{\text{PCS}}, s^\star) = 2\sigma\sqrt{c/m}$ for
688 each agent i. In particular, $p_i(M_{\text{PCS}}, s^\star) < p_{\min}^{\text{IR}}$ and $P(M_{\text{PCS}}, s^\star) = 2\sigma\sqrt{cm}$.

C Proof of Theorem 3

690 We will use M_{CDED} to denote our mechanism in §4.2, as it *corrupts* the *deployed estimate* based on
691 the *difference*. We have stated this mechanism formally in Algorithm 3. We will also re-state the
692 recommended strategies $s_i^\star = \{(n_i^\star, f_i^\star)\}_i$ below:

$$n_i^\star = \frac{\sigma}{\sqrt{cm}}, \qquad f_i^\star = \mathbf{I}. \tag{26}$$

693 Throughout this section, $s_i^\star$ will refer to (26) (and not (8) or (25)).

694 We will now present the proof of Theorem 3. First, in §C.1, we show that $s^\star$ is a Nash equilibrium of
695 M_{CDED} as the incentive compatibility result. Then, in §C.2, we show individual rationality at $s_i^\star$. In
696 §C.3, we conclude by showing that M_{CDED} is approximately efficient by showing that its social penalty
697 at most a $(1 + \epsilon)$ factor of the global minimum.

C.1 Algorithm 3 is incentive compatible

699 **Step 1.** We will first show that fixing any n_i, the best strategy is to submit the raw data, i.e. for all
700 $(n_i, f_i) \in \mathbb{N} \times \mathcal{F}$.

$$p_i(M_{\text{CDED}}, ((n_i, f_i^\star), s_{-i}^\star)) \leq p_i(M_{\text{CDED}}, ((n_i, f_i), s_{-i}^\star)). \tag{27}$$

701 Let $e_{z,i} = \epsilon_{z,i}/\eta_i$, where η_i, and $\epsilon_{z,i}$ are as given in lines 9 and 10 respectively. We have that $e_{z,i}$'s
702 are i.i.d. standard Normal samples. Because the cost term cn_i is fixed when n_i is fixed, we only need
703 to consider the risk term. We will first define,

$$\widehat{\mu}(X_i) := \frac{1}{|X_i|}\sum_{x \in X_i} x, \quad \widehat{\mu}(Y_i) := \frac{1}{|Y_i|}\sum_{x \in Y_i} x, \quad \widehat{\mu}(Y_{-i}) := \frac{1}{|Y_{-i}|}\sum_{x \in Y_{-i}} x. \tag{28}$$

704 Via some algebraic manipulations, we can express the maximum risk as:

$$\sup_\mu \mathbb{E}\left[\left(\frac{1}{|Y_i| + (m-1)n_i^\star}\left(\sum_{y \in Y_i}(y - \mu) + \sum_{z \in Y_{-i}}(z + e_{z,i}\eta_i - \mu)\right)\right)^2 \middle| \mu\right]$$

$$= \frac{1}{(|Y_i| + (m-1)n_i^\star)^2}\sup_\mu \mathbb{E}\left[\left(\sum_{y \in Y_i}(y - \mu)\right)^2 + \left(\sum_{z \in Y_{-i}}(z + e_{z,i}\eta_i - \mu)\right)^2 \middle| \mu\right]$$

$$= \frac{1}{(|Y_i| + (m-1)n_i^\star)^2}\sup_\mu \mathbb{E}\left[|Y_i|\left(\widehat{\mu}(Y_i) - \mu\right)^2 + \left(\sum_{z \in Y_{-i}}(z - \mu)\right)^2 + \left(\sum_{z \in Y_{-i}}e_{z,i}\eta_i\right)^2 \middle| \mu\right]$$

$$= \frac{1}{(|Y_i| + (m-1)n_i^\star)^2}\sup_\mu \mathbb{E}\left[|Y_i|\left(\widehat{\mu}(Y_i) - \mu\right)^2 + (m-1)n_i^\star\beta_\epsilon^2(\widehat{\mu}(Y_i) - \widehat{\mu}(Y_{-i}))^{2k_\epsilon} \middle| \mu\right]$$

$$+ \frac{(m-1)n_i^\star\sigma^2}{(|Y_i| + (m-1)n_i^\star)^2}$$

705 Recall that β_ϵ also involves $|Y_i|$. Note that as we have fixed n_i and $s_{-i} = s_{-i}^\star$, the maximum risk
706 depends only on $|Y_i|$ and $\widehat{\mu}(Y_i)$, that is, the agent's maximum risk and hence penalty only depends on
707 the number of points she submitted, and their average value. Hence, to find the optimal submission

23

Algorithm 3 M_{CDED}

Require: Approximation parameter $\epsilon > 0$ # to obtain a $1 + \epsilon$ bound on PR.
 1: **Mechanism designer publishes:**   The procedure in lines 5 –11.
 2: **Each agent i:**
 3:       Choose strategy $s_i = (n_i, f_i)$.
 4:       Sample n_i points $X_i = \{x_{i,j}\}_{j=1}^{n_i}$ and submit $Y_i = f_i(X_i)$ to the mechanism.
 5: **Mechanism:**
 6:       $k_\epsilon \leftarrow \min\{\lceil\frac{1}{2\epsilon}\rceil, 2\}, \quad \beta_\epsilon \leftarrow \sqrt{\dfrac{\left(\sum_{i=1}^m |Y_i|\right)^2 (m-1)^{k_\epsilon - 1}}{k_\epsilon (2k_\epsilon - 1)!! \sigma^{k_\epsilon} c^{\frac{k_\epsilon - 2}{2}} m^{3k_\epsilon/2}}}$
 7:       **For** each agent $i \in [m]$:   # can be done simultaneously for all agents
 8:             $Y_{-i} \leftarrow \bigcup_{j \neq i} Y_j$.
 9:             $\eta_i^2 \leftarrow \beta_\epsilon^2 \left(\frac{1}{|Y_i|} \sum_{y \in Y_i} y - \frac{1}{|Y_{-i}|} \sum_{y \in Y_{-i}} y\right)^{2k_\epsilon}$.
10:             $Z_i \leftarrow \{z + \epsilon_{z,i}, \text{ for all } z \in Y_{-i} \text{ where } \epsilon_{z,i} \sim \mathcal{N}(0, \eta_i^2)\}$
11:             Deploy estimate $\left(\frac{1}{|Y_i \cup Z_i|} \sum_{u \in Y_i \cup Z_i} u\right)$ for agent i.

Y_i, we will first fix the size of the agent's submission $|Y_i|$ and optimize for the sample mean $\widehat{\mu}(Y_i)$ (step 1.1), and then we will optimize for $|Y_i|$ (step 1.2).

