Decision-aid or Controller? Steering Human Decision Makers with Algorithms

Ruqing Xu, Sarah Dean

March 3, 2023

Vignettes

Consider an algorithm-aided lending decision. The algorithm takes in the features of a loan applicant (e.g. race, age, income, past credit history) and recommends a repayment probability to the loan manager. The loan manager receives the recommendation, also observes some features of the applicant, and then makes a decision (accept/reject; interest rate).

Racial bias

Suppose that we learn from the loan manager's past decisions that he is unjustifiably lenient to the applicants of his own racial group.

Cognitive bias

As most humans, a loan manager systematically overweighs small probabilities of default and makes lending decisions that are too conservative for the bank to maximize its profit.

Motivation

Data $\rightarrow$ Algorithm $\rightarrow$ Recommendation $\rightarrow$ Human DM $\rightarrow$ Decision

Motivation

- Until recently, the majority of ML literature is only concerned with things in this dashed box.
- What matters to the society more is the quality and fairness of final decisions.
- In this structure, can the algorithm have control over final decisions with recommendations alone?
- Can the algorithm learn about the human DM and adjust its recommendations accordingly?

Related Literature

Literature on human-machine complementarity focuses on designing optimal rules to combine the independent decisions of the algorithm and the human into a joint decision (Rastogi et al. 2022, Donahue et al. 2022).

Our work differs from this literature in two ways.

- Human remains the (only) final decision maker in the system while the algorithm merely acts as an "advisor."
- "Different information" versus "different objectives."

Contribution and Limitations

- We provide a formal model of the interaction between decision-aid algorithms and human DMs...
- ...in a noiseless setting where the algorithm knows the ideal decision for each case, the human DM has no private information, and the algorithm recommends a continuous-valued recommendation.
- Our results can be interpreted as the "fundamental limit" of controllability and identifiability and serves as a benchmark.
- Question for the audience: which extensions are feasible and economically important? In extended scenarios, what should the algorithm do?

Problem Setting

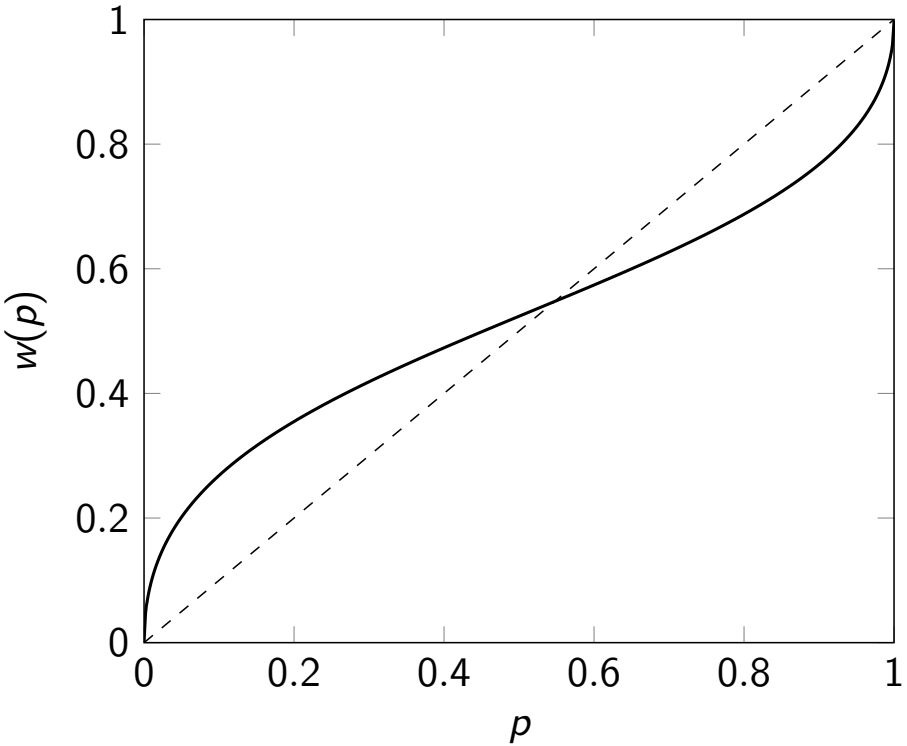
- Individual, algorithm, judge.
- For each individual i, the algorithm observes high-dimensional features $x_i \in \mathcal{X} \subset \mathbb{R}^D$.
- The judge can only process (or access) a subset of these features denoted by $x_i' = A(x_i)$. We call $A : \mathcal{X} \to \{0,1\}^D \odot \mathcal{X}$ the "attention function," where $\odot$ denotes entry-wise multiplication.
- The ideal decision for individual i is $m_i = f(x_i)$. We assume that this is known to the algorithm.
- After observing features x_i, the algorithm makes a recommendation n_i to the judge.
- The judge makes a final decision y_i according to the decision rule $y_i = g(n_i, x_i')$.
- WLOG we normalize n, m, y to be in $[0,1]$.

