Eidgenössische Technische Hochschule Züerich

Master thesis

Master in Mathematics

Incentive compatible mechanisms for robust estimators

Author:
Eliott MEA

Supervisors:
Yonatan SOMPOLINSKY
Roger WATTENHOFER

Abstract

In statistics, one of the straightforward goals of an estimator is to minimize error. However, the presence of outliers often challenges the performance of estimators. This is where the notion of robustness appears. Robust estimators are designed to withstand outliers or malicious observations. However, their performance is lower bounded by limits on error. This paper explores the potential of incentive mechanisms to foster accurate reporting and improved estimation.

We introduce a new estimator, called "Local Density Estimator " (LDE), which is a member of the density estimation methods family. We explore how it functions well within the asymptotic continuous model.We devise a mechanism for the allocation of a budget W to the participating oracles, that exhibits desirable properties, including robustness and efficiency. We prove these properties within both endogenous and exogenous incentive frameworks.

The significance of this research is especially important in finance, where the reliability of oracles and data feeds is crucial, and where there are clear incentives to misreport and manipulate (e.g., through MEV techniques; see Daian (2022)). Our findings underscore the importance of well-designed incentive mechanisms in enhancing the accuracy and reliability of financial data and estimation processes.

Contents

List of Figures

Chapter 1

Introduction

1.0.1 Estimators in Finance and Other Critical Systems

In crypto, reliable external data feeds are crucial as they ensure the accuracy and integrity of price feeds. The price feeds are foundational for financial systems. Indeed, smart contracts often handle financial transactions and require off-chain data such as exchange rates or commodity prices. For instance, stablecoins depend on regular updates of exchange rates to maintain their value by attaching the token to an external currency's value.

Entities providing data feeds are known as *oracles*, while the mechanism that selects the feed is called the *estimator*. Oracles supply external data to blockchain systems, and estimators determine how this data is processed and integrated. Accurate and reliable oracles are essential for maintaining trust in the information used by smart contracts and decentralized applications. However, as oracles cannot be blindly trusted, estimators are needed to ensure the overall quality and reliability of the data feed.

One major challenge is limiting the influence of intentional deviations in data feeds. Ensuring accuracy involves detecting and mitigating manipulations that could skew the data. This is critical to prevent adverse impacts on systems relying on these feeds. This issue is particularly pressing in finance due to the clear incentives for misreporting. Misreporting can lead to significant financial losses and undermine trust in the system. Ensuring data reliability is essential to mitigate these risks and maintain financial stability.

In crypto, the problem is compounded by (1) the absence of regulation and (2) the immediate potential for exploitation from misreporting. Without regulatory oversight, malicious actors can exploit inaccuracies for financial gain. This vulnerability is part of the Miner Extractable Value (MEV) surface, where exploiters can profit from discrepancies in data. Addressing these issues requires robust mechanisms to secure data feeds and prevent exploitation.

1.0.2 Oracles in Crypto

Trusted Oracles - This approach relies on a single, centralized entity to provide data feeds. While it ensures high accuracy and consistency, it creates a single point of failure and

introduces centralization risks. If the trusted oracle is compromised or fails, the entire system's integrity can be jeopardized. This also makes the system vulnerable to manipulation and reduces trust in its decentralization. If a smart contract call is included in a block and the external API it relies on becomes unavailable, nodes cannot execute the block or compute the updated context. As a result, the entire blockchain could fail.

Reputation based oracles : this method, used by systems like Chainlink Lorenz Breidenbach et al. (2021), involves multiple oracles providing data and being evaluated based on their historical accuracy and reliability. The reputation of each oracle influences their impact on the final data feed, reducing the risk of manipulation. However this relies very much on centralization and the trustworthiness of these oracles.

Our approach to solving the oracle problem involves several key strategies. We leverage a permissionless system that relies on an honest majority, which is a common practice in decentralized environments. To enhance data accuracy and resilience, we utilize multiple oracles within small time units, which is enabled through DAG (Directed Acyclic Graph) environments, where each round features reports from numerous oracles. The large number of oracles n is chosen to increase robustness and mitigate risks of manipulation. Additionally, we investigate the game-theoretic aspects of oracle participation and estimator design to ensure that the system is both secure and efficient, addressing potential vulnerabilities and optimizing overall performance.

Chapter 2

Adversarially robust estimators

2.1 Estimators

This chapter discusses the fundamental concepts related to statistical estimators, including the definition of maximum error, bias, and various metrics used to compare them. Roughly, an estimator is a function that provides an estimate of an unknown parameter based on observed or reported data.

Let us write $\mathcal{P}$ a given collection of probability measures, the so-called model class. We have $\mathcal{P} = \{\mathcal{P}_\mu : \mu \in \mathbb{R}\}$.

Definition 2.1.0.1. *An estimator $T_n : \mathcal{P} \to \mathbb{R}$ is some given (measurable) function $T_n(\cdot)$ evaluated at the observations $X = (X_1, \ldots, X_n)$. The function $T_n(\cdot)$ is not allowed to depend on unknown parameters*

There are several properties that can qualify how good an estimator is. Indeed we need metrics to compare estimators and choose which one might be more appropriate in different scenarios. There are a few measures of quality of estimators, primarily: the bias and what we call the error of an estimator.

Now one of the measure will we use the most in this paper is the estimator error. This measures the absolute worst case deviation from the ground truth of the estimator

$$\epsilon_T = \sup_{X \in \mathcal{P}} |T_n(X) - \mu|, \tag{2.1.0.1}$$

Furthermore, we also define the bias of an estimator, which measures the systematic deviation of the estimated predictions from the ground truth

$$\text{bias}_\mu(T_n) = \mathbb{E}[T(X)] - \mu \tag{2.1.0.2}$$

where the expectation is taken wrt a given probability measure of the observed data.

Let us look at another measure of how good an estimator is. In our scenario, given we insist on the importance of permissionlessness in crypto, a measure of a good estimator is robustness. Robustness in statistical estimators refers to their ability to provide reliable and accurate results even when the underlying assumptions about the data are violated or when the data contains outliers. We assess estimators based on their sensitivity to extreme deviations in the data. Suppose $X_1, \ldots, X_n$ are i.i.d. samples from a random variable X. Consider T_n an estimator that is symmetric with respect to X.

Definition 2.1.0.2. *We define the influence function as the influence of a single additional observation on the result of the estimator*

$$l(x) := (n+1)\left[T_{n+1}(X_1, \ldots, X_n, x) - T_n(X_1, \ldots, X_n)\right], \quad x \in \mathbb{R}.$$

We define the breakdown point of an estimator is an intuitive measure of its robustness, indicating the smallest proportion of contamination that can cause the estimator to produce arbitrarily large incorrect values. It quantifies the estimator's resilience to outliers and provides insight into its reliability under data corruption.

Definition 2.1.0.3. *Let $m \leq n$, the breakdown point (BP) ϵ^* of T_n is defined as follows:*

$$\epsilon(m) := \sup_{x_1^*, \ldots, x_m^*} \left|T(x_1^*, \ldots, x_m^*, X_{m+1}, \ldots, X_n)\right|.$$

If $\epsilon(m) = \infty$, we say that with m outliers, the estimator can break down. The breakdown point is :

$$\epsilon^* := \frac{\min\{m : \epsilon(m) = \infty\}}{n}.$$

An estimator is robust if it has a bounded influence function and/or a BP.

Intuitively, the breakdown point is the maximum fraction of the data that can be contaminated before the estimator loses its reliability entirely. A high breakdown point indicates a robust estimator that can handle a significant amount of contamination without giving misleading results. The above measure of robustness will reveal itself to be an important aspect of why we will define a new estimator different than the traditional ones.