Step 1.1. Since the other agents have each collected $\sigma/\sqrt{cm} = n_i^\star$ points and submitted it truthfully, we have $\widehat{\mu}(Y_{-i}) \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{(m-1)n_i^\star}\right)$. Via a binomial expansion , we can write,

$$\mathbb{E}\left[(\widehat{\mu}(Y_i) - \widehat{\mu}(Y_{-i}))^{2k_\epsilon}\right] = \mathbb{E}\left[((\widehat{\mu}(Y_i) - \mu) - (\widehat{\mu}(Y_{-i}) - \mu))^{2k_\epsilon}\right]$$

$$= \sum_{j=0}^{2k_\epsilon} (-1)^j \binom{2k_\epsilon}{j} \mathbb{E}\left[(\widehat{\mu}(Y_i) - \mu)^j\right] \mathbb{E}\left[(\widehat{\mu}(Y_{-i}) - \mu)^{2k_\epsilon - j}\right]$$

$$= \sum_{j=0}^{k_\epsilon} \binom{2k_\epsilon}{2j} \mathbb{E}\left[(\widehat{\mu}(Y_i) - \mu)^{2j}\right] \mathbb{E}\left[(\widehat{\mu}(Y_{-i}) - \mu)^{2k_\epsilon - 2j}\right]$$

Thus the maximum risk can be written as:

$$\sup_\mu \mathbb{E}\left[\sum_{j=0}^{k_\epsilon} A_j (\widehat{\mu}(Y_i) - \mu)^{2j} \,\middle|\, \mu\right] \tag{29}$$

where $A_0, \ldots, A_{k_\epsilon}$ is a sequence of positive coefficients.

Similar to the proof of Theorem 1, we construct a lower bound on the maximum risk using a sequence of Bayesian risks. Let $\Lambda_\ell := \mathcal{N}(0, \ell^2)$, $\ell = 1, 2, \ldots$ be a sequence of prior for μ. For fixed ℓ, the posterior distribution is:

$$p(\mu|X_i) \propto p(X_i|\mu)p(\mu) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{x \in X_i} (x - \mu)^2\right) \exp\left(-\frac{1}{2\ell^2}\mu^2\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{n_i}{\sigma^2} + \frac{1}{\ell^2}\right)\mu^2 + \frac{1}{2}2\frac{\sum_{x \in X_i} x}{\sigma^2}\mu\right).$$

This means the posterior of μ given X_i is Gaussian with:

$$\mu|X_i \sim \mathcal{N}\left(\frac{n_i\widehat{\mu}(X_i)/\sigma^2}{n_i/\sigma^2 + 1/\ell^2}, \frac{1}{n_i/\sigma^2 + 1/\ell^2}\right) =: \mathcal{N}(\mu_\ell, \sigma_\ell^2).$$

Therefore, the posterior risk is:

$$\mathbb{E}\left[\sum_{j=0}^{k_\epsilon} A_j (\widehat{\mu}(Y_i) - \mu)^{2j} \,\middle|\, X_i\right] = \mathbb{E}\left[\sum_{j=0}^{k_\epsilon} A_j ((\widehat{\mu}(Y_i) - \mu_\ell) - (\mu - \mu_\ell))^{2j} \,\middle|\, X_i\right]$$

24

$$= \int_{-\infty}^{\infty} \underbrace{\sum_{j=0}^{k_\epsilon} A_j (e - (\widehat{\mu}(Y_i) - \mu_\ell))^{2j}}_{=:F_1(e-(\widehat{\mu}(Y_i)-\mu_\ell))} \underbrace{\frac{1}{\sigma_\ell \sqrt{2\pi}} \exp\left(-\frac{e^2}{2\sigma_\ell^2}\right)}_{=:F_2(e)} de$$

719 Because:

720   • $F_1(\cdot)$ is even function and increases on $[0, \infty)$;

721   • $F_2(\cdot)$ is even function and decreases on $[0, \infty$, and $\int_{\mathbb{R}} F_2(e) de < \infty$

722   • For any $a \in \mathbb{R}$, $\int_{\mathbb{R}} F_1(e-a) F_2(e) de < \infty$,

723 By the corollary of Hardy-Littlewood inequality in Lemma 9,

$$\int_{\mathbb{R}} F_1(e-a) F_2(e) de \geq \int_{\mathbb{R}} F_1(e) F_2(e) de,$$

724 which means the posterior risk is minimized when $\widehat{\mu}(Y_i) = \mu_\ell$. In Lemma 11, we have stated
725 expressions for the expected value of the power of a normal random variable. Using this, we can
726 write the Bayes risk as:

$$R_\ell := \mathbb{E}\left[\sum_{j=0}^{k_\epsilon} A_j \mathbb{E}\left[(\mu - \mu_\ell)^{2j} \Big| X_i\right]\right] = \sum_{j=0}^{k_\epsilon} A_j (2j-1)!! \sigma_\ell^{2j}$$

727 and the limit of Bayesian risk as $\ell \to \infty$ is

$$R_\infty := \lim_{\ell \to \infty} \sum_{j=0}^{k_\epsilon} A_j (2j-1)!! \frac{\sigma^{2j}}{n_i^j}$$

728 When $\widehat{\mu}(Y_i) = \widehat{\mu}(X_i)$, the maximum risk is:

$$\sup_\mu \mathbb{E}\left[\sum_{j=0}^{k_\epsilon} A_j (\widehat{\mu}(Y_i) - \mu)^{2j} \Big| \mu\right] = \sup_\mu \mathbb{E}\left[\sum_{j=0}^{k_\epsilon} A_j (\widehat{\mu}(X_i) - \mu)^{2j} \Big| \mu\right]$$

$$= \sum_{j=0}^{k_\epsilon} A_j (2j-1)!! \sigma^{2j} n_i^{-j} = R_\infty.$$

729 This means, fixing n_i and $|Y_i|$, agent i achieves minimax risk when choosing $\widehat{\mu}(Y_i) = \widehat{\mu}(X_i)$; as the
730 maximum is larger than the average, this follows using a similar argument to Step 3 in §A.3.

731 **Step 1.2.** Next, we will show that the best size of the submission is $|Y_i| = |X_i| = n_i$, assuming
732 $\widehat{\mu}(Y_i) = \widehat{\mu}(X_i)$. For this, we will first use $n_i^\star$ to rewrite β_ϵ^2 as

$$\beta_\epsilon^2 = \frac{n_i^{\star k_\epsilon - 2}(m-1)^{k_\epsilon - 1}(|Y_i| + (m-1)n_i^\star)^2}{k_\epsilon (2k_\epsilon - 1)!! m^{k_\epsilon + 1} \sigma^{2k_\epsilon - 2}}.$$

733 Because

$$\widehat{\mu}(X_i) - \widehat{\mu}(Y_{-i}) \sim \mathcal{N}\left(0, \left(\frac{1}{n_i} + \frac{1}{(m-1)n_i^\star}\right)\sigma^2\right),$$

734 the risk term in the penalty can be rewritten and lower bounded as follows:

$$\frac{1}{(|Y_i| + (m-1)n_i^\star)^2}\left(|Y_i|^2 \sigma^2/n_i + (m-1)n_i^\star \beta_\epsilon^2 (2k_\epsilon - 1)!! \left(\frac{1}{n_i} + \frac{1}{(m-1)n_i^\star}\right)^{k_\epsilon} \sigma^{2k_\epsilon}\right)$$

$$+ \frac{(m-1)n_i^\star \sigma^2}{(|Y_i| + (m-1)n_i^\star)^2}$$

$$= \frac{|Y_i|^2 \frac{\sigma^2}{n_i} + (m-1)n_i^\star \sigma^2}{(|Y_i| + (m-1)n_i^\star)^2} + \frac{n_i^{\star k_\epsilon - 1}(m-1)^{k_\epsilon}}{k_\epsilon m^{k_\epsilon + 1}}\left(\frac{1}{n_i} + \frac{1}{(m-1)n_i^\star}\right)^{k_\epsilon} \sigma^2$$