Informational Structure

- The algorithm knows everything that the judge knows: A, x_i'. In addition, the algorithm knows x_i, f, and thus m_i.
- When the algorithm knows judge's decision rule $g(n_i, x_i')$
  $\rightarrow$ controllability under perfect information.
- When the algorithm does not know judge's decision rule $g(n_i, x_i')$
  $\rightarrow$ identifiability of g from data.

Linear in log odds model

Linear in log odds (LLO) is a well-known model in behavioral economics to capture people's non-linear perception of probabilities.



Algorithmic recommendations can be viewed as a probability in many scenarios. Therefore, it is reasonable to assume that the judge's decision function is LLO in the recommendation n.

Linear in log odds model

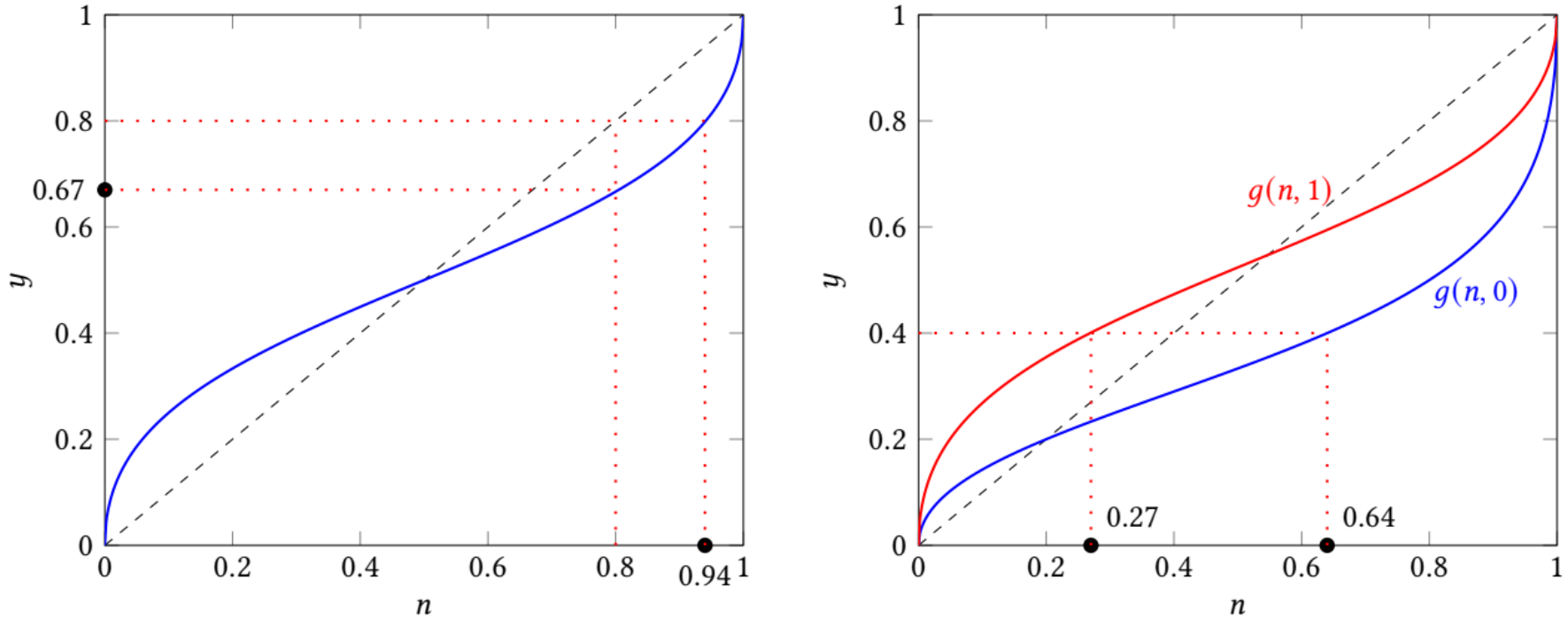


Figure: Left panel: Distortion of probabilities. Right panel: Biases based on group membership.

