Eliciting Private Information with Noise: The Case of Randomized Response

Andreas Blume Ernest K. Lai Wooyoung Lim

Maximal Transmittable Information under Direct Response and Randomized Response

1 Defining Mutual Information

The naive use of a garbling scheme (trusting that it results in truthful behavior) may lead to misleading interpretations when there are multiple equilibria and the analyst is unaware of which equilibrium is being played. This is the worst case scenario for randomized response. At the other extreme, it is worth considering how well randomized response does as a garbling scheme when the analyst knows the equilibrium behavior of senders. In particular, one may ask how much information can be transmitted under the best randomized response equilibrium versus the best direct response equilibrium. We use *mutual information*, a standard measure from information theory (Shannon [7]), to address this question.

We begin with a brief discussion of the nature and properties of mutual information in the context of our environment. Suppose $\Pr(\theta') > 0$ is the prior of the sender's type θ' and $\Pr(\theta'|r')$ the posterior upon observation of message r'. When r' is observed at θ', there is an informational gain if $\Pr(\theta'|r') > \Pr(\theta')$ or $\frac{\Pr(\theta'|r')}{\Pr(\theta')} > 1$. Similarly, an informational loss occurs at θ' if $\frac{\Pr(\theta'|r')}{\Pr(\theta')} < 1$. One can assign numerical values $v\left(\frac{P(\theta'|r')}{P(\theta')}\right)$ to the informational gains and losses by introducing a function $v : \mathbb{R} \to \mathbb{R}$ that is strictly monotonic, continuous and satisfies $v(1) = 0$. One such function is the logarithm. Using $\log(\cdot)$ for $v(\cdot)$, the expected net informational gain about the random variable θ due to the observation of the random variable r is thus

$$I(\theta; r) = \sum_{(\theta',r')\in\{s,t\}\times\{y,n\}} P(\theta', r') \log \frac{P(\theta'|r')}{P(\theta')}, \tag{1}$$

which is precisely the definition of mutual information.[1] Note that the above expression can be rewritten as $I(\theta; r) = H(\theta) - H(\theta|r)$, where $H(\theta) = -\sum_{\theta' \in \{s,t\}} \Pr(\theta') \log \Pr(\theta')$ is the entropy of the sender's type and

$$H(\theta|r) = -\sum_{r' \in \{y,n\}} \Pr(r') \sum_{\theta' \in \{s,t\}} \Pr(\theta'|r') \log \Pr(\theta'|r')$$

is the conditional entropy of the sender's type given the message r. Entropy is a measure of the uncertainty of a random variable. Mutual information therefore measures, quite intuitively, the reduction in the uncertainty about the sender's type θ due to the observation of the sender's message r; it ranges from zero to one.[2,3]

2 Maximal Mutual Information under Direct Response and Randomized Response

In light of the multiple equilibria under randomized response, we focus on the question: for a given value of ρ, what is the mutual information of the respective most informative equilibria under randomized response and under direct response, with "informativeness" evaluated with respect to mutual information? We denote the maximal mutual information by $\bar{I}_D(\rho)$ for direct response and $\bar{I}_R(\rho)$ for randomized response.

Under direct response, the uninformative and the informative equilibria exist complementarily under the respective cases $\rho \in (0, \frac{1}{2}]$ and $\rho \in (\frac{1}{2}, 1)$. Accordingly, the application

[1] By convention, $0 \log 0 = 0$, which can be justified by continuity.

[2] Mutual information is also referred to as relative entropy or Kullback-Leibler divergence, the divergence between the joint and product distributions of the random variable in question. If the base of the logarithm is 2, which is commonly adopted in information theory, then the unit of the entropy is in bits; if the base is e, the unit is in nats. Given that our model has a binary type space, we use 2 as our base. For an excellent reference in information theory, see Cover and Thomas [1].

[3] Given that in our model no payoff function is specified for the receiver, there is no obvious candidate for defining a value of information that would be less arbitrary than using mutual information. Also, pursuing the goal of maximizing the precision of the estimator of the population frequency of stigmatization subject to a truth-telling constraint, as in Ljungqvist [6], is compromised by the presence of multiple equilibria. This, and the fact that mutual information is widely used in information theory, motivate us to adopt it as our measure of informational gain. Jose, Nau and Winkler [4] investigate how entropy measures of information relate to utility. Kelly [5] links information-theoretic measures with the value of information in the case of a gambler who receives information through a noisy channel. Donaldson-Matasci, Bergstrom and Lachmann [3] identify uncertain environments in which the biological fitness value of information corresponds exactly to mutual information and show more generally that mutual information is an upper bound on the fitness value of information. Information-theoretic measures of information have been used in macroeconomics to study the consequences of information processing constraints (Sims [8]), and in organization theory to capture the idea that organizations have limited communication capacity (Dessein, Galeotti and Santos [2]).