25

$$\geq \frac{\sigma^2}{n_i + (m-1)n_i^\star} + \frac{n_i^{\star k_\epsilon - 1}(m-1)^{k_\epsilon}}{k_\epsilon m^{k_\epsilon + 1}}\left(\frac{1}{n_i} + \frac{1}{(m-1)n_i^\star}\right)^{k_\epsilon}\sigma^2.$$

Here, the last step follows from the fact that

$$\frac{|Y_i|^2\frac{\sigma^2}{n_i} + (m-1)n_i^\star\sigma^2}{(|Y_i| + (m-1)n_i^\star)^2} = \frac{|Y_i|^2\frac{\sigma^2}{n_i} + (m-1)n_i^\star\sigma^2}{n_i\frac{|Y_i|^2}{n_i} + 2|Y_i|(m-1)n_i^\star + (m-1)^2 n_i^{\star 2}}$$

$$\geq \frac{|Y_i|^2\frac{\sigma^2}{n_i} + (m-1)n_i^\star\sigma^2}{n_i\frac{|Y_i|^2}{n_i} + \left(n_i + \frac{|Y_i|^2}{n_i}\right)(m-1)n_i^\star + (m-1)^2 n_i^{\star 2}} = \frac{|Y_i|^2\frac{\sigma^2}{n_i} + (m-1)n_i^\star\sigma^2}{(n_i + (m-1)n_i^\star)\left(\frac{|Y_i|^2}{n_i} + (m-1)n_i^\star\right)}$$

$$= \frac{\sigma^2}{n_i + (m-1)n_i^\star}.$$

Equality holds in this inequality if and only if $|Y_i| = n_i$.

In conclusion, fixing n_i, the agent can minimize her penalty by submitting n_i points with the same sample mean as the dataset X_i she collected. One way to achieve this is set $f_i = \mathbf{I}$. This completes the proof of (27).

Step 2: Our next step is to show that the agent's best strategy is to collect $n_i^\star$ data points. That is, we will show for all $n_i \in \mathbb{N}$.

$$p_i(M_{\text{CDED}}, ((n_i^\star, f_i^\star), s_{-i}^\star)) \leq p_i(M_{\text{CDED}}, ((n_i, f_i^\star), s_{-i}^\star)). \tag{30}$$

In the following, we will use $p(n_i)$ as a shorthand for $p_i(M_{\text{CDED}}, ((n_i, f_i^\star), s_{-i}^\star))$. The penalty can be rewritten as:

$$p(n_i) = \frac{\sigma^2}{n_i + (m-1)n_i^\star} + \frac{n_i^{\star k_\epsilon - 1}(m-1)^{k_\epsilon}}{k_\epsilon m^{k_\epsilon + 1}}\left(\frac{1}{n_i} + \frac{1}{(m-1)n_i^\star}\right)^{k_\epsilon}\sigma^2 + cn_i$$

We need to show that $p_i(n_i)$ achieves minimum at $n_i = n_i^\star$. The derivative of $p_i(\cdot)$ is:

$$p'(n_i) = -\frac{\sigma^2}{(n_i + (m-1)n_i^\star)^2} + \frac{n_i^{\star k_\epsilon - 1}(m-1)^{k_\epsilon}}{m^{k_\epsilon + 1}}\left(\frac{1}{n_i} + \frac{1}{(m-1)n_i^\star}\right)^{k_\epsilon - 1}\sigma^2\left(-\frac{1}{n_i^2}\right) + c$$

Because $p'(n_i)$ increase in n_i, $p(n_i)$ is convex. Moreover, because

$$p'(n_i^\star) = -\frac{\sigma^2}{m^2 n_i^{\star 2}} + \frac{n_i^{\star k_\epsilon - 1}(m-1)^{k_\epsilon}}{m^{k_\epsilon + 1}}\left(\frac{1}{n_i^\star} + \frac{1}{(m-1)n_i^\star}\right)^{k_\epsilon - 1}\sigma^2\left(-\frac{1}{n_i^{\star 2}}\right) + c$$

$$= -\frac{\sigma^2}{m^2 n_i^{\star 2}} - \frac{(m-1)\sigma^2}{m^2 n_i^{\star 2}} + c = -\frac{\sigma^2}{m n_i^{\star 2}} + c = 0,$$

we know $p(n_i)$ reaches minimum at $n_i = n_i^\star$. This concludes the proof for (30).

C.2   Algorithm 3 is individually rational

The penalty of an agent at the recommended strategies can be expressed as:

$$p_i(M_{\text{CDED}}, s_i^\star) = p(n_i^\star) = \frac{\sigma^2}{m n_i^\star} + \frac{n_i^{\star k_\epsilon - 1}(m-1)^{k_\epsilon}}{k_\epsilon m^{k_\epsilon + 1}}\left(\frac{1}{n_i^\star} + \frac{1}{(m-1)n_i^\star}\right)^{k_\epsilon}\sigma^2 + cn_i^\star$$

$$= \frac{\sigma^2}{m n_i^\star} + \frac{n_i^{\star k_\epsilon - 1}(m-1)^{k_\epsilon}}{k_\epsilon m^{k_\epsilon + 1}}\frac{m^{k_\epsilon}}{n_i^{\star k_\epsilon}(m-1)^{k_\epsilon}}\sigma^2 + cn_i^\star$$

$$= \frac{\sigma^2}{m n_i^\star} + \frac{1}{k_\epsilon}\frac{\sigma^2}{m n_i^\star} + cn_i^\star = \left(2 + \frac{1}{k_\epsilon}\right)\frac{\sigma\sqrt{c}}{\sqrt{m}}. \tag{31}$$

We have that M_{CDED} is IR when $m \geq 2$, via the following simple calculation:

$$\left(2 + \frac{1}{k_\epsilon}\right)\frac{\sigma\sqrt{c}}{\sqrt{m}} \leq \left(2 + \frac{1}{2}\right)\frac{\sigma\sqrt{c}}{\sqrt{2}} < 2\sigma\sqrt{c} = p_{\min}^{\text{IR}}$$

26

C.3 **Algorithm 3 is approximately efficient**

751 Using the expression for $p_i(M_{\text{CDED}}, s_i^\star)$ in (31), the penalty ratio can be bounded by:

$$\text{PR}(M_{\text{CDED}}, s^\star) = \frac{\left(2 + \frac{1}{k_\epsilon}\right)\sigma\sqrt{cm}}{2\sigma\sqrt{cm}} = 1 + \frac{1}{2k_\epsilon} \leq 1 + \epsilon.$$

D Additional Materials for Section 4.2

D.1 Mechanism detail

754 See Algorithm 3.

D.2 Using a weighted average under the original strategy space from §2

756 In this section, we will consider a variation of M_{CDED} when applied to our original strategy space
757 $\mathbb{N} \times \mathcal{F} \times \mathcal{H}$. For this, we will assume that M_{CDED} will return $A_i = Z_i$ as the agent's allocation, and
758 then an agent can use X_i, Y_i, Z_i to estimate μ. In this situation, below we show that the agent can
759 achieve a smaller penalty using a weighted average over $X_i \cup Z_i$ instead of the sample mean used
760 by the mechanism. Here, the weights are proportional to the inverse of the variance of each data
761 point. (Our mechanism purposefully uses the sub-optimal sample mean in the restricted strategy
762 space $\mathbb{N} \times \mathcal{F}$ as a way to shape the agent's penalty and incentivize good behavior.)