Controllability

Definition (Full control)

The algorithm attains full control of the judge at $x \in \mathcal{X}$ if there exists an $n \in [0, 1]$ such that $g(n, A(x)) = f(x)$.

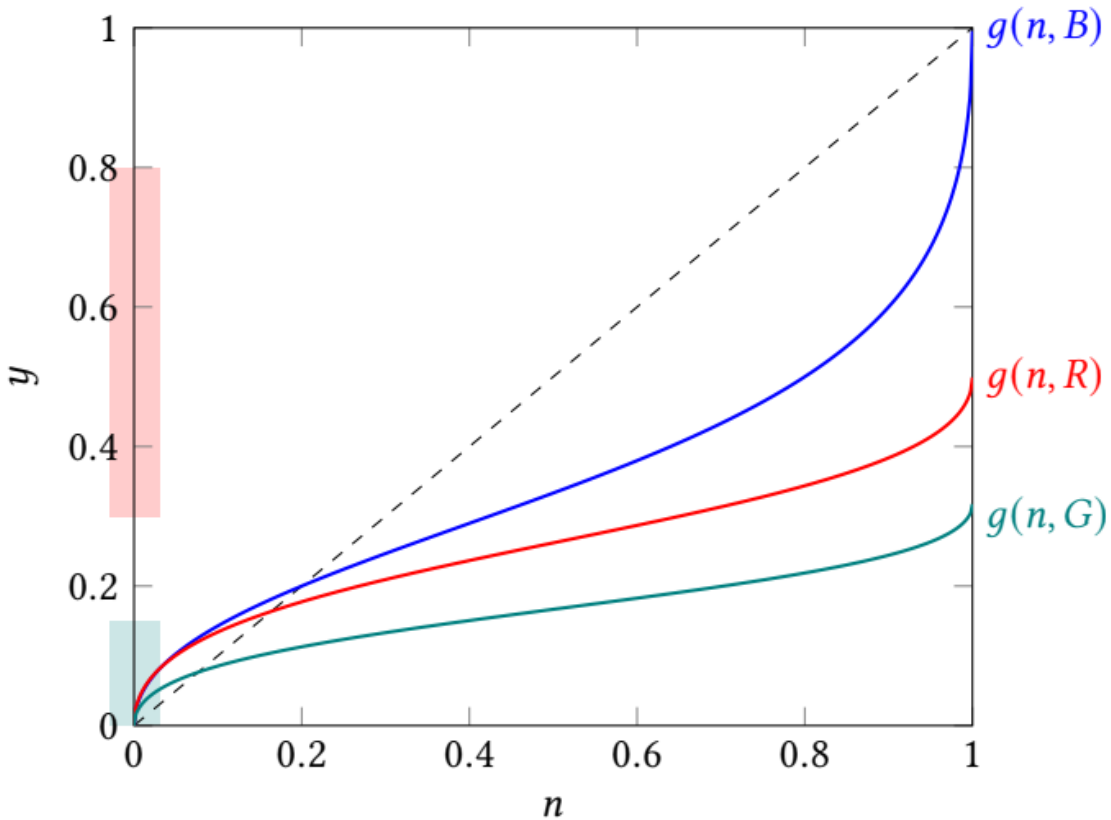


Figure: Full control with full range. Failing to control. Full control with limited range.

Parameter Identification

Question

If the judge's decision function is unknown but takes a parametric form with parameters $\beta \in \mathbb{R}^{D'}$, $\gamma \in \mathbb{R}$ and the form of h known:

$$g(n, x') = h(\beta^\top x', \gamma, n)$$

Can we identify β and γ from the dataset of past interactions:

$$\{x_i, x'_i, n_i, y_i\}_{i=1}^N \text{ where } y_i = h(\beta^\top x'_i, \gamma, n_i)$$

Parameter Identification

Question

If the judge's decision function is unknown but takes a parametric form with parameters $\beta \in \mathbb{R}^{D'}$, $\gamma \in \mathbb{R}$ and the form of h known:

$$g(n, x') = h(\beta^\top x', \gamma, n)$$

Can we identify β and γ from the dataset of past interactions:

$$\{x_i, x_i', n_i, y_i\}_{i=1}^N \text{ where } y_i = h(\beta^\top x_i', \gamma, n_i)$$

Answer

Yes! As long as two general conditions are satisfied:
Rank condition: $\text{rank}(X) = D$, standard assumption in linear regression.
Independence condition: Need independent variation in x_i' and n. (If we observe two individuals with $x_i' = x_j'$ but $n_i \neq n_j$, this is satisfied.)