of the definition of mutual information in (1) gives the following characterization:

Proposition 1. *Under direct response, the maximal mutual information allowed by any equilibrium is*

$$\bar{I}_D(\tfrac{\lambda}{\xi}) = \begin{cases} 0, & \textit{if } \rho \in (0, \frac{1}{2}], \\ 1 + \frac{1}{2}[(\frac{1}{\rho} - 1)\log(1-\rho) + \log \rho], & \textit{if } \rho \in (\frac{1}{2}, 1). \end{cases}$$

***Proof*.** If $\rho \in (0, \frac{1}{2}]$, any equilibrium must be uninformative. The receiver's posterior beliefs are the same as the prior, which implies that $H(\theta|r) = H(\theta) = 1$, and thus $I(\theta; r) = 0$. Note that the out-of-equilibrium beliefs do not enter into the calculation because for the unused answer r', $\Pr(r') = 0$.

If $\rho \in (\frac{1}{2}, 1)$, in the unique equilibrium $\sigma(y|t) = 1$ and $\sigma(y|s) = \frac{1}{\rho} - 1$, and thus $\Pr(y) = \frac{1}{2\rho}$. Bayes' rule implies that $\mu_s(y) = 1 - \rho$ and $\mu_s(n) = 1$. Accordingly, $H(\theta|r) = -(\frac{1}{2\rho})[(1-\rho)\log(1-\rho) + \rho \log \rho]$, where $0 \log 0 = 0$ is used. Thus, for the unique equilibrium under $\rho \in (\frac{1}{2}, 1)$, $I(\theta|r) = 1 + \frac{1}{2}[(\frac{1}{\rho} - 1)\log(1-\rho) + \log \rho]$.

□

With the continuum of informative equilibria, the determination of maximal mutual information is less straightforward under randomized response. To facilitate the exposition, we start with the following lemma:

Lemma 1. *Under randomized response,*

1. *for $\rho \in (\frac{1}{2}, 1)$ and $p_s = \frac{1}{\rho} - 1$ or $p_s = 2 - \frac{1}{\rho}$, there exist equilibria whose mutual information coincides with $\bar{I}_D(\frac{\lambda}{\xi})$; and,*

2. *for $\rho \in (0, 1)$, the maximal mutual information among the truthful equilibria is $\bar{I}_{R-T}(\rho) = \frac{1}{2}[(1-\rho)\log(1-\rho) + (1+\rho)\log(1+\rho)]$, achieved at $p_s = \frac{1-\rho}{2}$.*

Furthermore, there exists a $c \approx 0.743$ such that $\bar{I}_{R-T}(\rho) > \bar{I}_D(\rho)$ for $\rho \in (0, c)$ and $\bar{I}_{R-T}(\rho) \leq \bar{I}_D(\rho)$ for $\rho \in [c, 1)$ with strict inequality except at $\rho = c$.

***Proof*.** For Part 1, note that from Proposition 1, we have that for $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$, $\bar{I}_D(\rho) = 1 + \frac{1}{2}[(\frac{1}{\rho} - 1)\log(1-\rho) + \log \rho]$, which is derived from the equilibrium in which $\sigma(y|t) = 1$ and $\sigma(y|s) = \frac{1}{\rho} - 1$. The strategy profile implies the following components for mutual information, $\mu_s(y) = 1 - \rho$, $\mu_s(n) = 1$ and $\Pr(y) = \frac{1}{2\rho}$. We first show that there is an equilibrium under randomized response that has the same components. Consider the equilibrium in which $\sigma(y|s, q_s) = \sigma(y|t, q_t) = \sigma(y|t, q_s) = 1$ and $\sigma(y|s, q_t) = 0$, which exists if and only if $p_s = \frac{1}{\rho} - 1$ and $\rho > \frac{1}{2}$. It is immediate that $\mu_s(n) = 1$, $\mu_s(y) = 1 - \rho$, and $\Pr(y) = \frac{1}{2}(1 + p_s) = \frac{1}{2\rho}$. We show next that there is another equilibrium under randomized response that has the same components up to rotation of the answers, and thus has the same mutual information. Consider the equilibrium in which $\sigma(y|s, q_t) = \sigma(y|t, q_s) = \sigma(y|t, q_t) = 0$ and $\sigma(y|s, q_s) = 1$, which exists if and only if $p_s = 2 - \frac{1}{\rho}$ and $\rho > \frac{1}{2}$. It is immediate that $\mu_s(y) = 1$, $\mu_s(n) = 1 - \rho$, and $\Pr(n) = \frac{1}{2}(2 - p_s) = \frac{1}{2\rho}$. Thus, for $\rho \in (\frac{1}{2}, 1)$ and $p_s \in \{\frac{1}{\rho} - 1, 2 - \frac{1}{\rho}\}$, there exist equilibria under randomized response whose mutual information is $1 + \frac{1}{2}[(\frac{1}{\rho} - 1)\log(1 - \rho) + \log\rho]$.