763 This shows that M_{CDED} (with the above modification) is not IC in this more general strategy space.
764 The agent can obtain a lower penalty using a better estimator (such as the weighted average we show
765 over here) and achieve a lower penalty. More importantly, as the agent knows that she can achieve a
766 lower estimation error via a better estimator instead of more data, she can leverage this insight to
767 collect less data and reduce her penalty even further.

768 We should emphasize that it is unclear if this weighted average is minimax optimal. It is also unclear
769 if there exists a Nash equilibrium for M_{CDED} (or any straightforward modification of M_{CDED}) in the
770 expanded strategy space.

771 **The weighted average estimator:** We will now present the weighted average estimator that
772 achieves a lower maximum risk. To show this, first note that for all $x \in X_i$, $\mathbb{V}[x] = \sigma^2$; when
773 $(n_i, f_i) = (n_i^\star, f_i^\star)$, for all $x \in Z_i$,

$$
\begin{aligned}
\mathbb{V}[x] &= \mathbb{E}\left[(z + \epsilon_{z,i} - \mu)^2\right] = \sigma^2 + \beta_\epsilon^2 \mathbb{E}\left[(\widehat{\mu}(X_i) - \widehat{\mu}(Y_{-i}))^{2k_\epsilon}\right] \\
&= \sigma^2 + \frac{n_i^{\star k_\epsilon - 2}(m-1)^{k_\epsilon - 1}(mn_i^\star)^2}{k_\epsilon(2k_\epsilon - 1)!!m^{k_\epsilon + 1}\sigma^{2k_\epsilon - 2}}(2k_\epsilon - 1)!!\left(\frac{1}{n_i^\star} + \frac{1}{(m-1)n_i^\star}\right)^{k_\epsilon}\sigma^{2k_\epsilon} \\
&= \sigma^2 + \frac{n_i^{\star k_\epsilon}(m-1)^{k_\epsilon - 1}}{k_\epsilon m^{k_\epsilon - 1}}\frac{m^{k_\epsilon}}{(m-1)^{k_\epsilon}n_i^{\star k_\epsilon}}\sigma^2 \\
&= \sigma^2 + \frac{1}{k_\epsilon}\frac{m}{m-1}\sigma^2
\end{aligned}
$$

774 Consider the following weighted-average estimator:

$$
h_i(X_i, Y_i, (Z_i, \eta_i^2)) = \frac{\frac{1}{\sigma^2}\sum_{x \in X_i} x + \frac{1}{\sigma^2 + \frac{1}{k_\epsilon}\frac{m}{m-1}\sigma^2}\sum_{x \in Z_i} x}{\frac{n_i^\star}{\sigma^2} + \frac{(m-1)n_i^\star}{\sigma^2 + \frac{1}{k_\epsilon}\frac{m}{m-1}\sigma^2}}
$$

775 The maximum risk of h_i is

$$
\begin{aligned}
\mathbb{E}\left[\left(h_i(X_i, Y_i, (Z_i, \eta_i^2)) - \mu\right)^2\right] &= \frac{1}{\frac{n_i^\star}{\sigma^2} + \frac{(m-1)n_i^\star}{\sigma^2 + \frac{1}{k_\epsilon}\frac{m}{m-1}\sigma^2}} = \frac{1}{1 + \frac{m-1}{1 + \frac{1}{k_\epsilon}\frac{m}{m-1}}}\frac{\sigma^2}{n_i^\star} = \frac{1 + \frac{1}{k_\epsilon}\frac{m}{m-1}}{m + \frac{1}{k_\epsilon}\frac{m}{m-1}}\frac{\sigma^2}{n_i^\star} \\
&< \frac{\left(1 + \frac{1}{k_\epsilon}\right)\left(1 + \frac{1}{k_\epsilon}\frac{1}{m-1}\right)}{m + \frac{1}{k_\epsilon}\frac{m}{m-1}}\frac{\sigma^2}{n_i^\star} = \left(1 + \frac{1}{k_\epsilon}\right)\frac{\sigma^2}{mn_i^\star} \quad (32)
\end{aligned}
$$

27

776 Note that the RHS of (32) is the risk of the sample average deployed by M_{CDED}. This means, suppose
777 all other agents choose $s^\star$, then agent i can choose a weighted average to reduce her penalty without
778 collecting more data.

E High dimensional mean estimation with bounded variance

780 In this section, we will study estimating a d–dimensional mean $\mu(\theta) \in \mathbb{R}^d$ for distributions θ with
781 bounded variance. We will focus on our original setting in §2, but will outline the modifications to
782 the formalism to accommodate the generality. For $x \in \mathbb{R}^d$, let $x^{(i)}$ denote the ith dimension.

783 **Modifications to the setting in §2:** First, we should change the definitions of $\mathcal{F}, \mathcal{H}$ and $\mathcal{M}$ in
784 equations 1 and (2) to account for the fact that the data is d dimensional. For instance, the space
785 of functions mapping the dataset collected to the dataset submitted should be defined as $\mathcal{F} = \{f :$
786 $\bigcup_{n \geq 0} \mathbb{R}^{d \times n} \to \bigcup_{n \geq 0} \mathbb{R}^{d \times n}\}$. Next, let $\Theta = \{\theta; \ \mathbf{supp}\,(\theta) \subset \mathbb{R}^d, \ \mathbb{E}_{x \sim \theta}\left[(x^{(i)} - \mu(\theta)^{(i)})^2\right] \leq$
787 $\sigma^2, \ \forall i \in [d]\}$ be the class of all d–dimensional distributions where the variance along each dimension
788 is bounded by σ^2. Here, the maximum variance σ^2 is known and is public information. Note that
789 we do not assume that the individual dimensions are independent. An agent's penalty p_i is defined
790 similar to (3) but considers the maximum risk over Θ, i.e

$$p_i(M, s) = \sup_{\theta \in \Theta} \mathbb{E}\left[\|h_i(X_i, Y_i, A_i) - \mu(\theta)\|_2^2 \ \big| \ \theta\right] \ + \ cn_i. \tag{33}$$