2.1.1 Huber's contamination model

Now to tackle our problem, one great paper has already set a very strong model. This paper was written by Huber (1963) where he talks about how estimators can be resilient to malicious voters in a sampled distribution. This is called the Huber contamination model. It assumes data comes from a mixture of two distributions: a majority $> \frac{1}{2}$ from a "clean" or "honest" distribution (e.g., normal distribution $\mathcal{N}(\mu, \sigma^2)$) and a minority from a "contaminating" or "attacker" distribution. Thus we assume here that honest nodes are normally distributed around the ground truth μ. Indeed there can be some noise in their observations which yields a distribution $H \sim N(\mu, \sigma)$. Suppose attacker nodes have $0 \leq \alpha < \frac{1}{2}$ percent of all voting power. We call this contamination power, then we can write

$$X_i \sim F = (1-\alpha)H + \alpha A \tag{2.1.1.1}$$

where H is normally distributed as we saw above, and the distribution A of attacker votes have no restriction as to how they are distributed. Now this model explains the setup of an observed distribution F by a mechanism. The goal of the mechanism is now, through an estimator, to output a value as close to the ground truth μ as possible. To that end we have to pick an estimator that can detect this while being robust to the α attacker votes. So what are the candidate estimators that we can chose from ? Now before we dive into that, let us explain one of the proofs that is most relevant to this model, presented by Huber in his paper. We call it *Huber's bound*.

2.1.2 Huber Bound

Huber shows the following argument: Suppose honest agents are normally distributed. Then there exists an attacker distribution, such that any translation invariant estimator has a bias of at least c. Now in Huber proof we have that $c = \alpha(1-\alpha)^{-1}(\varphi(0))^{-1}$. Here φ is the cdf of a $\mathcal{N}(0,1)$ distribution.

Theorem 2.1.2.1. *Let $\mathcal{P}$ be the set of probability distribution. $T : \mathcal{P} \to \mathbb{R}$ a translation invariant estimator of the location parameter. Then the median gives the smallest bias in the sense*

$$\sup_{f \in \mathcal{P}} bias(T(f)) \geq \alpha(1-\alpha)^{-1}(\varphi(0))^{-1}$$

Proof. Suppose we take the following density function,

$$f(t) = \begin{cases} (1-\alpha)\varphi(0) & |t| < c \\ (1-\alpha)\varphi(|t|-c) & |t| \geq c \end{cases} \tag{2.1.2.1}$$

where c is chosen in a way that f remains in $\mathcal{P}$. We show T cannot have a bias smaller that c. $f \in \mathcal{P}$ being a probability distribution, we have that $\int f(t)dt = 1$, thus we can check this implies $c = \alpha(1-\alpha)^{-1}(\varphi(0))^{-1}$. Now we notice that $f(t+c)$ and $f(t-c)$ are symmetric functions around $t = 0$, thus we have T should have the same bias for both these densities. By being translation invariant, we have that $T(f(t+c)) = T(f(t)) + c$. Now if T has a bias smaller than c for both of these densities we get

$$2c = T(f(t+c)) - T(f(t-c)) \leq |T(f(t+c))| + |T(f(t-c))| < 2c$$

hence our contradiction. So any translation invariant estimator T has a bias of at least c for this f. □

It can be shown that the proof holds for any symmetric distribution F around $t = \text{med}(F)$. We call this the Huber bound, if honest nodes have density $(1-\alpha)\cdot h$, and an attacker we can take the weighted median (meaning the value t such that $(1-\alpha)H(t) = \frac{1}{2}$, and recreate symmetry around that point by having $\forall x \geq 0, \alpha \cdot a(x) = (1-\alpha)\cdot(h(t-x) - h(x))$. This means that for any translation invariant estimator, under contamination, it is impossible to beat the above error.

Now this proof relies very much on symmetry, and it seems that it is only feasible for an attacker if he knows everything about the distribution of honest nodes. Namely the exact value of μ and σ. In the real world this doesn't seem very realistic, why would the

attacker know something that honest players don't ? Is this distribution still realistic to mimic when we have an error on one of the above parameters ? This brings us to the next section : choosing an estimator.

2.1.3 The leading candidates

Let us look at the estimators that are central in the paper where the model is introduced, namely "Robust estimation of a location parameter".

Mean – For the sample mean $T_n = \frac{1}{n}\sum_{i=1}^n X_i$, the influence function shows that the mean is highly sensitive to outliers. We can calculate the influence function of an observation x

$$l(x) = (n+1)\left[\frac{1}{n+1}\left(\frac{1}{n}\sum_{i=1}^n X_i + x\right) - \frac{1}{n}\sum_{i=1}^n X_i\right] = x \cdot \frac{n^2-1}{n^2} - \frac{n+1}{n}\cdot\frac{1}{n}\sum_{X_i \neq x} X_i.$$

thus we can already see that the influence function is unbounded, therefore that the breakdown point of the mean is 0.

Median – It can handle up to $\lfloor n/2 \rfloor$ outliers

$$\epsilon^* = \frac{\min\{m : \phi(m) = \infty\}}{n} = \frac{\lfloor n/2 \rfloor}{n}.$$

For large n, this approximates to $\frac{1}{2}$. Thus, the breakdown point of the median is 0.5. This highlights the robustness of the median compared to the mean. However, the median is much less precise than the mean when the sample size is small. This leads us to the following estimator defined by Huber.

Huber estimator – The Huber estimator is the middle-ground between the mean and the median. Let F be the distribution of X.

$$T_n = \text{argmin}_\mu \sum_{i=1}^n \rho_k(X_i - \mu),$$

where the Huber loss function ρ_k is:

$$\rho_k(u) = \begin{cases} \frac{u^2}{2} & \text{if } |u| \leq k, \\ k(|u| - \frac{k}{2}) & \text{if } |u| > k. \end{cases}$$

We define $\gamma := \arg\min_c P_\theta \rho_c$. Then by van de Geer (2022), we have that the influence function of the Huber estimator is

$$l(x) = \frac{1}{F(k+\gamma) - F(-k+\gamma)} \begin{cases} x - \gamma & |x-\gamma| \leq k \\ +k & x - \gamma > k \\ -k & x - \gamma < -k \end{cases}.$$

Thus it is bounded. Notice that for $k \to 0$, this corresponds to the influence function of the median.

We ask the following question: how easy is it to reach this bound if the attacker doesn't know the exact distribution of the attacker ? How easy is it for him to replicate the exact distribution proposed by Huber ?

Claim 1. With unknown μ, the median and Huber loss have an error ($\geq$ Huber bound) under the following strategy: attacker votes everything in $M \in \mathbb{R}$ arbitrarily large.

Proof. Clearly by definition of the median, we have that it outputs $\Phi_{\mu}^{-1}\left(\frac{1}{2(1-\alpha)}\right)$ which is the value such that there are a $\frac{n}{2}$ honest nodes $\leq$ than that value. For the Huber estimator, if $k = 0$ this is simply the median, thus the same result applies. Now if $k > 0$, by definition, this means that the estimator is less robust. This leads to the influence function evaluated in M to be larger. Therefore the final skew is larger than the Huber bound. □

This motivates the need for a new estimator, one that would make it tough to replicate this upper bound as soon as it doesn't know the exact form of the distribution of honest nodes. This is because these estimators do not rely on symmetry, thus we are tempted of finding an estimator that would require extreme precision to replicate Huber bound. Before that let us present our model which motivates this new estimator.