For Part 2, consider the truthful equilibria in which $\sigma(y|s, q_s) = \sigma(y|t, q_t) = 1$ and $\sigma(y|s, q_t) = \sigma(y|t, q_s) = 0$, which exist if and only if $p_s \geq \frac{1-\rho}{2}$.

The strategy profiles imply that $\mu_s(y) = p_s$, $\mu_s(n) = 1 - p_s$, and $\Pr(y) = \Pr(n) = \frac{1}{2}$. The resulting mutual information is thus $1 + p_s \log p_s + (1 - p_s)\log(1 - p_s)$, which attains its minimum at $p_s = \frac{1}{2}$ and is strictly convex in p_s. This implies that for $p_s \in [\frac{1-\rho}{2}, \frac{1}{2})$, the mutual information attains its maximum when $p_s = \frac{1-\rho}{2}$. Substituting $p_s = \frac{1-\rho}{2}$ into $1 + p_s \log p_s + (1 - p_s)\log(1 - p_s)$, we obtain $\bar{I}_{R-T}(\rho) = \frac{1}{2}[(1 - \rho)\log(1 - \rho) + (1 + \rho)\log(1 + \rho)]$.

Finally, we compare the two values of mutual information, $1 + \frac{1}{2}[(\frac{1}{\rho} - 1)\log(1 - \rho) + \log\rho]$ and $\frac{1}{2}[(1 - \rho)\log(1 - \rho) + (1 + \rho)\log(1 + \rho)]$. Subtracting the latter from the former, we define $\Delta\bar{I}(\rho) = \frac{1}{2}[(2 - \rho - \frac{1}{\rho})\log(1 - \rho) + (1 + \rho)\log(1 + \rho) - \log\rho] - 1$ for $\rho = [\frac{1}{2}, 1]$, using the fact that the expression is well-defined at the endpoints of the interval $[\frac{1}{2}, 1]$. Note that $\Delta\bar{I}(\frac{1}{2}) = \frac{3}{4}\log 3 - 1 > 0$, $\Delta\bar{I}(1) = 0$, and $\frac{d\Delta\bar{I}(\rho)}{d\rho} = \frac{(1-\rho^2)\ln(1-\rho) + \rho^2\ln(1+\rho)}{\rho^2 \ln 4} > 0$ at $\rho = 1$. Hence, there exists $x \in (0, \frac{1}{2})$ for which $\Delta\bar{I}(x) < 0$, and, by the intermediate value theorem, there exists a $c \in (\frac{1}{2}, 1)$ with $\Delta\bar{I}(c) = 0$. Since $\frac{d^2\Delta\bar{I}(\rho)}{d\rho^2} = -\frac{\rho(1+2\rho) + 2(1+\rho)\ln(1-\rho)}{\rho^3(1+\rho)\ln 4} > 0$ for $\rho \in [\frac{1}{2}, 1]$, this c is unique. It can be verified numerically that $c \approx 0.743$.

□

Under direct response, the mutual information is solely determined by the sender's strategy, which in the informative equilibrium consists of truth-telling by type t, $\sigma(y|t) = 1$, and randomization by type s, $\sigma(n|s) = 2 - \frac{1}{\rho}$. Under randomized response, the probabilities of the questions also contribute to determining the mutual information. This suggests the possibility that the non-degenerate question probabilities may serve as an exogenous randomization to mimic the equilibrium randomization under direct response, resulting in the same set of anwer probabilities and posteriors that enter into the computation of mutual information.

The first part of Lemma 1 says that this is indeed the case. The analysis boils down to finding p_s, $\sigma(y|t, q_s)$, $\sigma(y|t, q_s)$, $\sigma(n|s, q_s)$, and $\sigma(n|s, q_t)$ under randomized response so that $p_s\sigma(y|t, q_s) + (1 - p)\sigma(y|t, q_t) = 1$ and $p_s\sigma(n|s, q_s) + (1 - p)\sigma(n|s, q_t) = 2 - \frac{1}{\rho}$. These conditions are satisfied by $p_s = \frac{1}{\rho} - 1$ paired with $\sigma(y|t, q_s) = \sigma(y|t, q_t) = \sigma(n|s, q_t) = 1$ and $\sigma(n|s, q_s) = 0$, and the strategy forms an equilibrium under randomized response if and only if p_s is at that exact value. The two equilibria under two different responses result in the same posteriors; this is no coincidence because the incentive conditions behind one equilibrium carry over to the other.