791 Finally, the social penalty and ratio PR are as defined in (5), but with the above definition for p_i.

792 **Mechanism:** Our mechanism for this problem is the same as the one outlined in Algorithm 1, with
793 the following cosmetic modifications. First, the allocation space should now be $\mathcal{A} = \bigcup_{n \geq 0} \mathbb{R}^{d \times n} \times$
794 $\bigcup_{n \geq 0} \mathbb{R}^{d \times n} \times \mathbb{R}_+^d$. The noise modulating parameter α is determined by a similar equation as in (7),
795 but with c replaced with c/d. In line 12 of Algorithm 1, we should set the size of the dataset Z_i to
796 be $\min\{|Y_{-i}|, \sigma\sqrt{d/(cm)}\}$. Finally, the operations in lines 13 and 14 should be interpreted as d–
797 dimensional operations that are performed elementwise. The recommended strategy $s_i^\star = (n_i^\star, f_i^\star, h_i^\star)$
798 for agent i is as follows:

$$n_i^\star = \begin{cases} \frac{\sigma}{m}\sqrt{\frac{d}{c}} & \text{if } m \leq 4, \\ \sigma\sqrt{\frac{d}{cm}} & \text{if } m \geq 5 \end{cases}, \qquad\qquad f_i^\star = \mathbf{I}, \tag{34}$$

$$h_i^\star(X_i, Y_i, (Z_i, Z_i', \eta_i^2)) = \frac{\frac{1}{\sigma^2}\sum_{u \in X_i \cup Z_i} u + \frac{1}{\sigma^2 + \tau_i^2}\sum_{u \in Z_i'} u}{\frac{1}{\sigma^2}|X_i \cup Z_i'| + \frac{1}{\sigma^2 + \tau_i^2}|Z_i'|}, \qquad \text{where, } \tau_i^2 = \frac{2\alpha^2\sigma^2}{n_i^\star} \in \mathbb{R}_+.$$

799 Above, one difference worth highlighting is the change in the recommended estimator $h_i^\star$. Previously,
800 the weighting used the η_i^2 term returned by the mechanism, which is a function of Y_i and Z_i. This
801 data-dependent weighting was necessary to obtain an *exactly* (i.e including constants) minimax
802 optimal estimator for the corrupted dataset, which in turn was necessary to achieve an exact Nash
803 equilibrium. However, bounding the risk when using a data-dependent weighting is challenging
804 when the Gaussian assumption does not hold. Instead, here we use a deterministic weighting via the
805 quantity τ_i^2. While this is not exactly minimax optimal, we can show that its maximum risk is very
806 close to a lower bound, which helps us obtain an approximate Nash equilibrium. It is worth pointing
807 out that designing exactly minimax optimal estimators, even under i.i.d assumptions, is challenging
808 for general classes of distributions [23].

809 The following theorem states the main properties of this mechanism.

810 **Theorem 7.** *The following statements are true about the mechanism M_{C3D} in Algorithm 1 with the*
811 *above modifications. (i) The strategy profile $s^\star$ as defined in (34) is an approximate Nash equilibrium,*
812 *i.e if all agents except i are following $s^\star$, then for any alternative strategy s_i for agent i, we have*
813 *$p_i(M_{\text{C3D}}, s^\star) \leq p_i(M_{\text{C3D}}, (s_{-i}^\star, s_i))(1 + 5/m)$ (ii) The mechanism is individually rational at $s^\star$. (iii)*
814 *The mechanism is approximately efficient at $s_i^\star$, with $\text{PR}(M_{\text{C3D}}, s^\star) < 2 + 10/m$.*

815 We see that even under this more general setting, our mechanism retains its main properties with only
816 a slight weakening of the results. We now have approximate, instead of exact, IC, with the benefit of
817 deviation diminishing as there are more agents. Similarly, the bound on the efficiency is only slightly
818 weaker than the one in Theorem 1.

28

819**E.1   Proof of Theorem 7**

820   When $m \leq 4$, the claims follow using the exact steps in §A.1. Therefore, we focus on the case $m \geq 5$.
821   Moreover, some of the key steps of this proof follows along similar lines to Theorem 1, so we will
822   provide an outline and focus on the differences.

823   **Approximate incentive compatibility.** We will first prove the statement *(i)* of Theorem 7, which
824   states that $s_i^\star$, as defined in (34), is an approximate Nash equilibrium for M_{C3D}. That is, we will
825   show that the maximum possible reduction in penalty for an agent i when deviating from $s_i^\star$ is small,
826   provided that all other agents are following $s_{-i}^\star$.

827   For this, we will first lower bound the penalty p_i (33) using the family of independent Gaussian distri-
828   butions. Let $\Theta_{\mathcal{N}} = \{\mathcal{N}(\mu, \sigma^2 I_d) : \mu \in \mathbb{R}^d\}$ denote the space of d–dimensional normal distributions
829   with identity covariance matrix. For any mechanism M and strategy profile $s \in \mathcal{S}^m$, we define the
830   penalty of agent i restricted to $\Theta_{\mathcal{N}}$ as:

$$p_i^{\mathcal{N}}(M, s) = \sup_{\theta \in \Theta_{\mathcal{N}}} \mathbb{E}\left[\|h_i(X_i, Y_i, A_i) - \mu(\theta)\|_2^2 \mid \theta\right] + cn_i.$$

831   Since $\Theta_{\mathcal{N}} \subset \Theta$, it is straightforward to see that for all $M \in \mathcal{M}$ and $s \in \mathcal{S}^m$,

$$p_i^{\mathcal{N}}(M, s) \leq p_i(M, s). \tag{35}$$

832   We will now use this result to lower bound the penalty of an agent for any other alternative strategy.
833   First note that, by independence, the mean estimation problem on $\Theta_{\mathcal{N}}$ can be viewed as d independent
834   copies of the univariate normal mean estimation problem considered in Theorem 1 but with c replaced
835   with c/d. Let $\tilde{h}_i^\star$ be the weighted average that applies the estimator in (8) along each dimension.
836   And let $\tilde{s}_i^\star = (n_i^\star, f_i^\star, \tilde{h}_i^\star)$. We can now lower bound the penalty of agent i when following any
837   (alternative) strategy $s_i \in \mathcal{S}$, provided that other agents are following $s_{-i}^\star$. We have:

$$
\begin{aligned}
p_i(M_{\text{C3D}}, (s_i, s_{-i}^\star)) &= p_i\big(M_{\text{C3D}}, \big(s_i, (n_{-i}^\star, f_{-i}^\star, h_{-i}^\star)\big)\big) \\
&\geq p_i^{\mathcal{N}}\big(M_{\text{C3D}}, \big(s_i, (n_{-i}^\star, f_{-i}^\star, h_{-i}^\star)\big)\big) && \text{(By (35))} \\
&= p_i^{\mathcal{N}}\Big(M_{\text{C3D}}, \Big(s_i, \big(n_{-i}^\star, f_{-i}^\star, \tilde{h}_{-i}^\star\big)\Big)\Big) \\
&\qquad\text{(As agent i's penalty will not be affected by other agents' estimators)} \\
&\geq p_i^{\mathcal{N}}\Big(M_{\text{C3D}}, \Big(\big(n_i^\star, f_i^\star, \tilde{h}_i^\star\big), \big(n_{-i}^\star, f_{-i}^\star, \tilde{h}_{-i}^\star\big)\Big)\Big) \\
&\qquad(\text{ By adapting the analysis in §A.3. }) \\
&= p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star) \tag{36}
\end{aligned}
$$