2.2 Continuous Model

In the continuous model we consider a large number of oracles n, such that the histogram of the reported observations of honest oracles can be considered to match a normal distribution's pdf, for all practical matters. Similarly, for the adversary oracle, a strategy would amount to producing a pdf (with mass points, potentially), rather than a discrete amount of $\alpha \cdot n$ data points. Honest nodes' pdf then integrates to $1 - \alpha$, whereas the attacker's pdf integrates to α, representing their relative power.

2.3 A new estimator

We would like an estimator T satisfying the following property. Let c be the Huber bound, then if the attacker only knows $\mu \in [\mu \pm \epsilon]$, then $\epsilon_T \leq c$. This motivates the following estimator.

Definition 2.3.0.1. *Let $f \in \mathbb{R}$ be a probability density function. Let $\mu \subseteq \mathbb{R}$ be the space of parameters. The Local Density Estimator (LDE) estimator T of μ is defined as*

$$T = \operatorname*{argmax}_{t \in \mathbb{R}} \int_{\mathbb{R}} \min(f(x), k_t(x))dx$$

where k_t is called the kernel function. If the argmax is not unique, the estimator returns the center of symmetry of the argmax set.

In this specific case, we will use this estimator with the kernel $k_t(x) = (1 - \alpha) \cdot \varphi_{t,\sigma}(x)$. Wlog we assume the ground truth $\mu = 0$. Intuition of the above estimator, finding the largest normal distribution of size $1-\alpha$ that lies under the density curve of all votes. Thus we slide a kernel, and at each mean t, we see if the votes create a normal distribution at that mean. Let f be the density of the attacker.

Lemma 2.3.0.2. *Let $0 \leq \eta \leq \Phi_{0,\sigma}^{-1}\left(\frac{1}{2\cdot(1-\alpha)}\right)$. Let a be the density of the attacker, and define*

$$a_{\eta}^{*}(x) = \begin{cases} 0 & \text{if } x \leq \eta \\ (1-\alpha) \cdot (\varphi_{2\eta,\sigma}(x) - \varphi_{0,\sigma}(x)) & \text{else} \end{cases}$$

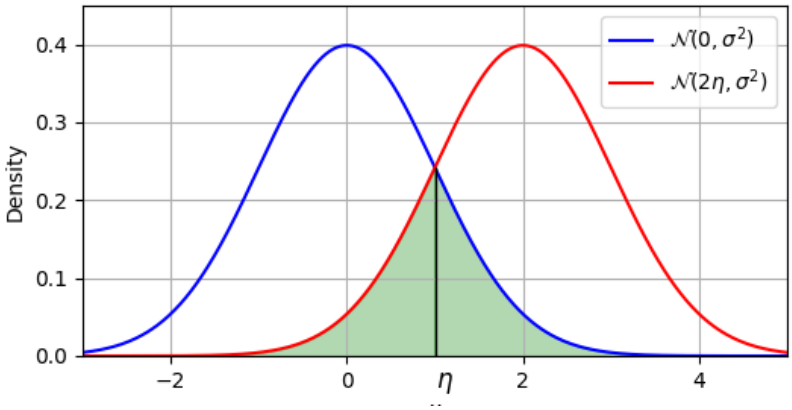

Figure 2.1: The blue distribution represents honest nodes, while the red represents the confusion attackers are attempting to generate. Green area represents honest votes which are used to confuse the mechanism

If $\exists x \in \mathbb{R}$ *such that* $a(x) < a^*_\eta(x)$*, then* $\hat{\mu} < \eta$*. Furthermore* $2\Phi_{0,\sigma}(\eta) - 1$ *percent of total votes are required to be able to create an error of* η*. Thus* η *is achievable iff* $\alpha > 2\Phi_{0,\sigma}(\eta) - 1 \iff \eta < \Phi^{-1}_{0,\sigma}\left(\frac{1+\alpha}{2}\right)$.

Proof. The integrand of the estimator cannot integrate to more than $1-\alpha$ (upper bound the integrand by $k_t(x)$). Hence for the estimator to output η, the integral needs to be equal to $1-\alpha$ in $t = 2\eta$. This means that we need to have $f(x) \geq (1-\alpha) \cdot \varphi_{2\eta,\sigma}(x), \forall x \in \mathbb{R}$, where f is the density of the histogram. Now having $f(x) = h(x) + a(x)$, then we need $a(x) \geq (1-\alpha) \cdot \varphi_{2\eta,\sigma}(x) - h(x)$ where $h(x) = (1-\alpha) \cdot \varphi_{0,\sigma}(x)$. □

Remark. only $2\Phi_{0,\sigma}(\eta) - 1$ percent of votes are needed to create this symmetry in η.

Suppose the attacker knows the exact value of σ, but only knows approximately the value of μ. Meaning it knows μ is somewhere in $[\mu - \epsilon, \mu + \epsilon]$. Note that we have that the error of our estimator is given by

$$\epsilon_{LDE} = \mathbb{E}_{\mathcal{P}_\mu}[|X - \mu|] = \frac{\sigma}{\sqrt{2\pi}}$$

In this case we show that with our estimator, it is impossible for the attacker to create an error of Hubers bound for our estimator. Indeed suppose I want to recreate Hubers bound. By what we have shown, this can only be achieved if I recreate a perfect symmetry in $\eta = Huber$. Therefore, if I want to take a shot at recreating this bound, I need to guess the value of μ. Hence I get this value correct with probability 0 as I have no information apart from the interval in which it lies. Instead of guessing, let us describe a safer way to vote if I only know μ up to an error of ϵ. How can we guarantee, that at η any value of μ in the interval will work. We want to create the largest possible η such that

$$\forall x \in \mathbb{R}, \quad f(x) \geq (1-\alpha) \cdot \varphi_{2\eta,\sigma}(x)$$

as in this case we can guarantee that when computing my estimator in $t = 2\eta$, then the integral will be $1-\alpha$ as we ensured that $f \geq \varphi_{2\eta,\sigma}$. However recall we want to use honest

votes as our own, and to ensure we don't expect more honest votes than there are, we end up with the final voting distribution

$$a_{\eta,\epsilon}(x) = \max\left\{(1-\alpha)\cdot\left(\varphi_{2\eta,\sigma}(x) - \min\left\{\varphi_{-\epsilon,\sigma}(x), \varphi_{\epsilon,\sigma}(x)\right\}\right), 0\right\}$$

This yeilds a final histogram distribution of $a_{\eta,\epsilon}(x) + (1-\alpha)\cdot\varphi_{0,\sigma}(x), x \in \mathbb{R}$ The above

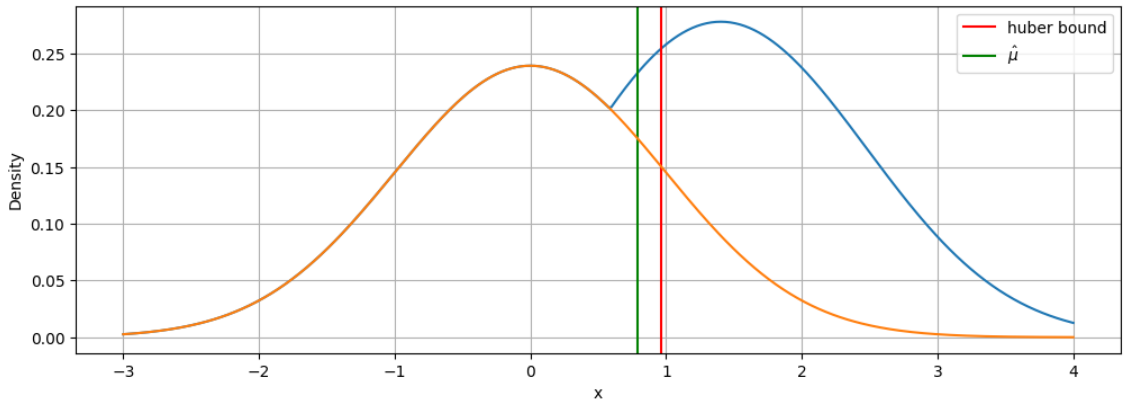

Figure 2.2: Plot of the histogram density for the optimal strategy when μ is unknown. Honest votes in orange, attacker votes are added to make the blue distribution

plot shows us that as soon as there is an error ϵ, our estimator achieves a better error than the Huber estimator or the median. Now in the real case scenario, as we said above there is an expected error of $\epsilon = \frac{\sigma}{\sqrt{2\pi}}$ on μ.