The intuition behind the second part of Lemma 1 can be seen from the fact that when $p_s = \frac{1}{2}$, the uninteresting case that we ruled out, no information is transmitted regardless of the sender's strategy; the receiver's posteriors will remain at $\frac{1}{2}$. More information is transmitted, and thus the mutual information increases, when p_s decreases from $\frac{1}{2}$. Given the constraint that the truthful equilibria can be supported only for $p_s \geq \frac{1-\rho}{2}$, the maximal mutual information is achieved when the constraint binds.

We proceed to characterize the maximal mutual information under randomized response, covering all equilibria:

Proposition 2. *Under randomized response, the maximal mutual information allowed by any equilibrium is*

$$\bar{I}_R(\rho) = \begin{cases} \frac{1}{2}[(1-\rho)\log(1-\rho) + (1+\rho)\log(1+\rho)], & \text{if } \rho \in (0, c), \\ 1 + \frac{1}{2}[(\frac{1}{\rho} - 1)\log(1-\rho) + \log \rho], & \text{if } \rho \in [c, 1), \end{cases}$$

where $c \approx 0.743$.

Proof. We prove the proposition by solving a constrained maximization problem. We first show that on the equilibrium path of any equilibrium, we must have

$$|\mu_s(y) - \mu_s(n)| \leq \rho. \tag{2}$$

Suppose, on the contrary, that $|\mu_s(y) - \mu_s(n)| > \rho$ on the equilibrium path. If $\mu_s(y) - \mu_s(n) > \rho$, then $\mu_s(n) > \rho - \mu_s(y)$ and $\rho - \mu_s(n) > -\mu_s(y)$. Regardless of whether it is q_s or q_t, both s and t strictly prefer to answer with n. This implies that $\mu_s(y)$ is not on the equilibrium path, a contradiction. If $\mu_s(y) - \mu_s(n) < -\rho$, then $\rho - \mu_s(y) > -\mu_s(n)$ and $-\mu_s(y) > \rho - \mu_s(n)$. Then, both s and t strictly prefer to answer with y, which again leads to the contradiction that $\mu_s(n)$ is not on the equilibrium path.

We maximize mutual information subject to (2). Since our objective is to find the maximal mutual information allowed by any equilibria under randomized response, it follows from Lemma 1 that for truthful equilibria we can focus on the case where $p_s = \frac{1-\rho}{2}$, which implies that $|\mu_s(y) - \mu_s(n)| = \mu_s(n) - \mu_s(y) = \rho$; for the other equilibria involving randomization, the indifference also requires that $|\mu_s(y) - \mu_s(n)| = \rho$. For our purpose, it is thus without loss of generality to consider that (2) binds.

The objective function is

$$\begin{aligned} I(\theta;r) = 1+ &\left[\frac{\Pr(y|s)+\Pr(y|t)}{2}\right](\mu_s(y)\log\mu_s(y) + [1-\mu_s(y)]\log[1-\mu_s(y)]) \\ +&\left[\frac{\Pr(n|s)+\Pr(n|t)}{2}\right](\mu_s(n)\log\mu_s(n) + [1-\mu_s(n)]\log[1-\mu_s(n)]). \end{aligned} \tag{3}$$

Note that as a function, (3) has six variables. We use the fact that these are probabilities to reduce the number of variables. First of all, by Bayes' rule, we have that

$$\mu_s(y) = \frac{\Pr(y|s)}{\Pr(y|s)+\Pr(y|t)} \Leftrightarrow \Pr(y|s)+\Pr(y|t) = \frac{\Pr(y|s)}{\mu_s(y)}, \tag{4}$$

$$\mu_s(n) = \frac{\Pr(n|s)}{\Pr(n|s)+\Pr(n|t)} \Leftrightarrow \Pr(n|s)+\Pr(n|t) = \frac{\Pr(n|s)}{\mu_s(n)}. \tag{5}$$

Substituting (4) and (5) into (3), we obtain

$$\begin{aligned} I(\theta;r) = 1+ &\left[\frac{\Pr(y|s)}{2\mu_s(y)}\right](\mu_s(y)\log\mu_s(y) + [1-\mu_s(y)]\log[1-\mu_s(y)]) \\ +&\left[\frac{\Pr(n|s)}{2\mu_s(n)}\right](\mu_s(n)\log\mu_s(n) + [1-\mu_s(n)]\log[1-\mu_s(n)])]. \end{aligned} \tag{6}$$

We use the fact that $\Pr(n|\cdot) = 1 - \Pr(y|\cdot)$ to further eliminate $\Pr(n|s)$ and $\mu_s(n)$. Note that (5) can be rewritten as

$$\mu_s(n) = \frac{1-\Pr(y|s)}{2-[\Pr(y|s)+\Pr(y|t)]} = \frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}, \tag{7}$$

where in the second equality we use (4) for $\Pr(y|s)+\Pr(y|t)$. Using (7) and the fact that $\frac{\Pr(n|s)}{2\mu_s(n)} = 1 - \frac{\Pr(y|s)}{2\mu_s(y)}$, (6) becomes

$$\begin{aligned} I(\theta;r) = 1+ &\left[\frac{\Pr(y|s)}{2\mu_s(y)}\right](\mu_s(y)\log\mu_s(y) + [1-\mu_s(y)]\log[1-\mu_s(y)]) \\ +&\left[1-\frac{\Pr(y|s)}{2\mu_s(y)}\right]\left[\left(\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\log\left(\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\right. \\ +&\left.\left(1-\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\log\left(1-\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\right]. \end{aligned} \tag{8}$$