838   Above, the second step uses (35) and the third step uses the fact that other agent's *estimator* will
839   not affect agent i's penalty. The fourth step uses the fact that for estimation problems in $\Theta_{\mathcal{N}}$, the
840   strategy profile $\tilde{s}^\star = \{(n_i^\star, f_i^\star, \tilde{h}_i^\star)\}_i$ is a Nash equilibrium; in §A.3, we showed this for the one
841   dimensional case, but this proof can be easily adapted to d dimensions since we are assuming an
842   identity covariance matrix in $\Theta_{\mathcal{N}}$. Finally, by adapting the analysis in §A.5, we can obtain the
843   following expression for agent i's penalty $p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star)$ in $\Theta_{\mathcal{N}}$:

$$p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star) = d\sigma\sqrt{\frac{c/d}{m}}\left(\frac{\frac{10\alpha^2}{n_i^\star} - 1}{\frac{4\alpha^2}{n_i^\star}\frac{m+1}{m} - 1} + 1\right) \tag{37}$$

844   To state the approximate IC result, we will now upper bound the penalty of the agent when following
845   $s_i^\star$. Using the bounded variance assumption, we have:

$$p_i(M, s^\star) = \sup_{\theta \in \Theta} \mathbb{E}\left[\left\|\frac{\frac{1}{\sigma^2}\sum_{u \in X_i \cup Z_i} u + \frac{1}{\sigma^2 + \tau_i^2}\sum_{u \in Z_i'} u}{\frac{1}{\sigma^2}|X_i \cup Z_i'| + \frac{1}{\sigma^2 + \tau_i^2}|Z_i'|} - \mu(\theta)\right\|_2^2 \middle| \theta\right] + cn_i^\star$$

$$= \sup_{\theta \in \Theta}\sum_{k=1}^d \mathbb{E}\left[\left(\frac{\frac{1}{\sigma^2}\sum_{u \in X_i \cup Z_i}\big(u^{(k)} - \mu(\theta)^{(k)}\big) + \frac{1}{\sigma^2 + \tau_i^2}\sum_{u \in Z_i'}\big(u^{(k)} - \mu(\theta)^{(k)}\big)}{\frac{1}{\sigma^2}|X_i \cup Z_i'| + \frac{1}{\sigma^2 + \tau_i^2}|Z_i'|}\right)^2 \middle| \theta\right] + cn_i^\star$$

29
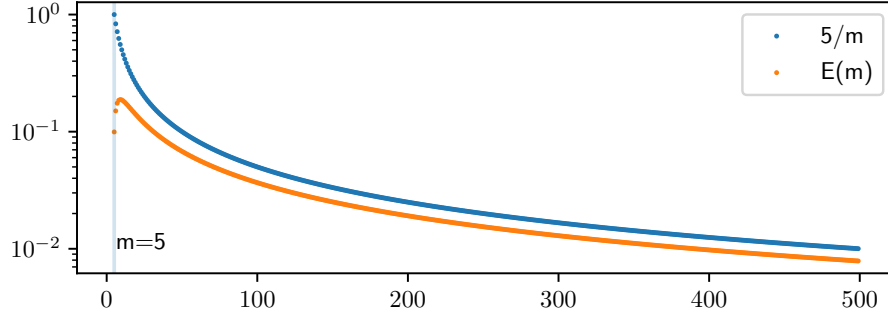
Figure 2: $E(m)$ plot. See `G_em_plot.py`.

$$= \sup_{\theta \in \Theta} \sum_{k=1}^{d} \frac{\frac{1}{\sigma^2} \sum_{u \in X_i \cup Z_i} \mathbb{E}\left[\left(u^{(k)} - \mu(\theta)^{(k)}\right)\right] + \frac{1}{\sigma^2 + \tau_i^2} \sum_{u \in Z_i'} \mathbb{E}\left[\left(u^{(k)} - \mu(\theta)^{(k)}\right)\right]}{\frac{1}{\sigma^2}|X_i \cup Z_i'| + \frac{1}{\sigma^2 + \tau_i^2}|Z_i'|} + cn_i^\star \quad (38)$$

$$\leq \frac{d}{\frac{2n_i^\star}{\sigma^2} + \frac{(m-2)n_i^\star}{\sigma^2 + \frac{2\alpha^2\sigma^2}{n_i^\star}}} + cn_i^\star = \frac{\sigma^2}{n_i^\star} \frac{d}{2 + \frac{m-2}{1 + \frac{2\alpha^2}{n_i^\star}}} + cn_i^\star = \sigma\sqrt{\frac{cd}{m}} \left(\frac{m}{2 + \frac{m-2}{1 + \frac{2\alpha^2}{n_i^\star}}} + 1 \right), \quad (39)$$

where (38) is because: for all $k \in [d]$, $\forall x_1^{(k)}, x_2^{(k)} \in X_i \cup Z_i$, $\forall z_1^{(k)}, z_2^{(k)} \in Z_i'$, $x_1^{(k)} - \mu^{(k)}, x_2^{(k)} - \mu^{(k)}, z_1^{(k)} - \mu^{(k)}, z_2^{(k)} - \mu^{(k)}$ are uncorrelated pairwise. The final inequality is due to the bounded variance assumption.

Next, for brevity, let us write $A_m := \frac{\alpha}{\sqrt{n_i^\star}}$ where α is as defined in (7). By adapting the analysis in §A.2, we can show that

$$A_m := \frac{\alpha}{\sqrt{n_i^\star}} \in \left(1, 1 + \frac{C_m}{m}\right), \qquad \text{where,} \quad C_m = \begin{cases} 20, & \text{if } m \leq 20 \\ 5, & \text{if } m > 20 \end{cases}. \quad (40)$$

By combining the results in (36), (39), and (40), we obtain the following bound:

$$\frac{p_i(M_{\text{C3D}}, s^\star)}{\inf_{s_i} p_i(M_{\text{C3D}}, (s_i, s_{-i}^\star))} - 1 \leq \frac{p_i(M_{\text{C3D}}, s^\star)}{p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star)} - 1$$

$$\leq \frac{\sigma\sqrt{\frac{cd}{m}} \left(\frac{m}{2 + \frac{m-2}{1 + 2A_m^2}} + 1 \right)}{d\sigma\sqrt{\frac{c/d}{m}} \left(\frac{10A_m^2 - 1}{4A_m^2 \frac{m+1}{m} - 1} + 1 \right)} - 1 = \frac{\frac{m}{2 + \frac{m-2}{1 + \frac{2\alpha^2}{n_i^\star}}} + 1}{\frac{\frac{10\alpha^2}{n_i^\star} - 1}{\frac{4\alpha^2}{n_i^\star} \frac{m+1}{m} - 1} + 1} - 1$$

$$= \frac{4A_m^2\left((A_m^2 - 1)m + 1 - 4A_m^2\right)m}{(4A_m^2 + m)((7A_m^2 - 1)m + 2A_m^2)} =: E(m). \quad (41)$$