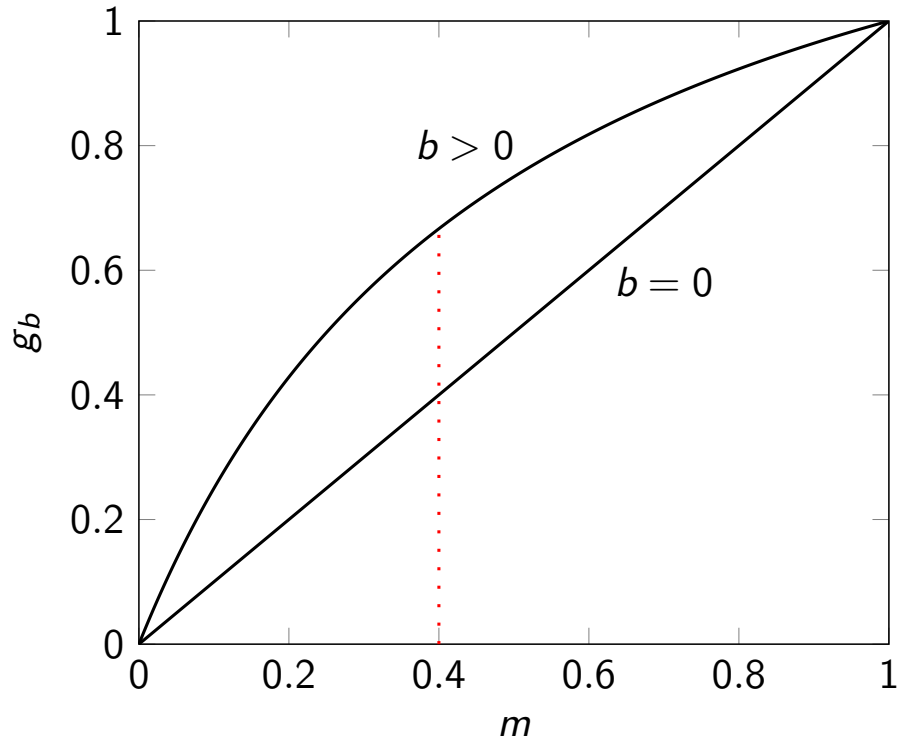
A Strategic Judge

A strategic judge is able to adapt his decision rule to best respond to the algorithm.

- As before, x_i, x_i', and m_i, the "true state."
- The algorithm has a utility function $U^A(y, m) = -(y - m)^2$, i.e., the algorithm wants the decision to be close to the true state.
- The judge has a utility function $U^J(y, m, b) = -(y - g_b(m))^2$, i.e., the judge wants the decision to be close to his "ideal decision function" $g_b(m)$ if the true state is m.

A Strategic Judge

When $b = 0$, $g_b(m)$ is the identity line (= the ideal decision function of the algorithm). When $b > 0$, $g_b(m)$ shifts away from the identity line.



The Sequential Game

We first consider a sequential game where the algorithm and the judge take turns to react.

1. $T = 1$, the algorithm is constrained to behave "truthfully," i.e., $n = m$, while the judge best responds according to his utility.
2. $T = 2$, the judge's decision rule is fixed as in $T = 1$ and the algorithm changes its recommendation rule in best response.
3. $T = 3$, the algorithm's recommendation rule is fixed as in $T = 2$ and the judge changes his decision rule in best response.
4. ...so on and so forth.

The Sequential Game

At $T = 1$, the algorithm is mandated to provide the true state of the world. The judge's best response is simply $y = g_b(n)$.

At $T = 2$, the algorithm learns about the decision rule of the judge in the first period and best responds by choosing $n = g_b^{-1}(m)$. Since under this recommendation, the final decision of the judge at $T = 2$ is