Chapter 3

Incentive compatible robust estimator

In the previous section we showed that if the attacker only knows $\mu \in [\mu \pm \epsilon]$, then the error of the mechanism is lower bounded. This section discusses whether incentives can help induce truthful behavior, encouraging malicious agents to vote closer to the ground truth. The challenge being : how should the mechanism distribute the budget W in a way that attackers get more profit from being truthful rather than inducing an error in the mechanism.

We define a model where the mechanism now receives a budget W. It then outputs an estimator of the ground truth and a payment function that redistributed the budget W amongst agents based on a scoring function. The source of the budget can be, for example, payments from external users of the system, such as the Aave Aave (2024) DeFi platform that uses oracles.

Definition 3.0.0.1. *A mechanism is said to be budget-balanced if the sum of all payments received by the participants is fixed. Formally, let n be the number of participants, and let p_i denote the payment received by participant i. The mechanism of budget W is budget-balanced if: $\sum_{i=1}^{n} p_i = W$*

Notice that compared to classical betting or wagering mechanisms, the ground truth (unknown by both mechanism and players) cannot be used to reward players. If it were known by the mechanism we could just reward based on the distance from μ, but here we cannot reveal it at the end of the vote in order to reward them.

One straightforward approach could be to give all W to the player closest to the estimator. We assume that no honest nodes ends up exactly on the estimator and only an all-knowing attacker agent can profit W as long as it keeps one vote on the final estimator value.

Claim 2. Under a rational attacker, then the sample mean estimator has a worst case error of $+\infty$. Whereas the sample median and LDE (local density estimator) have a worst case error of $\Phi_{\mu,\sigma}\left(\frac{1}{2(1-\alpha)}\right)$.

Indeed even if the attacker introduces its maximal possible error for a given estimator, it can just keep one vote x_i such that $X_i = T_n(X \backslash X_i)$. Hence the attacker ends up skewing the estimator to the Huber bound for the median and LDE (or $+\infty$ for the mean) and is

rewarded the whole budget W.

We now provide a good truthful-inducing mechanism. The mechanism fails with estimators such as the median or the Huber estimator. Meanwhile it induces truthfulness with our estimator, demonstrating its usefulness. Define the mechanism M^* which distributes the budget uniformly to all agents with value in the γ neighbourhood of $T_n(X) = \hat{\mu}$, then

$$\Gamma = \{i : X_i \in [\hat{\mu} - \gamma, \hat{\mu} + \gamma]\} \implies \forall i, p_i = \frac{W}{|\Gamma|} \cdot \mathbf{1}\{i \in \Gamma\}$$

continuous case, we want an exponential between 0 and 1 ensuring that we have a zero sum game. So if we take

$$e^{-\beta \cdot |x-\mu|}$$

now we distribute the budget W according to this score. A higher β represents a lower γ as our score will become negligible as soon as we are a bit too far from $\hat{\mu}$. The above integrates to $\frac{2}{\beta}$, thus if we want everything to integrate to 1, we decide

$$p_i = \frac{\beta \cdot e^{-\beta \cdot |x_i - \hat{\mu}|}}{2}$$

We observe that sample median fails as attacker can use $\alpha \cdot n - 1$ votes in an arbitrarily large value so that there are no honest votes in Γ. Then use its last vote to be in Γ. For the median, attacker can vote everything at the Huber bound. Let us now study the strategic reaction of the attacker with the LDE estimator.

Denote by $h(\cdot)$ and $H(\cdot)$ the honest pdf and cdf resp., and by $a_2(\cdot)$ and $A_2(\cdot)$ attacker's continuous pdf and cdf. Recall then that $a_{2,\eta}(x) = \phi_{2\eta}(x) - \phi_0(x)$, defined for $x \geq \eta$, and $a_{1,\eta}(x) = \alpha - \int_\eta^\infty a_{2,\eta}(x) = \alpha - A_2(\infty)$, defined for $x = \eta$. Now recall from Chapter 1 the remark (2.3), this gives us the following lemma

Lemma 3.0.0.2. *The optimal density for the attacker is given by*

$$a_\eta(x) = \begin{cases} 0 & x \leq \eta \\ \alpha - (1-\alpha) \cdot (2\Phi_0(\eta) - 1) & x = a_\gamma \\ (1-\alpha) \cdot (2\Phi_0(\eta) - 1) \cdot (\phi_{2\eta}(x) - \phi_0(x)) & x > \eta \end{cases} \tag{3.0.0.1}$$

where $a_\gamma \in [-\gamma, \gamma]$ ensure that the excess votes that aren't needed to recreate symmetry are a source of profit.

So if the attacker knows the ground truth ($\mu = 0$ here), it can skew it to $\hat{\mu} = \eta$ and use the remainder of its votes in $X_i = \hat{\mu}$. This means we know how the attacker will act if we assume it is rational and thus will pick the most profitable strategy.

Claim 3. For an error η of the estimator, the optimal utility for an attacker is given by

$$util_a(\eta) = W \cdot \frac{a_{1,\eta}(\eta) + A_2(\eta + \gamma)}{a_{1,\eta}(\eta) + A_2(\eta + \gamma) + H(\eta + \gamma) - H(\eta - \gamma)}$$

$$= W \cdot \left(1 + \frac{H(\eta + \gamma) - H(\eta - \gamma)}{a_{1,\eta}(\eta) + A_2(\eta + \gamma)}\right)^{-1}$$

Our goal is to show that the attacker's optimal strategy provides a lower bias than that of the median estimator (which in this case is the Huber bound $\Phi_{0,\sigma}^{-1}\left(\frac{1}{2(1-\alpha)}\right)$). Denote by η^* the argument that achieves optimum (maximum). To argue that our estimator is better than median estimator, we need to show that for some γ (note that η^* depends on γ), the attacker is choosing a smaller error ϵ_T than Huber's bound. It suffices to show that there exists a γ such that the derivative of the utility, wrt η, is (strictly) negative for $\eta \geq \eta_{hub}$. The derivative is (ignoring W):

$$-\frac{(util_a(\eta))^2}{(a_{1,\eta}(\eta)+A_2(\eta+\gamma)))^2}\cdot\Big[(a_{1,\eta}(\eta)+A_2(\eta+\gamma))\cdot(H'(\eta+\gamma)-H'(\eta-\gamma)) - (H(\eta+\gamma)-H(\eta-\gamma))\cdot\Big(a'_{1,\eta}(\eta)+A'_2(\eta+\gamma)\Big)\Big] \quad (3.0.0.2)$$

Now we want to show that for any positive γ, M^* is truthful or IC (incentive compatible) only up to γ. That is, the attacker is incentivized to deviate by γ. Formally we have