Let $E(m)$ denote the final upper bound obtained above. Next, we will prove $E(m) < 5/m$. When $m \in [5, 500]$, this can be individually verified for each value of $E(m)$ (see Figure 2). When $m \geq 500$, we have $A_m \leq 1.01$ (see (40)). From this we can conclude,

$$E(m) \leq \frac{4 \times 1.01^2 \times (2.01 \times \frac{5}{m}m - 3)m}{6m^2} < \frac{5}{m}. \quad (42)$$

Combining the results in (41) and (42), we obtain the following approximate IC result:

$$\forall i \in [m], \ s_i \in \mathcal{S}, \qquad p_i(M_{\text{C3D}}, s^\star) \leq p_i(M_{\text{C3D}}, (s_i, s_{-i}^\star)) \left(1 + \frac{5}{m}\right).$$

Individual rationality: This proof is very similar to the proof in §A.4. In particular, using calculations similar to (39), we can show that regardless of the choice of n_i, the agent's penalty is strictly smaller when using the uncorrupted (Z_i) and corrupted (Z_i') datasets along with the weighted average in (34).

30

859 **Approximate efficiency:** To bound the penalty ratio, first note that by (37) and using the same
860 reasoning as §A.5, we have that

$$\frac{\sum_i p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star)}{\inf_{M \in \mathcal{M}, s \in \mathcal{S}^m} \sum_i p_i^{\mathcal{N}}(M, s)} = \frac{m p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star)}{\inf_{M \in \mathcal{M}, s \in \mathcal{S}^m} \sum_i p_i^{\mathcal{N}}(M, s)} = \frac{m p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star)}{2\sigma\sqrt{cmd}} \le 2. \quad (43)$$

861 Next, as $\Theta_{\mathcal{N}} \subset \Theta$, and noting that $P(M, s) = \sum_i p_i(M, s)$ for all M, s, we can also write,

$$\inf_{M \in \mathcal{M}, s \in \mathcal{S}^m} \sum_i p_i^{\mathcal{N}}(M, s) \le \inf_{M \in \mathcal{M}, s \in \mathcal{S}^m} P(M, s). \quad (44)$$

862 We can combine the above results to obtain the following upper bound on PR:

$$\begin{aligned}
\text{PR}(M_{\text{C3D}}, s^\star) = \frac{P(M_{\text{C3D}}, s^\star)}{\inf_{M \in \mathcal{M}, s \in \mathcal{S}^m} P(M, s)} &\le \frac{m p_i(M_{\text{C3D}}, s^\star)}{\inf_{M \in \mathcal{M}, s \in \mathcal{S}^m} \sum_i p_i^{\mathcal{N}}(M, s)} && \text{(By (44))} \\
&= \frac{m p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star)}{\inf_{M \in \mathcal{M}, s \in \mathcal{S}^m} \sum_i p_i^{\mathcal{N}}(M, s)} \frac{p_i(M_{\text{C3D}}, s^\star)}{p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star)} && \\
&\le 2 \frac{p_i(M_{\text{C3D}}, s^\star)}{p_i^{\mathcal{N}}(M_{\text{C3D}}, \tilde{s}^\star)} && \text{(By (43))} \\
&= 2(1 + E(m))) && \text{(By definition of } E(m)\text{, see (41))} \\
&< 2 + \frac{10}{m}. && \text{(By (42))}
\end{aligned}$$

863 This establishes approximate efficiency for M_{C3D} for the high dimensional setting.

F   Useful Results

865 In this section, we will state some useful results that we have used throughout this proof.

866 **Lemma 8** (Hardy-Littlewood inequality, Lemma 1.6 in Burchard [10]). *Let f and g be non-negative*
867 *measurable functions that vanish at infinity. Let f^* and g^* to denote the symmetric decreasing*
868 *rearrangement of f and g. If $\int f^* g^* < \infty$, then,*

$$\int fg \le \int f^* g^*.$$

869 Next, we will use the above result to derive a corollary that will be useful in our proofs.

870 **Lemma 9** (A corollary of Hardy-Littlewood). *Let f, g be nonnegative even functions such that,*

871   • *f is monotonically increasing on $[0, \infty)$.*

872   • *g is monotonically decreasing on $[0, \infty)$, and has a finite integral $\int_{\mathbb{R}} g(x)dx < \infty$.*

873   • *$\forall a$, $\int_{\mathbb{R}} f(x - a)g(x)dx < \infty$.*

874 *Then for all a,*

$$\int_{\mathbb{R}} f(x)g(x)dx \le \int_{\mathbb{R}} f(x - a)g(x)dx$$

875 *Proof.* We will break this proof into two cases. The first is when $\sup f < \infty$ and the second is when
876 $\sup f = \infty$. First consider the case $\sup f < \infty$. Let

$$M := \lim_{x \to \infty} f(x).$$

877 By using Lemma 8, $\forall a$,

$$\int_{\mathbb{R}} (M - f(x))g(x)dx \ge \int_{\mathbb{R}} (M - f(x - a))g(x)dx.$$

878 The result follows after rearrangement.

31

879 If $\sup f = \infty$, let $f_n(x) := \min\{f(x), n\}$. For all n and a, by Lemma 8,

$$\int_{\mathbb{R}} (n - f_n(x))g(x)dx \geq \int_{\mathbb{R}} (n - f_n(x - a))g(x)dx,$$

880 thus

$$\int_{\mathbb{R}} f_n(x)g(x)dx \leq \int_{\mathbb{R}} f_n(x - a)g(x)dx.$$

881 Note that $|f_n(x)g(x)| \leq f(x)g(x)$, the result follows by letting $n \to \infty$ on both sides and using
882 dominated convergence theorem. □

883 Below, we provide a brief example on using Lemma 9 to calculate the Bayes risk in a normal mean
884 estimation problem with i.i.d data. While it is not necessary to use Hardy-Littlewood for this problem,
885 this example will illustrate how we have used it in our proofs.

886 **Example 1.** *Consider the Normal mean estimation problem given samples* $X_{[n]} \sim \mathcal{N}(\mu, \sigma^2)$*, where*
887 μ *admits a prior distribution* $\mathcal{N}(0, \ell^2)$*. The goal is to minimize the average risk:*

$$\mathbb{E}_{\mu \sim \mathcal{N}(0,\ell^2)}\Big[\mathbb{E}_{X_{[n]} \sim \mathcal{N}(\mu,\sigma^2)}[L(\hat{\mu} - \mu)|\mu]\Big],$$

888 *where the loss function,* $L(\cdot)$*, is an even function that increases on* $[0, \infty)$*. By a standard argument,*
889 *one can show that the posterior distribution of* μ *conditioned on* $X_{[n]}$ *is Gaussian with data-dependent*
890 *parameters* $\bar{\mu}, \bar{\sigma}^2$*:*

$$\mu|X_{[n]} \sim \mathcal{N}(\overline{\mu}, \overline{\sigma}^2).$$

891 *The posterior risk is:*

$$\mathbb{E}_{\mu|X_{[n]}}[L(\hat{\mu} - \mu)] = \mathbb{E}_{\mu|X_{[n]}}[L((\mu - \overline{\mu}) + (\overline{\mu} - \hat{\mu}))] = \int_{\mathbb{R}} \underbrace{L(x + (\overline{\mu} - \hat{\mu}))}_{=:f(x+(\overline{\mu}-\hat{\mu}))} \underbrace{\frac{\exp\left(-\frac{x^2}{2\overline{\sigma}^2}\right)}{\overline{\sigma}\sqrt{2\pi}}}_{=:g(x)} dx$$