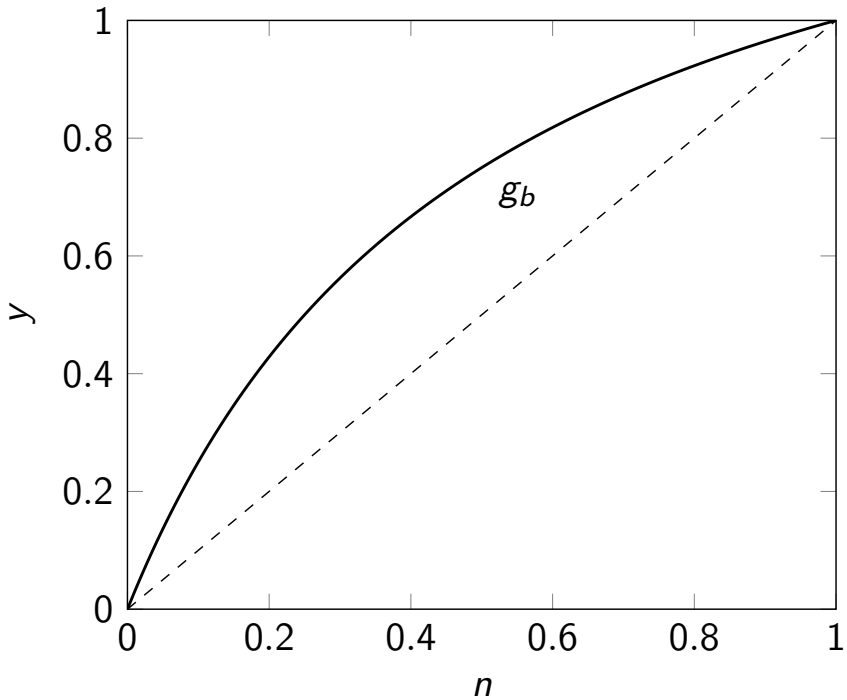
$$y = g_b(n) = g_b(g_b^{-1}(m)) = m$$

At $T = 3$, the algorithm's recommendation rule is fixed at $n = g_b^{-1}(m)$. This function is invertible, so the judge deduces that $m = g_b(n)$. Therefore, the judge's decision function is

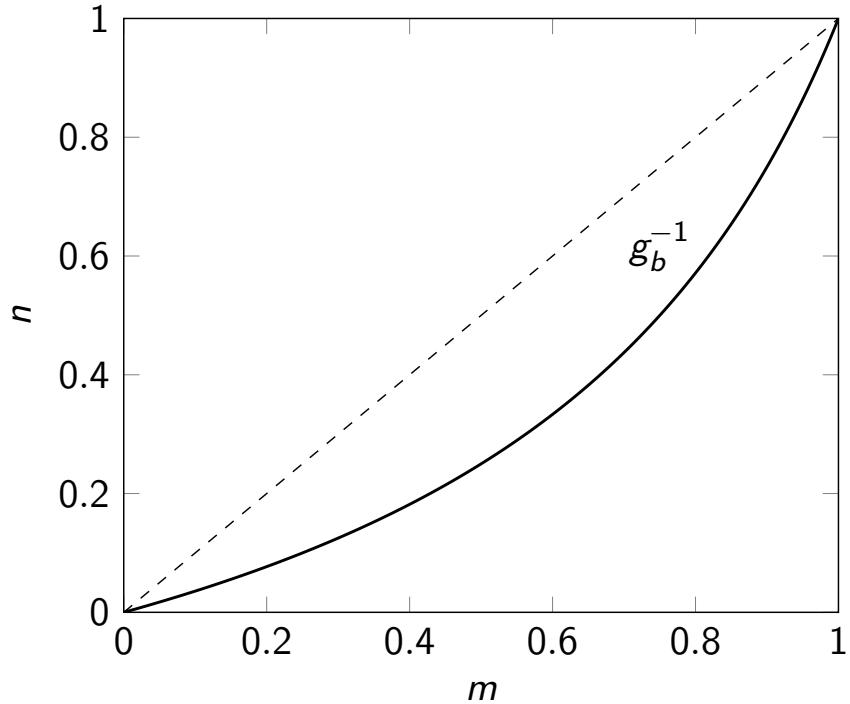
$$y = g_b(m) = g_b(g_b(n)) = g_b^2(n)$$

Note that g_b^2 is a more distorted curve, i.e., further away from the identity line.