Finally, we eliminate $\Pr(y|s)$ by using the binding (2). Without loss of generality, we consider the case where $\mu_s(n) > \mu_s(y)$ so that the constraint is $\mu_s(n) - \mu_s(y) = \rho$. Using

(7), the constraint becomes

$$\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)} - \mu_s(y) = \rho \Leftrightarrow \Pr(y|s) = \mu_s(y)\left(\frac{1}{\rho}\right)(2\left[\rho+\mu_s(y)\right]-1)\,. \tag{9}$$

Substituting (9) into (8), we obtain the following function in terms of $\mu_s(y)$ only:

$$\begin{aligned} I(\theta;r) =& \hat{I}(\mu_s(y)) \\ =& 1+\left(1-\frac{1}{2\rho}[1-2\mu_s(y)]\right)(\mu_s(y)\log\mu_s(y)+[1-\mu_s(y)]\log[1-\mu_s(y)]) \\ &+\left(\frac{1}{2\rho}[1-2\mu_s(y)]\right)[(\mu_s(y)+\rho)\log(\mu_s(y)+\rho) \\ &+(1-\mu_s(y)-\rho)\log(1-\mu_s(y)-\rho)]\,. \end{aligned} \tag{10}$$

Note that there are also the box constraints that $\mu_s(y) \in [0,1]$ and $\mu_s(n) \in [0,1]$. And given the binding (2), these box constraints are satisfied if and only if $\mu_s(y) \in [0, 1-\rho]$. Thus, our maximization problem is

$$\operatorname*{Max}_{\mu_s(y)\in[0,1-\rho]} \hat{I}(\mu_s(y)).$$

Note that $\hat{I}(\cdot)$ is symmetric at $\frac{1-\rho}{2}$, i.e., $\hat{I}\big(\frac{1-\rho}{2}+x\big) = \hat{I}\big(\frac{1-\rho}{2}-x\big)$.

The first-order condition for an extremum is

$$[1-2\rho-4\mu_s(y)]\ln\left(\frac{\mu_s(y)+\rho}{\mu_s(y)}\right) = [3-2\rho-4\mu_s(y)]\ln\left(\frac{1-\mu_s(y)-\rho}{1-\mu_s(y)}\right). \tag{11}$$

Equation (11) is satisfied at the point of symmetry, $\mu_s(y) = \frac{1-\rho}{2}$. The second derivative of $\hat{I}(\mu_s(y))$ is

$$\hat{I}''(\mu_s(y)) = \frac{\frac{1-2\mu_s(y)}{2(\mu_s(y)+\rho)(1-\mu_s(y)-\rho)} - \frac{1-2[\mu_s(y)+\rho]}{2\mu_s(y)[1-\mu_s(y)]} - 2\ln\Big(\Big[\frac{1-\mu_s(y)}{\mu_s(y)}\Big]\Big[\frac{\mu_s(y)+\rho}{1-\mu_s(y)-\rho}\Big]\Big)}{\rho\ln 2}.$$

It can be verified that

$$\hat{I}''(\tfrac{1-\rho}{2}) = \frac{4\left[\frac{\rho}{(1-\rho)(1+\rho)}+\ln\left(\frac{1-\rho}{1+\rho}\right)\right]}{\rho\ln 2} \gtreqless 0 \quad \text{for} \quad \rho \gtreqless d,$$

where $d \approx 0.796$. Thus, $\mu_s(y) = \frac{1-\rho}{2}$ corresponds to a local maximum for $\rho < d$ and a local minimum for $\rho > d$.

We further derive the third derivative:

$$
\begin{aligned}
\hat{I}'''(\mu_s(y)) = & \frac{1}{([\mu_s(y)][1-\mu_s(y)][\mu_s(y)+\rho][1-\mu_s(y)-\rho])^2 \ln 4} \\
& \times (1-2\mu_s(y)-\rho)\Big[2\rho^3[1-2\mu_s(y)] - 3\rho^2(1-2\mu_s(y)[1-\mu_s(y)]) \\
& + \rho(1-2\mu_s(y)(2-\mu_s(y)[3-2\mu_s(y)]) \\
& + 2\mu_s(y)[1-\mu_s(y)](1-\mu_s(y)[1-\mu_s(y)])\Big].
\end{aligned} \tag{12}
$$

We evaluate the values of the third derivative for $\rho \in [0,1)$, which in turns allows us to infer the properties of the second derivative and to establish the global maxima of the objective function.