Claim 4. The mechanism M^* is such that

$$\forall\gamma>0,\ \operatorname*{argmax}_{\eta\in\mathbb{R}_+} util_a(\eta)=\gamma$$

Proof. We show that for any $\eta \geq \gamma$, the derivative of the utility of the attacker is ≤ 0. Therefore the attacker should skew at most by γ (as he gains nothing from skewing more). The derivative of the utility is negative if and only if

$$(a_{1,\eta}(\eta)+A_2(\eta+\gamma))\cdot(H'(\eta+\gamma)-H'(\eta-\gamma))-(H(\eta+\gamma)-H(\eta-\gamma))\cdot\Big(a'_{1,\eta}(\eta)+A'_2(\eta+\gamma)\Big)>0$$

Now, we have the following

$$a_2(x)=(1-\alpha)(\phi_{2\eta}(x)-\phi_0(x))=(1-\alpha)(\phi(x-2\eta)-\phi(x))$$

$$A_2(x)=\int_\eta^x a_2(t)dt$$

The mass of the attacker on η is: $a_1=\alpha-(A_2(\infty)-A_2(\eta))=\alpha-A_2(\infty)$ and the mass in the γ environment of η is thus: $a_1+A_2(\eta+\gamma)=\alpha-A_2(\infty)+A_2(\eta+\gamma)$, which is indeed smaller than α. It is:

$$\begin{aligned}
&\alpha-(1-\alpha)\int_\eta^\infty(\phi(t-2\eta)-\phi(t))\,dt+(1-\alpha)\int_\eta^{\eta+\gamma}(\phi(t-2\eta)-\phi(t))\,dt\\
&=\alpha-(1-\alpha)\int_{\eta+\gamma}^\infty(\phi(t-2\eta)-\phi(t))\,dt\\
&=\alpha-(1-\alpha)\left[\Phi(\eta+\gamma)-\Phi(\gamma-\eta)\right]\\
&=\alpha-(1-\alpha)\left[\Phi(\eta-\gamma)-1+\Phi(\eta+\gamma)\right]\\
&=1-(1-\alpha)(\Phi(\eta-\gamma)+\Phi(\eta+\gamma))
\end{aligned}$$

The derivative wrt η is given by : $-(1-\alpha)(\phi(\eta-\gamma)+\phi(\eta+\gamma))$ We define

$$H(\eta+\gamma)-H(\eta-\gamma)=(1-\alpha)\cdot(\Phi(\eta+\gamma)-\Phi(\eta-\gamma))$$

$$H'(\eta+\gamma) - H'(\eta-\gamma) = (1-\alpha)\cdot(\phi(\eta+\gamma) - \phi(\eta-\gamma))$$

Combining the above, we get negative utility iff

$$(1-\alpha)\cdot(\phi(\eta+\gamma) - \phi(\eta-\gamma))\cdot[1-(1-\alpha)(\Phi(\eta-\gamma)+\Phi(\eta+\gamma))] -$$
$$(1-\alpha)\cdot(\Phi(\eta+\gamma) - \Phi(\eta-\gamma))\cdot[-(1-\alpha)(\phi(\eta-\gamma)+\phi(\eta+\gamma))] > 0$$

or if we divide by $(1-\alpha)$ on both sides

$$(\phi(\eta+\gamma) - \phi(\eta-\gamma))\cdot[1-(1-\alpha)(\Phi(\eta-\gamma)+\Phi(\eta+\gamma))] +$$
$$(\Phi(\eta+\gamma) - \Phi(\eta-\gamma))\cdot[(1-\alpha)(\phi(\eta-\gamma)+\phi(\eta+\gamma))] > 0$$

and then develop the parentheses

$$(\phi(\eta+\gamma) - \phi(\eta-\gamma)) - (1-\alpha)\cdot(\phi(\eta+\gamma) - \phi(\eta-\gamma))\cdot(\Phi(\eta-\gamma)+\Phi(\eta+\gamma))$$
$$+(1-\alpha)\cdot(\phi(\eta-\gamma)+\phi(\eta+\gamma))\cdot(\Phi(\eta+\gamma) - \Phi(\eta-\gamma)) > 0$$

if we develop this further we get

$$(\phi(\eta+\gamma) - \phi(\eta-\gamma)) + (1-\alpha)\cdot(-2\cdot\phi(\eta+\gamma)\cdot\Phi(\eta-\gamma) + 2\cdot\phi(\eta-\gamma)\cdot\Phi(\eta+\gamma)) > 0$$

which can be written as

$$\phi(\eta+\gamma) - \phi(\eta-\gamma) - 2(1-\alpha)\phi(\eta+\gamma)\Phi(\eta-\gamma) + 2(1-\alpha)\phi(\eta-\gamma)\Phi(\eta+\gamma) > 0 \tag{3.0.0.3}$$

and evaluating in $\alpha = 0.5$ the above holds if

$$\phi(\eta+\gamma) - \phi(\eta-\gamma) - \phi(\eta+\gamma)\Phi(\eta-\gamma) + \phi(\eta-\gamma)\Phi(\eta+\gamma) > 0$$
$$\phi(\eta+\gamma)(1-\Phi(\eta-\gamma)) - \phi(\eta-\gamma)(1-\Phi(\eta+\gamma)) > 0$$

$$\frac{1-\Phi(\eta-\gamma)}{\phi(\eta-\gamma)} > \frac{1-\Phi(\eta+\gamma)}{\phi(\eta+\gamma)}$$

This holds due to Mills Ratio[1] being monotonically decreasing Árpád Baricz (2007).

[1] Mills ratio is defined as the ratio of the tail probability of the standard normal distribution to its density function:

$$\begin{aligned}
&\phi(\eta+\gamma)\Phi(\eta-\gamma) - \phi(\eta-\gamma)\Phi(\eta+\gamma) > 0 \\
&\iff \frac{\Phi(\eta-\gamma)}{\phi(\eta-\gamma)} > \frac{\Phi(\eta+\gamma)}{\phi(\eta+\gamma)} \\
&\iff \frac{1-\Phi(\gamma-\eta)}{\phi(\eta-\gamma)} > \frac{1-\Phi(-\eta-\gamma)}{\phi(\eta+\gamma)} \\
&\iff \frac{1-\Phi(\gamma-\eta)}{\phi(\gamma-\eta)} < \frac{1-\Phi(-\eta-\gamma)}{\phi(-\eta-\gamma)}
\end{aligned}$$

Millis ratio decreasing (so what now). Millis->Mills'. The above inequality holds due to Mills' ratio

For other α suffice it to show that the derivative with respect to α of the above previous expression (3.0.0.3) is negative:

$$2\phi(\eta+\gamma)\Phi(\eta-\gamma) - 2\phi(\eta-\gamma)\Phi(\eta+\gamma) < 0 \iff$$
$$\frac{\Phi(\eta-\gamma)}{\phi(\eta-\gamma)} < \frac{\Phi(\eta+\gamma)}{\phi(\eta+\gamma)} \iff$$
$$\frac{1-\Phi(\gamma-\eta)}{\phi(\gamma-\eta)} < \frac{1-\Phi(-\eta-\gamma)}{\phi(-\eta-\gamma)}$$

This again holds due to the Mills ratio being monotonically decreasing. We have thus shown that the derivative of the utility wrt η is negative for all $\eta \geq \gamma$. This proves that the best strategy for the attacker is to act honestly, that is, vote at $\eta = \gamma$. □

For now, we see supposedly no trade-off for picking an infinitesimally small γ.