The Sequential Game

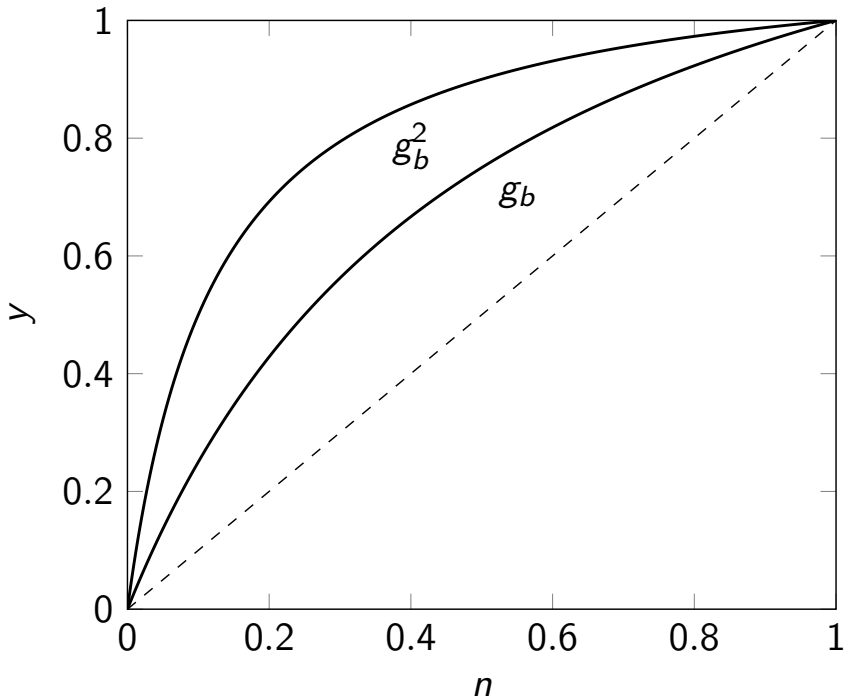


(a) Judge's strategy

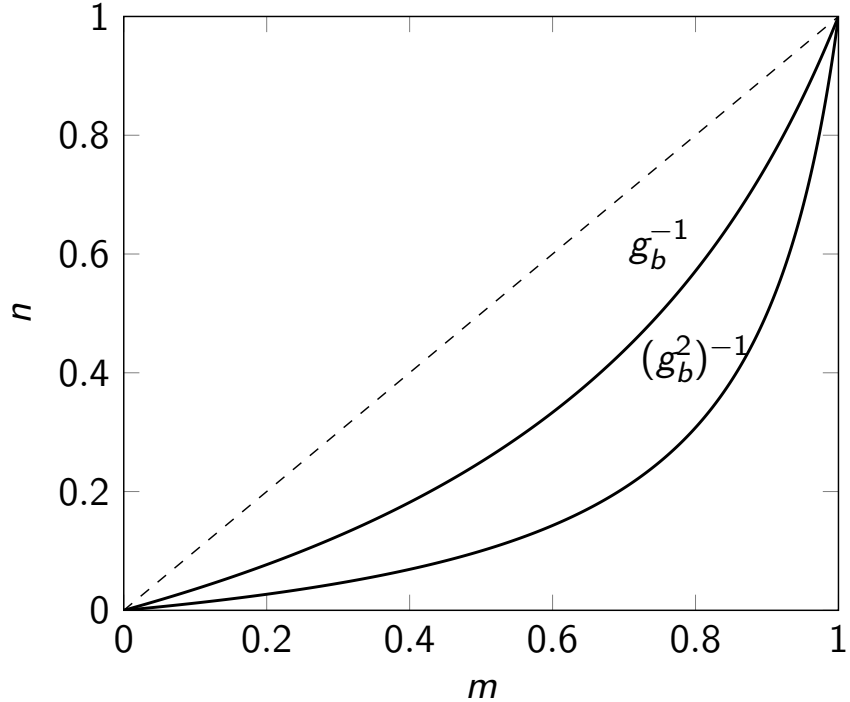


(b) Algorithm's strategy

The Sequential Game

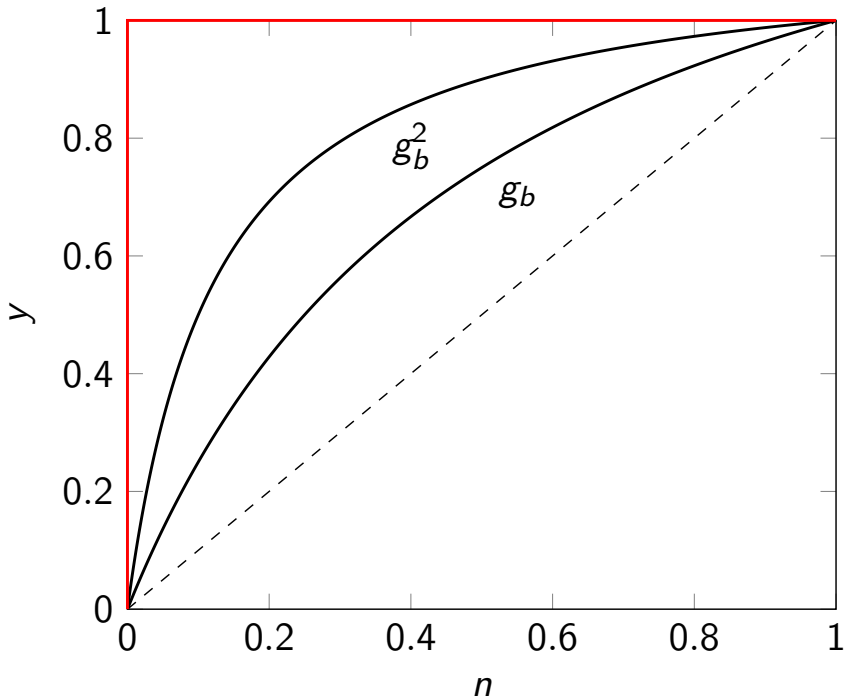


(a) Judge's strategy

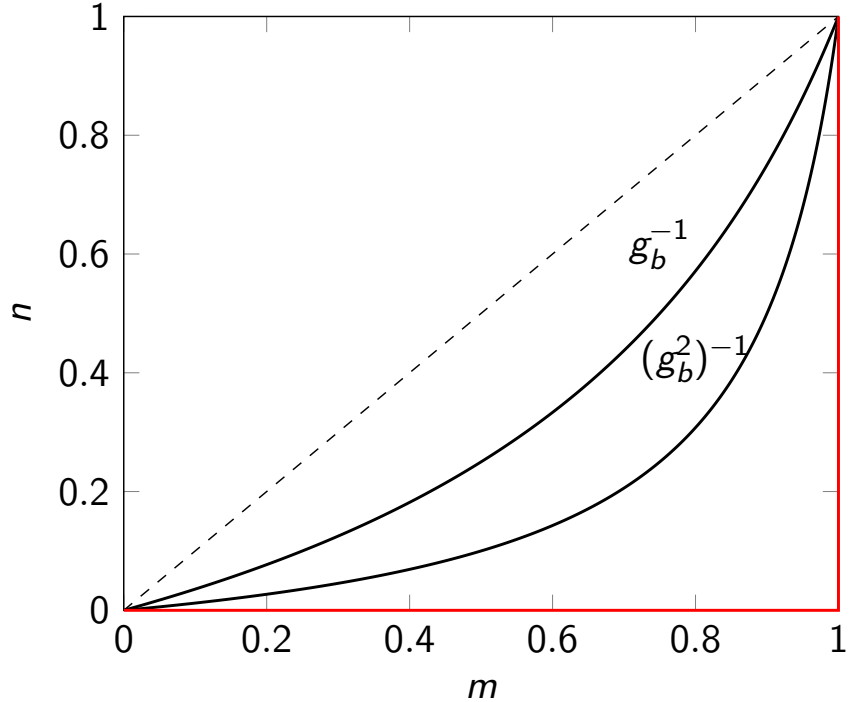


(b) Algorithm's strategy

The Sequential Game



(a) Judge's strategy



(b) Algorithm's strategy

The Simultaneous Game

At the limit, more like a simultaneous game where each player anticipates the other player's strategy and best responds in the same period. We call this an equilibrium.

One insight we can glean from the sequential game is that the algorithm should not choose a one-to-one correspondence between m and n in equilibrium. Otherwise, the judge can perfectly back out the true state of the world and achieve his first-best.

Indeed, with reference to the classical cheap talk game in Crawford & Sobel (1982), we show that all equilibria in our game are "partition equilibria," where the algorithm makes the same recommendation for an interval of true states.