Solving $\hat{I}'''(\mu_s(y)) = 0$ gives three real solutions:

$$\hat{\mu}_s(y) = \frac{1}{2}\left(1-\rho-\sqrt{2\sqrt{4\rho^4-\rho^2+1}-3\rho^2-1}\right), \tag{13}$$

$$\bar{\mu}_s(y) = \frac{1-\rho}{2}, \tag{14}$$

$$\tilde{\mu}_s(y) = \frac{1}{2}\left(1-\rho+\sqrt{2\sqrt{4\rho^4-\rho^2+1}-3\rho^2-1}\right). \tag{15}$$

We first consider $\rho \in [0, \frac{1}{2}]$. Note that for $\rho \in [0, \frac{1}{2}]$, $\bar{\mu}_s(y) = \frac{1-\rho}{2}$ in (14) is the only point in $(0, 1-\rho)$ at which the third derivative vanishes. Evaluating the expression in (12) for $\rho \in [0, \frac{1}{2}]$ then gives that $\hat{I}'''(\mu_s(y)) \gtreqless 0$ for $\mu_s(y) \lesseqgtr \frac{1-\rho}{2}$. And for $\rho \in [0, \frac{1}{2}]$, $\lim_{\mu_s(y)\to 0} \hat{I}''(\mu_s(y)) = \lim_{\mu_s(y)\to 1-\rho} \hat{I}''(\mu_s(y)) = -\infty$. Accordingly, with $\frac{1}{2} < d$, for $\rho \in [0, \frac{1}{2}]$, $\hat{I}''(\mu_s(y)) \leq \hat{I}''(\frac{1-\rho}{2}) < 0$ for all $\mu_s(y) \in [0, 1-\rho]$. Thus, $\hat{I}(\mu_s(y))$ is strictly concave on $[0, 1-\rho]$ for $\rho \in [0, \frac{1}{2}]$, and $\mu_s(y) = \frac{1-\rho}{2}$ corresponds to a global maximum for $\rho \in [0, \frac{1}{2}]$.

We consider next $\rho \in [\sqrt{3/7}, 1]$. Note that for $\rho \in (\sqrt{3/7}, 1)$, $\bar{\mu}_s(y) = \frac{1-\rho}{2}$ in (14) is the only point in $[0, 1-\rho]$ at which the third derivative vanishes. And for $\rho \in \{\sqrt{3/7}, 1\}$, the three solutions in (13)-(15) coincide. Evaluating the expression in (12) for $\rho \in [\sqrt{3/7}, 1]$ then gives that $\hat{I}'''(\mu_s(y)) \gtreqless 0$ for $\mu_s(y) \gtreqless \frac{1-\rho}{2}$. Accordingly, with $\sqrt{3/7} < d$, for $\rho \in (d, 1]$, $\hat{I}''(\mu_s(y)) \geq \hat{I}''(\frac{1-\rho}{2}) > 0$ for all $\mu_s(y) \in [0, 1-\rho]$. Thus, $\hat{I}(\mu_s(y))$ is strictly convex on $[0, 1-\rho]$ for $\rho \in (d, 1]$, and the global maxima lie at, given the symmetry at $\frac{1-\rho}{2}$, the two boundaries, $\mu_s(y) = 0$ or $\mu_s(y) = 1-\rho$.

We further divide the remaining case $\rho \in (\frac{1}{2}, d]$ into two sub-cases, where $\rho \in (\frac{1}{2}, \sqrt{3/7})$ and where $\rho \in [\sqrt{3/7}, d]$. We consider the latter case first. It follows from the above that for $\rho \in [\sqrt{3/7}, d]$, we have that $\hat{I}''(\mu_s(y)) \geq \hat{I}''(\frac{1-\rho}{2})$ for all $\mu_s(y) \in [0, 1-\rho]$. Given the symmetry of $\hat{I}(\mu_s(y))$, we without loss of generality focus on its behavior for $\mu_s(y) \in [0, \frac{1-\rho}{2}]$. Note that for $\rho \in [\sqrt{3/7}, d]$, $\lim_{\mu_s(y)\to 0} \hat{I}''(\mu_s(y)) = \infty$ and recall that for $\rho \leq d$, $\hat{I}''(\frac{1-\rho}{2}) \leq 0$. Given that for $\rho \in [\sqrt{3/7}, 1]$, $\hat{I}'''(\mu_s(y)) < 0$ for $\mu_s(y) < \frac{1-\rho}{2}$, there exists a unique $k \in (0, \frac{1-\rho}{2}]$ such that $\hat{I}''(k) = 0$. This further implies that there is at most one point in $(0, \frac{1-\rho}{2})$ such that the first-order condition is satisfied, in which case it corresponds to a local minimum; $\mu_s(y) = \frac{1-\rho}{2}$ thus corresponds to a unique local maximum. Given that, for $\rho \in [\sqrt{3/7}, d]$, $\hat{I}(\mu_s(y))$ is strictly convex for $\mu_s(y)$ sufficiently close to zero and concave (strictly concave for $\rho < d$) in the neighborhood of $\frac{1-\rho}{2}$, the global maximum is achieved either at the unique local maximum at $\mu_s(y) = \frac{1-\rho}{2}$ or at the boundary $\mu_s(y) = 0$ or, by symmetry, $\mu_s(y) = 1 - \rho$.