3.0.1 Extension to exogenous incentives

We now discuss the dynamics of the mechanism in the face of external exogenous incentives. The previous section showed that any positive γ would suffice to induce truthful behaviour. However, what happens when a player also has an external incentives to deviate and skew the estimator? In such a case, can a budget of W, coupled with a correct mechanism, suffice to counter such incentives and induce truthfulness? Of course, if W is too small, no mechanism would suffice to disincentivize deviations. Indeed with a big potential off-chain profit, the attacker won't mind losing on a potential profit W. We will show below the minimal γ that is needed. We will then see how there now exists a trade-off with picking γ.

Assume the incentive to skew the mechanism is given by a linear factor of $\hat{\mu} : c_1 \cdot \hat{\mu}$. It is clear that when γ goes to 0, the attacker's best strategy is to deviate to the Huber bound

decreasing

Derivative wrt α is negative iff

$$2\phi(\eta+\gamma)\Phi(\eta-\gamma) - 2\phi(\eta-\gamma)\Phi(\eta+\gamma) < 0 \iff$$

Millis ratio decreasing so doesn't hold true (same results but with a negative sign so all false

$$\text{Mills Ratio}(x) = \frac{1-\Phi(x)}{\phi(x)}$$
$$\frac{1-\Phi(x)}{\phi(x)} \leq \frac{1}{x}$$
$$\frac{1-\Phi(x)}{\phi(x)} \geq \frac{1}{x} - \frac{1}{x^3}$$

and gain: The loss in in-mechanism profit from deviation goes to 0 with γ going to 0, and the attacker will gain $c_1 \cdot \hat{\mu}$ for $\hat{\mu} = \eta_{hub}$. Thus, we need a stronger γ:

The above analysis needs now to be repeated with the utility having another component, the external incentive.

$$util_a(\eta) = W \cdot \frac{a_{1,\eta}(\eta) + A_2(\eta+\gamma)}{a_{1,\eta}(\eta) + A_2(\eta+\gamma) + H(\eta+\gamma) - H(\eta-\gamma)} + c_1 \cdot \eta =$$

$$W \cdot \left(1 + \frac{H(\eta+\gamma) - H(\eta-\gamma)}{a_{1,\eta}(\eta) + A_2(\eta+\gamma)}\right)^{-1} + c_1 \cdot \eta \tag{3.0.1.1}$$

$$\begin{aligned}
\frac{\partial util_a}{\partial \eta} &= \frac{W}{\left(a_{1,\eta}(\eta) + A_2(\eta+\gamma) + H(\eta+\gamma) - H(\eta-\gamma)\right)^2} \\
&\quad \cdot \Bigg[\left(a'_{1,\eta}(\eta) + A'_2(\eta+\gamma)\right) \left(a_{1,\eta}(\eta) + A_2(\eta+\gamma) + H(\eta+\gamma) - H(\eta-\gamma)\right) \\
&\quad - \left(a_{1,\eta}(\eta) + A_2(\eta+\gamma)\right) \left(a'_{1,\eta}(\eta) + A'_2(\eta+\gamma) + H'(\eta+\gamma) - H'(\eta-\gamma)\right) \Bigg] + c_1 \\
&= \frac{W}{\left(a_{1,\eta}(\eta) + A_2(\eta+\gamma) + H(\eta+\gamma) - H(\eta-\gamma)\right)^2} \\
&\quad \cdot \Bigg[\left(a'_{1,\eta}(\eta) + A'_2(\eta+\gamma)\right) \left(H(\eta+\gamma) - H(\eta-\gamma)\right) \\
&\quad - \left(a_{1,\eta}(\eta) + A_2(\eta+\gamma)\right) \left(H'(\eta+\gamma) - H'(\eta-\gamma)\right) \Bigg] + c_1
\end{aligned}$$

There exist W and γ such that the mechanism induces truthfulness. Moreover, for any W above a certain lower bound, we should have the existence of a γ that induces truthfulness. Indeed if W is too small, then d goes to infinity and we will see in the proof that this is problematic. Moreover we find the asymptotic relationship between γ and W.

Theorem 3.0.1.1. *There exist $\gamma, W > 0$ such that*

$$\operatorname*{argmax}_{\eta \in \mathbb{R}_+} util_a(\eta) = \gamma \tag{3.0.1.2}$$

Moreover, if $c_1 = 1$ and

$$W \leq \operatorname*{argmax}_{\gamma \in \mathbb{R}_+} \; \phi(2\gamma) - \sqrt{\frac{2}{\pi}} \cdot (1 - \Phi(2\gamma))$$

then there exits no γ such that (3.0.1.2) holds. Moreover, γ is asymptotically inversely related to W.

Proof. the derivative of the utility is negative if and only if

$$\begin{aligned}&\frac{c_1}{W}\cdot\left(a_{1,\eta}(\eta)+A_2(\eta+\gamma)+H(\eta+\gamma)-H(\eta-\gamma)\right)^2\\&<\left(a_{1,\eta}(\eta)+A_2(\eta+\gamma)\right)\left(H'(\eta+\gamma)-H'(\eta-\gamma)\right)\\&-\left(a'_{1,\eta}(\eta)+A'_2(\eta+\gamma)\right)\left(H(\eta+\gamma)-H(\eta-\gamma)\right)\end{aligned}\tag{3.0.1.3}$$

We saw that the RHS equals

$$(1-\alpha)\cdot(\phi(\eta+\gamma)-\phi(\eta-\gamma))\cdot[1-(1-\alpha)(\Phi(\eta-\gamma)+\Phi(\eta+\gamma))]-$$
$$(1-\alpha)\cdot(\Phi(\eta+\gamma)-\Phi(\eta-\gamma))\cdot[-(1-\alpha)(\phi(\eta-\gamma)+\phi(\eta+\gamma))]$$

and that it is positive for any $\gamma>0$. We now need conditions on γ for it to dominate the LHS. From prvs calcs it follows that the term in the LHS parenthesis in (3.0.1.3) equals:

$$\begin{aligned}&1-(1-\alpha)(\Phi(\eta-\gamma)+\Phi(\eta+\gamma))+(1-\alpha)\cdot(\Phi(\eta+\gamma)-\Phi(\eta-\gamma))\\&=1-2(1-\alpha)\Phi(\eta-\gamma)\end{aligned}$$

Thus we need to solve

$$\begin{aligned}&(1-\alpha)\cdot(\phi(\eta+\gamma)-\phi(\eta-\gamma))\cdot[1-(1-\alpha)(\Phi(\eta-\gamma)+\Phi(\eta+\gamma))]\\&-(1-\alpha)\cdot(\Phi(\eta+\gamma)-\Phi(\eta-\gamma))\cdot[-(1-\alpha)(\phi(\eta-\gamma)+\phi(\eta+\gamma))]\\&-\frac{c_1}{W}\cdot(1-2(1-\alpha)\Phi(\eta-\gamma))^2>0\end{aligned}\tag{3.0.1.4}$$

Or

$$\begin{aligned}&\frac{(\phi(\eta+\gamma)-\phi(\eta-\gamma))\cdot[1-(1-\alpha)(\Phi(\eta-\gamma)+\Phi(\eta+\gamma))]}{(1-2(1-\alpha)\Phi(\eta-\gamma))^2}\\&-\frac{(\Phi(\eta+\gamma)-\Phi(\eta-\gamma))\cdot[-(1-\alpha)(\phi(\eta-\gamma)+\phi(\eta+\gamma))]}{(1-2(1-\alpha)\Phi(\eta-\gamma))^2}\\&-\frac{c_1}{W}\frac{1}{1-\alpha}>0\end{aligned}\tag{3.0.1.5}$$