892 *By applying Lemma 9 with* f *and* g*, the posterior risk above is minimized when* $\hat{\mu} = \overline{\mu}$*. So is the*
893 *average risk.*

894 The next Lemma shows that convexity is preserved under expectation under certain conditions.

895 **Lemma 10.** *Let* y *be a random variable and* $f(x, y)$ *be a function s.t.*

896   • $f(x, y)$ *is convex in* x*;*

897   • $\mathbb{E}_y[|f(x, y)|] < \infty$ *for all* x*.*

898 *Then* $\mathbb{E}_y[f(x, y)]$ *is also convex in* x*.*

899 *Proof.* For any x_1, x_2, we have

$$\frac{\mathbb{E}_y[f(x_1, y)] + \mathbb{E}_y[f(x_2, y)]}{2} = \mathbb{E}_y\left[\frac{f(x_1, y) + f(x_2, y)}{2}\right] \geq \mathbb{E}_y\left[f\left(\frac{x_1 + x_2}{2}, y\right)\right]$$

900 □

901 **Lemma 11** (Centered moments of normal random variable)**.** *Let* $X \sim \mathcal{N}(\mu, \sigma^2)$ *be a normal random*
902 *variable and* $p \in \mathbb{Z}_+$*, then*

$$\mathbb{E}[(X - \mu)^p] = \begin{cases} 0 & \text{if } p \text{ is odd} \\ \sigma^p(p-1)!! & \text{if } p \text{ is even} \end{cases}.$$

32

F.1 Some technical results

Next, we will state some technical results that were obtained purely using algebraic manipulations and are not central to the main proof ideas. The first result states upper and lower bounds on the Gaussian complementary error function using an asymptotic expansion.

Lemma 12 (Erfc bound). *For all* $x > 0$,

$$\mathrm{Erfc}(x) \leq \frac{1}{\sqrt{\pi}}\left(\frac{\exp(-x^2)}{x} - \frac{\exp(-x^2)}{2x^3} + \frac{3\exp(-x^2)}{4x^5}\right) \tag{45}$$

$$\mathrm{Erfc}(x) \geq \frac{1}{\sqrt{\pi}}\left(\frac{\exp(-x^2)}{x} - \frac{\exp(-x^2)}{2x^3}\right) \tag{46}$$

Proof. By integration by parts:

$$\frac{\sqrt{\pi}}{2}\mathrm{Erfc}(x) = \int_x^\infty \exp(-t^2)dt = \left(-\frac{\exp(-t^2)}{2t}\right)\Big|_x^\infty - \int_x^\infty \frac{\exp(-t^2)}{2t^2}dt$$

$$= \frac{\exp(-x^2)}{2x} - \left(\left(-\frac{\exp(-t^2)}{4t^3}\right)\Big|_x^\infty - \int_x^\infty \frac{3\exp(-t^2)}{4t^4}dt\right)$$

$$= \frac{\exp(-x^2)}{2x} - \frac{\exp(-x^2)}{4x^3} + \underbrace{\int_x^\infty \frac{3\exp(-t^2)}{4t^4}dt}_{\geq 0} \tag{47}$$

$$= \frac{\exp(-x^2)}{2x} - \frac{\exp(-x^2)}{4x^3} + \left(-\frac{3\exp(-t^2)}{8t^5}\right)\Big|_x^\infty - \int_x^\infty \frac{15\exp(-t^2)}{8t^6}dt$$

$$= \frac{\exp(-x^2)}{2x} - \frac{\exp(-x^2)}{4x^3} + \frac{3\exp(-x^2)}{8x^5} \underbrace{- \int_x^\infty \frac{15\exp(-t^2)}{8t^6}dt}_{\leq 0} \tag{48}$$

The results follow by (47) and (48). $\qquad\square$

Our next result, which is obtained using Mathematica, states an expression for the function $p(n_i)$ and its derivative as defined in (23).

Lemma 13 (Value and derivative of penalty function at $s^\star$). *Let* $p(n_i) = p_i(M_{\text{C3D}}, ((n_i, f_i^\star, h_i^\star), s_{-i}^\star))$ *(see (23)) and* $s_i^\star, f_i^\star, h_i^\star$ *be as specified in (8). The penalty of agent* i *in Algorithm 1 satisfies:*

$$p(n_i^\star) = \frac{\sqrt{\frac{\alpha^2}{mn_i^\star}}\sigma^2\left(2m\sqrt{2\pi}\sqrt{\frac{\alpha^2}{mn_i^\star}} - \exp\left(\frac{mn_i^\star}{8\alpha^2}\right)(m-2)\pi\,\mathrm{Erfc}\left(\frac{1}{2\sqrt{2}\sqrt{\frac{\alpha^2}{mn_i^\star}}}\right)\right)}{4\sqrt{2\pi}\alpha^2} + cn_i^\star \tag{49}$$

$$p'(n_i^\star) = -\frac{\sigma^2}{64\frac{\alpha^2}{m-2}\frac{\alpha}{\sqrt{mn_i^\star}}mn_i^\star}\left(\frac{4\alpha}{\sqrt{mn_i^\star}}\left(\frac{4\alpha^2 m}{(m-2)n_i^\star} - 1\right)\right.$$

$$\left. - \exp\left(\frac{mn_i^\star}{8\alpha^2}\right)\left(\frac{4\alpha^2}{mn_i^\star}(m+1) - 1\right)\sqrt{2\pi}\,\mathrm{Erfc}\left(\frac{1}{2\sqrt{2}\sqrt{\frac{\alpha^2}{mn_i^\star}}}\right)\right) + c. \tag{50}$$

Proof. We derived these expressions using Mathematica and have given our source code in `thm1pf_penalty.nb` and `thm1pf_penalty.pdf`. But, we will also outline how these results can be derived manually.

First, after rearranging the denominator inside the expectation in (23), we can write the RHS as $a + b\mathbb{E}[\frac{1}{d+x^2}]$, where the expectation is with respect to a standard normal $\mathcal{N}(0,1)$ variable, a, b, d are quantities that depend on $n_i, m, c, \sigma^2, \alpha^2$, and importantly, d is strictly larger than 0. Using properties of the chi-squared distribution, we have

$$\mathbb{E}\left[\frac{1}{d+x^2}\right] = \sqrt{\frac{\pi}{2d}}e^{d/2}\,\mathrm{Erfc}\left(\frac{d}{2}\right).$$

33

By plugging in these expressions and then substituting $n_i = n_i^\star$, we obtain (49). Then (50) is obtained by taking the derivative of (49) (with $n_i^\star$ replaced with n_i) with respect to n_i. $\qquad\square$

34