Finally, we consider $\rho \in (\frac{1}{2}, \sqrt{3/7})$. Note that for $\rho \in (\frac{1}{2}, \sqrt{3/7})$, the solution in (13) satisfies that $\hat{\mu}_s(y) \in (0, \frac{1-\rho}{2})$ and the solution in (15) satisfies that $\tilde{\mu}_s(y) \in (\frac{1-\rho}{2}, 1-\rho)$. Similar to the above paragraph, the following argument focuses on $\mu_s(y) \in [0, \frac{1-\rho}{2}]$ under the symmetry. Evaluating the expression in (12) gives that $\hat{I}'''(\mu_s(y)) \leq 0$ for $\mu_s(y) \leq \hat{\mu}_s(y)$ and $\hat{I}'''(\mu_s(y)) > 0$ for $\mu_s(y) \in (\hat{\mu}_s(y), \frac{1-\rho}{2})$. Recall that for ρ in this range, we have that $\hat{I}''(\frac{1-\rho}{2}) < 0$. Then, the fact that $\hat{I}'''(\mu_s(y)) > 0$ for $\mu_s(y) \in (\hat{\mu}_s(y), \frac{1-\rho}{2})$ implies that $\hat{I}''(\mu_s(y)) < 0$ for $\mu_s(y) \in (\hat{\mu}_s(y), \frac{1-\rho}{2})$. Note that for $\rho \in (\frac{1}{2}, \sqrt{3/7})$, $\lim_{\mu_s(y)\to 0} \hat{I}''(\mu_s(y)) = \infty$. Thus, given that $\hat{I}'''(\mu_s(y)) \leq 0$ for $\mu_s(y) \leq \hat{\mu}_s(y)$, there exists a unique $v \in (0, \hat{\mu}_s(y)]$ such that $\hat{I}''(v) = 0$. The argument from the above paragraph then applies to establish that the global maximum is again achieved either at the unique local maximum at $\mu_s(y) = \frac{1-\rho}{2}$ or at the boundary $\mu_s(y) = 0$ or, by symmetry, $\mu_s(y) = 1 - \rho$.

Substituting $\mu_s(y) = \frac{1-\rho}{2}$ into (10), we obtain $\frac{1}{2}[(1-\rho)\log(1-\rho) + (1+\rho)\log(1+\rho)]$, which is the mutual information of the truthful equilibrium; substituting $\mu_s(y) = 0$ or $\mu_s(y) = 1-\rho$ into (10) and using $0 \log 0 = 0$, we obtain $1 + \frac{1}{2}[(\frac{1}{\rho} - 1)\log(1-\rho) + \log \rho]$, which is the mutual information of the informative equilibrium under direct response replicable under randomized response. The result follows from the fact that $c < d$, where $c \approx 0.743$ is the critical value in Lemma 1.

□

The essence behind Proposition 2 is that the two values of mutual information in Lemma 1 form an upper envelope of the mutual information of all equilibria under randomized response. The following corollary, which compares the maximal mutual information under the two responses, is immediate:

Corollary 1. *For given $\rho \in (0, 1)$, the maximal mutual information under randomized response weakly dominates that under direct response, with strict dominance for $\rho \in (0, c)$, where $c \approx 0.743$.*