From 3.0.0.3 we get that the above is

$$\frac{\phi(\eta+\gamma)-\phi(\eta-\gamma)-2(1-\alpha)\phi(\eta+\gamma)\Phi(\eta-\gamma)+2(1-\alpha)\phi(\eta-\gamma)\Phi(\eta+\gamma)}{(1-2(1-\alpha)\Phi(\eta-\gamma))^2}-\frac{c_1}{W}\frac{1}{1-\alpha}>0$$

Or

$$\begin{aligned}&(1-\alpha)\phi(\eta+\gamma)(1-2(1-\alpha)\Phi(\eta-\gamma))-(1-\alpha)\phi(\eta-\gamma)(1-2(1-\alpha)\Phi(\eta+\gamma)\\&-(1-2(1-\alpha)\Phi(\eta-\gamma))^2\cdot\frac{W}{c_1}>0\end{aligned}$$

we have that $1-\alpha>1/2$, hence it is clear that the term $(1-2(1-\alpha)\Phi(\eta-\gamma))^2\cdot\frac{W}{c_1}$ is greater in α than in $1/2$ as there are no other terms in α. Now for the rest

$$\begin{aligned}&2(1-\alpha)\phi(\eta+\gamma)(1-2(1-\alpha)\Phi(\eta-\gamma))-2(1-\alpha)\phi(\eta-\gamma)(1-2(1-\alpha)\Phi(\eta+\gamma)\\&>\phi(\eta+\gamma)(1-\Phi(\eta-\gamma))-\phi(\eta-\gamma)(1-\Phi(\eta+\gamma)\end{aligned}$$

$$\begin{aligned}\frac{df(\alpha)}{d\alpha} &= -\phi(\eta+\gamma)\left(1-2(1-\alpha)\Phi(\eta-\gamma)\right)+2(1-\alpha)\phi(\eta+\gamma)\Phi(\eta-\gamma)\\ &\quad+\phi(\eta-\gamma)\left(1-2(1-\alpha)\Phi(\eta+\gamma)\right)-2(1-\alpha)\phi(\eta-\gamma)\Phi(\eta+\gamma)<0\\ &\iff \phi(\eta-\gamma)\left(1-2(1-\alpha)\Phi(\eta+\gamma)\right)-2(1-\alpha)\phi(\eta-\gamma)\Phi(\eta+\gamma)\\ &<\phi(\eta+\gamma)\left(1-2(1-\alpha)\Phi(\eta-\gamma)\right)+2(1-\alpha)\phi(\eta+\gamma)\Phi(\eta-\gamma)\end{aligned}$$

addition

$$\begin{aligned}&\iff \phi(\eta-\gamma)\left(1-4(1-\alpha)\Phi(\eta+\gamma)\right)<\phi(\eta+\gamma)\left(1-4(1-\alpha)\Phi(\eta-\gamma)\right)\\ &\iff \frac{1-4(1-\alpha)\Phi(\eta+\gamma)}{\phi(\eta+\gamma)}<\frac{1-4(1-\alpha)\Phi(\eta-\gamma)}{\phi(\eta-\gamma)}\iff\\ &\iff \frac{1-\Phi(\eta+\gamma)-3(1-\frac{4}{3}\alpha)\Phi(\eta+\gamma)}{\phi(\eta+\gamma)}<\frac{1-\Phi(\eta-\gamma)-3(1-\frac{4}{3}\alpha)\Phi(\eta-\gamma)}{\phi(\eta-\gamma)}\\ &\Leftarrow \frac{\Phi(\eta+\gamma)}{\phi(\eta+\gamma)}>\frac{\Phi(\eta-\gamma)}{\phi(\eta-\gamma)},\end{aligned}$$

and the latter inequality holds since $\Phi()$ is monotonically increasing and $\phi()$ is decreasing (after the mean val, so for $\gamma<\eta$). The implication in the line before ($\Leftarrow$) stems from the beloved Mills' ratio monotonically decreasing.

end of addition

$$\phi(\eta-\gamma)\cdot\frac{\Phi(\eta-\gamma)}{\Phi(\eta+\gamma)}-\phi(\eta+\gamma)<\frac{c_1}{W}<\frac{\phi(\eta+\gamma)}{1-\Phi(\eta-\gamma)}-\frac{\phi(\eta-\gamma)(1-\Phi(\eta+\gamma))}{(1-\Phi(\eta-\gamma))^2}$$

Or, as simplified in the prvs section, for $\alpha=0.5$,

$$\frac{\phi(\eta+\gamma)(1-\Phi(\eta-\gamma))-\phi(\eta-\gamma)(1-\Phi(\eta+\gamma))}{(1-\Phi(\eta-\gamma))^2}-\frac{2c_1}{W}>0 \tag{3.0.1.6}$$

$$\begin{aligned}W\cdot\phi(\eta+\gamma)(1-\Phi(\eta-\gamma))-W\cdot\phi(\eta-\gamma)(1-\Phi(\eta+\gamma))\\ -2c_1\left(\Phi(\gamma-\eta)\right)^2>0\end{aligned} \tag{3.0.1.7}$$

We will now evaluate the derivative of this at $\eta=\gamma$, and calculate the minimal γ for which the derivative is negative:

For $\eta=\gamma$ the expression simplifies to:

$$\phi(\eta+\gamma)(1-\Phi(\eta-\gamma))-\phi(\eta-\gamma)(1-\Phi(\eta+\gamma))-2d_1\left(\Phi(\gamma-\eta)\right)^2>0$$

Substituting $\eta=\gamma$:

$$\phi(2\gamma)(1-\underbrace{\Phi(0)}_{=1/2}) - \phi(0)(1-\Phi(2\gamma)) - 2d_1\left(\Phi(0)\right)^2 > 0$$
$$\iff \phi(2\gamma)(1-0.5) - \phi(0)(1-\Phi(2\gamma)) - 2d_1(0.5)^2 > 0$$

This simplifies to:

$$0.5\phi(2\gamma) - \phi(0)(1-\Phi(2\gamma)) - \frac{d_1}{2} > 0$$
$$\iff 0.5\phi(2\gamma) - \frac{1}{\sqrt{2\pi}}(1-\Phi(2\gamma)) - \frac{d_1}{2} > 0 \tag{3.0.1.8}$$

□

3.0.2 Visual representation

We now illustrate the implications of this trade-off and parameter relationship. From what we have seen above, we write Ext_Inc_Coef := c_1 as it represents the external incentive coefficient. By the same reasoning we write W as mechanism budget and γ as the estimator error. Furthermore we write Inc_Coef_Ratio when talking about $d = \frac{c_1}{W}$. Recall here that c_1 is a variable that is out of the control of the mechanism, the latter can only control γ and W. All of the plots deal with the case $\sigma = 1$ for the scale of honest nodes.