The Simultaneous Game

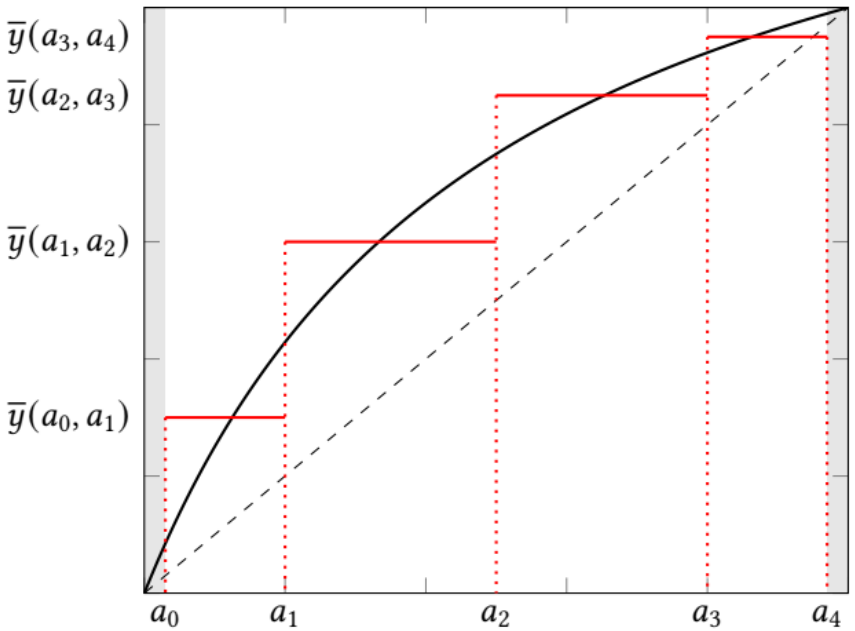


Figure: Illustrations of a partition equilibrium with four intervals. The algorithm sends the same recommendation for the states of the world in the same partition interval, and the judge chooses the decision that maximizes the expected utility knowing that the true state falls into this interval.

The algorithm says "the true state is between a_1 and a_2" even if it knows exactly where the true state is.

The Simultaneous Game

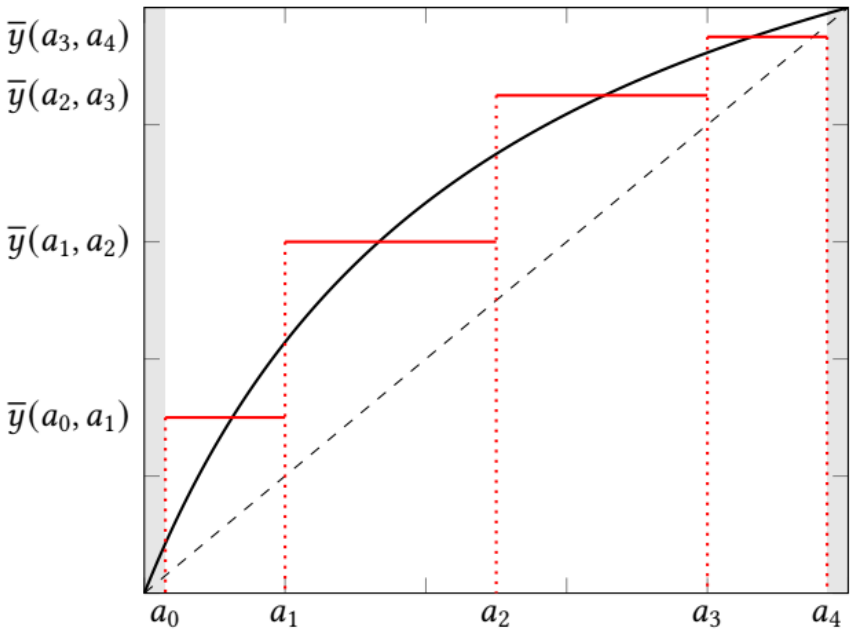


Figure: Illustrations of a partition equilibrium with four intervals. The algorithm sends the same recommendation for the states of the world in the same partition interval, and the judge chooses the decision that maximizes the expected utility knowing that the true state falls into this interval.

The algorithm says "the true state is between a_1 and a_2" even if it knows exactly where the true state is.

Wait, isn't this what algorithms have been doing in real-life...?

Discussion and Extensions

- We characterize the fundamental limits of algorithmic influence in decision-aid systems.
- The interpretation of these results can be two-fold.
- $(+)$ Correct human bias and help better decision making.
- $(-)$ "Accountability laundering." It is fairly easy for an algorithm to go from a decision-aid to a controller. If the algorithm is secretly used for such purposes, the human judges are held accountable for their decisions, while the goals of the algorithm's designer are realized.
- When viewed in this light, our results help to understand the extent to which we should worry about such misuses of algorithms.

References I

Crawford, V. P. & Sobel, J. (1982), 'Strategic information transmission', *Econometrica: Journal of the Econometric Society* pp. 1431–1451.

Donahue, K., Chouldechova, A. & Kenthapadi, K. (2022), Human-algorithm collaboration: Achieving complementarity and avoiding unfairness, *in* '2022 ACM Conference on Fairness, Accountability, and Transparency', FAccT '22, Association for Computing Machinery, New York, NY, USA, p. 1639–1656.
URL: *https://doi.org/10.1145/3531146.3533221*

Rastogi, C., Leqi, L., Holstein, K. & Heidari, H. (2022), 'A unifying framework for combining complementary strengths of humans and ML toward better predictive decision-making', *CoRR* **abs/2204.10806**.
URL: *https://doi.org/10.48550/arXiv.2204.10806*