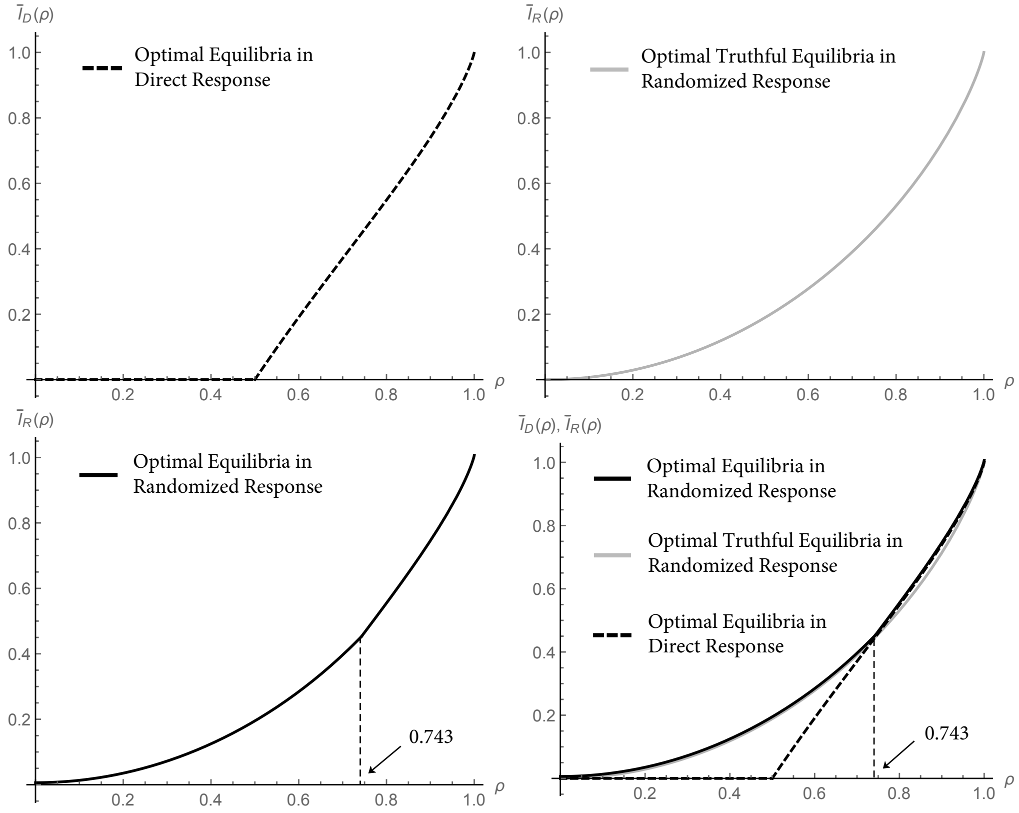

Figure 1: Maximal Mutual Information

Figure 1 summarizes the above results. The upper left-hand panel shows that the maximal mutual information achievable with direct response is zero for lower relative truth-telling preferences, $\rho < \frac{1}{2}$, becomes positive for moderate relative truth-telling preferences, $\rho > \frac{1}{2}$, and increases to one as $\rho \to 1$. The upper right-hand panel shows that, in contrast, the maximal mutual information achievable with truthful randomized response equilibria is strictly positive regardless of the relative truth-telling preference and thus improves on direct response for low relative truth-telling preferences. This is consistent with the rationale for using a garbling scheme like randomized response: with low to moderate relative truth-telling preferences we cannot expect any information transmission under direct response, whereas with randomized response some information can be transmitted regardless of the truth-telling incentives, as long as there is some preference for truth-telling.

While direct response is dominated by randomized response for lower relative truth-telling preferences, the same is not the case for moderate preferences, as shown in the two lower panels. Below $\rho \approx 0.743$ randomized response dominates direct response and the maximal mutual information is achieved with a truthful randomized response equilibrium. Beyond that value randomized response and direct response are tied and the optimal truth-telling randomized response equilibrium is dominated by the optimal direct response equilibrium (and also an informative non-truth-telling randomized response equilibrium).

In summary, consistent with the rationale for garbling in general and for randomized response specifically, for low to moderate values of ρ (below 0.743), more information can be obtained with randomized response than with direct response. For higher values of ρ, there is no loss from using direct response, as long as the equilibrium strategy, or equivalently the lying behavior of the sender, is known.

References

[1] Cover, Thomas M., and Joy A. Thomas [1991], *Elements of Information Theory*, John Wiley and Sons: New York, NY.

[2] Dessein, Wouter, Andrea Galeotti, and Tano Santo [2013] "Rational Inattention and Organizational Focus." Columbia University Working Paper.

[3] Donaldson-Matasci, Matina C., Carl T. Bergstrom, and Michael Lachmann [2010], "The Fitness Value of Information," *Oikos* **119** 219-230.

[4] Jose, Victor Richmond R., Robert F. Nau, and Robert L. Winkler [2008], "Scoring Rules, Generalized Entropy, and Utility Maximization," *Operations Research* **56**, 1146-1157.

[5] Kelly, J. L. [1956] "A New Interpretation of Information Rate," *Bell System Tech. J.* **35**, 917-926.

[6] Ljungqvist, Lars [1993], "A Unified Approach to Measures of Privacy in Randomized Response Models: A Utilitarian Perspective," *Journal of the American Statistical Association* **88**, 97-103.

[7] Shannon, Claude [1948], "A Mathematical Theory of Communication," *Bell System Technical Journal* **27**, 379-423, 623-656.

[8] Sims, Christopher A. [2003], "Implications of Rational Inattention," *Journal of Monetary Economics* **50**, 665-690.