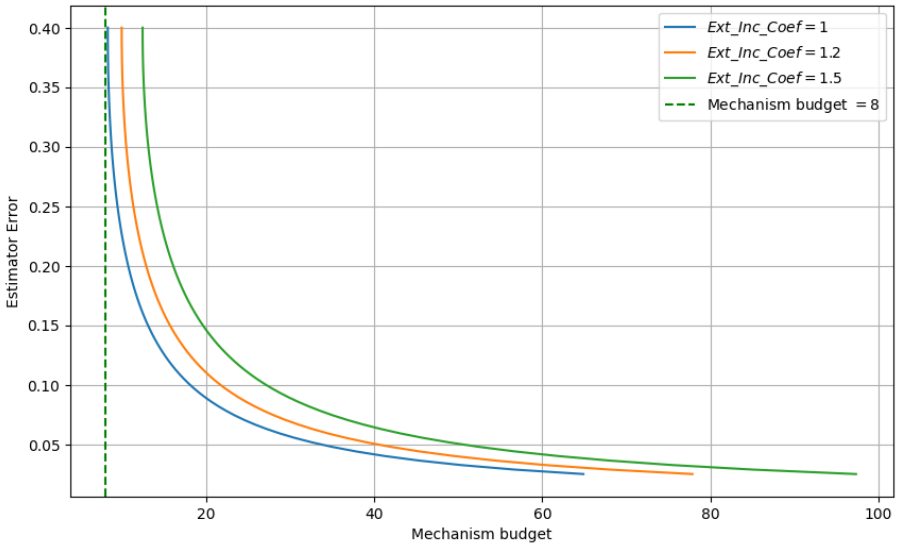

Figure 3.1: Plot of the budget W that is needed to achieve a certain estimator error when the external incentive coefficient $c_1 \in \{1, 1.2, 1.5\}$. The green line in $W = 8$ shows that with an arbitrarily small budget, the mechanism can't be fixed

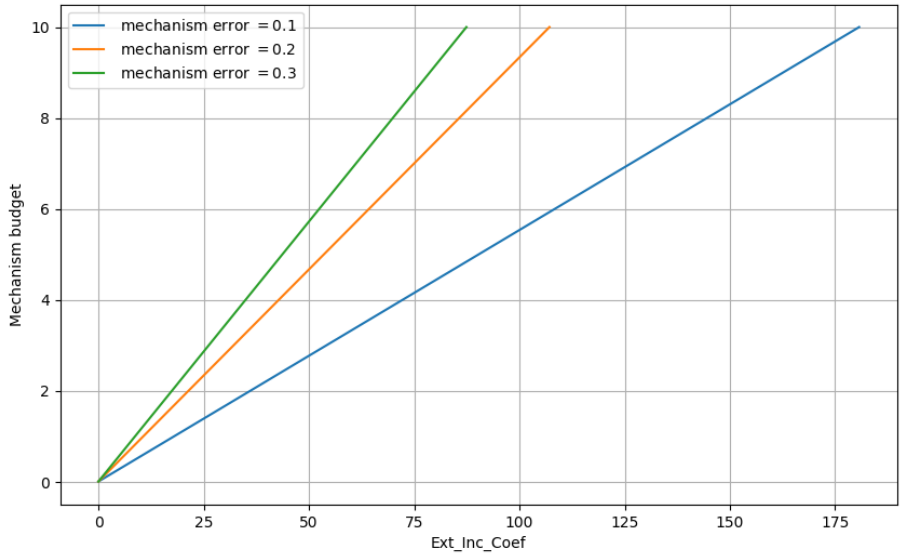

Figure 3.2: Plot of the mechanism budget W as a function of the External Incentive coefficient for different value of the mechanism error

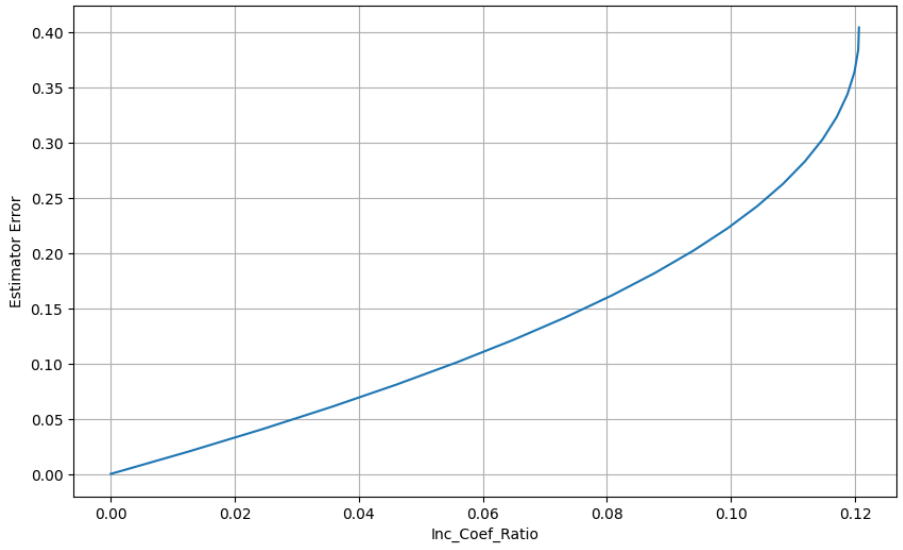

Figure 3.3: Plot the value of the Incentive Coefficient Ratio d as a function of the mechanism error γ. The cutoff is due to the fact that for any larger d, the budget W is too small to fix the mechanism

Chapter 4

Conclusion

Although robust estimators have been developed in the past, they seem to be sub-optimal in our setup. Indeed when facing a rational agent, or attacker, that has a large power, no mechanism can detect the ground truth with precision ϵ when using the traditional estimators. Although at first we could think the Huber bound cannot be improved upon, we realise that the proof requires the attacker to be all-knowing, which is a too strong hypothesis for the attacker.

This lead us to come up with a new estimator that would be more robust to adversarial influences. It proved to be more resilient than other known estimators whenever the attacker lacks precise knowledge of the distribution of the data feed. Through the utilization of a new estimator, we manage to induce truthfulness in rational agents.

Oracles are essential in cryptocurrency. One important use-case is to determine threshold prices for loan liquidations. Activity in DeFi often involves over-collateralized loans. The latter are dependent on the price feed of oracles to ensure the collateral remains above a specific threshold relative to the borrowed asset. For example, stablecoin contracts like MakerDAO, as well as lending liquidity pools such as Aave heavily rely on oracles for proper functioning. However these systems often rely on centralization points as they depend on a set of oracles that must be trusted. When the collateral threshold is breached, a liquidation event occurs, potentially resulting in millions of dollars in liquidations, creating a significant opportunity for manipulations under the umbrella term of MEV.

Our work introduces a new type of dynamic Oracle and corresponding contracts that do not rely on a single or set of permissioned oracles. Instead, it allows oracles to be permissionless while remaining robust against manipulations by entities such as MEV.

Bibliography

Aave (2024). https://aave.com/.

Daian, P. (2022). Mev for the next trillion, it's time to get serious.... https://writings.flashbots.net/mev-for-the-next-trillion.

Huber, P. (1963). Robust estimation of a location parameter. pp. 10.

Lorenz Breidenbach, C. C. et al. (2021). Chainlink 2.0: Next steps in the evolution of decentralized oracle networks. https://research.chain.link/whitepaper-v2.pdf.

van de Geer, S. (2022). Mathematical statistics.

Árpád Baricz (2007). Mills' ratio: Monotonicity patterns and functional inequalities. https://core.ac.uk/download/pdf/82775732.pdf.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. In consultation with the supervisor, one of the following three options must be selected:

- (●) I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].
- ( ) I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used and cited generative artificial intelligence technologies[2].
- ( ) I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used generative artificial intelligence technologies[3]. In consultation with the supervisor, I did not cite them.

Title of paper or thesis:

Incentive compatible mechanisms for robust estimation

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Mea	Eliott

With my signature I confirm the following:

- I have adhered to the rules set out in the Citation Guide.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
22/07/2024	Eliott Mea

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] E.g. ChatGPT, DALL E 2, Google Bard
[2] E.g. ChatGPT, DALL E 2, Google Bard
[3] E.g. ChatGPT, DALL E 2, Google Bard