Learning to balance fairness and self-interest: a reinforcement-learning account

Griffin M*, **Lebreton M*,** Gross J, & de Dreu CKW

Leiden University & University of Amsterdam

Fairness & Ultimatum game

1. Proposer gets an Initial endowment M

2. Proposer makes an offer $\mathbf{x} \in [0 - M]$

3. Receiver makes a decision $\boldsymbol{A}$ to Accept (1) or Rejects (0)the offer.
If Accepts:
- P gets $M - \mathbf{x}$
- R gets $\mathbf{x}$

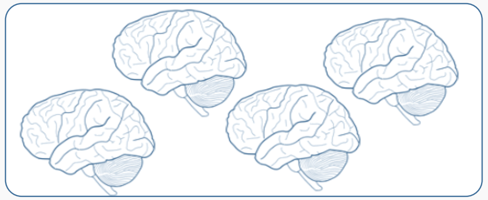
If Rejects
- P gets 0
- R gets 0

<u>Self-interest</u>: keep $\mathbf{x}$ as small as possible
- keep enough of the endowment M

<u>Fairness norm</u>: make $\mathbf{x}$ big enough
- Morally acceptable
- Offer do not get rejected

Fairness can be ambiguous, e.g. in
- different populations or
- different contexts

where different fairness norms prevail, but no repeated-interactions with specific individuals



Hypotheses

When the **fairness norm is ambiguous** (e.g. interacting with individuals from different populations or in new contexts) individual can ***learn fairness norms by trial-and-errors***, so as to propose offers that balance self-interest and compliance to norm.

To do so, individuals form ***expectations*** about the ***probability of individuals to accept*** offers, which are revised according to ***observed behavior***, via ***prediction-error correction mechanism*** (a.k.a. delta-rule)

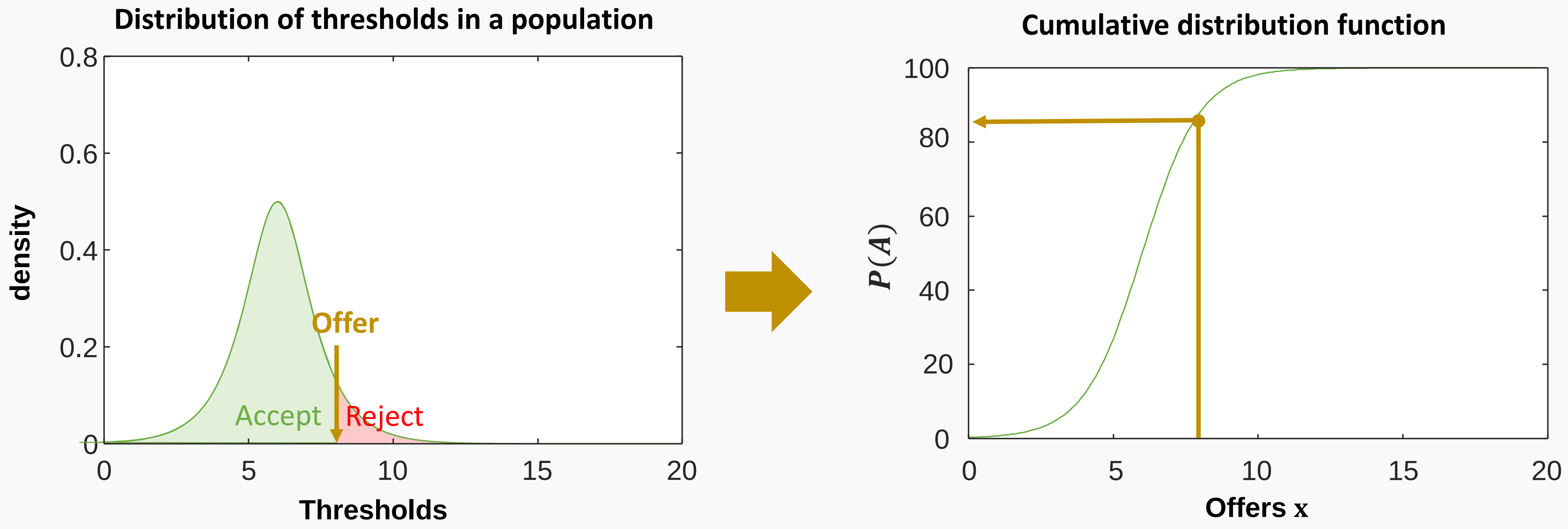
Prior norms, i.e. anterior to learning, ***bias learning***, making individual sub-optimal in certain situations.

Learning is impacted by the social context, and knowledge about those contexts.

What is fairness?

Individuals in a population have (hard or soft) “threshold”, which determine whether they accept or reject an offer.

Fairness norm: make an offer that would be considered acceptable by “enough” individuals



Core idea

Decision of the receivers is governed by a function g, indexing the probability of accepting an offer $\mathbf{x}$, depending on some true underlying "population" parameters θ

$$P(\boldsymbol{A}_t|,\mathbf{x}_t) = g(\theta, \mathbf{x}_t)$$

The proposer learn by trial-and error the parameters $\hat{\theta}$ of a function f predicting/estimating the receiver's decisions.

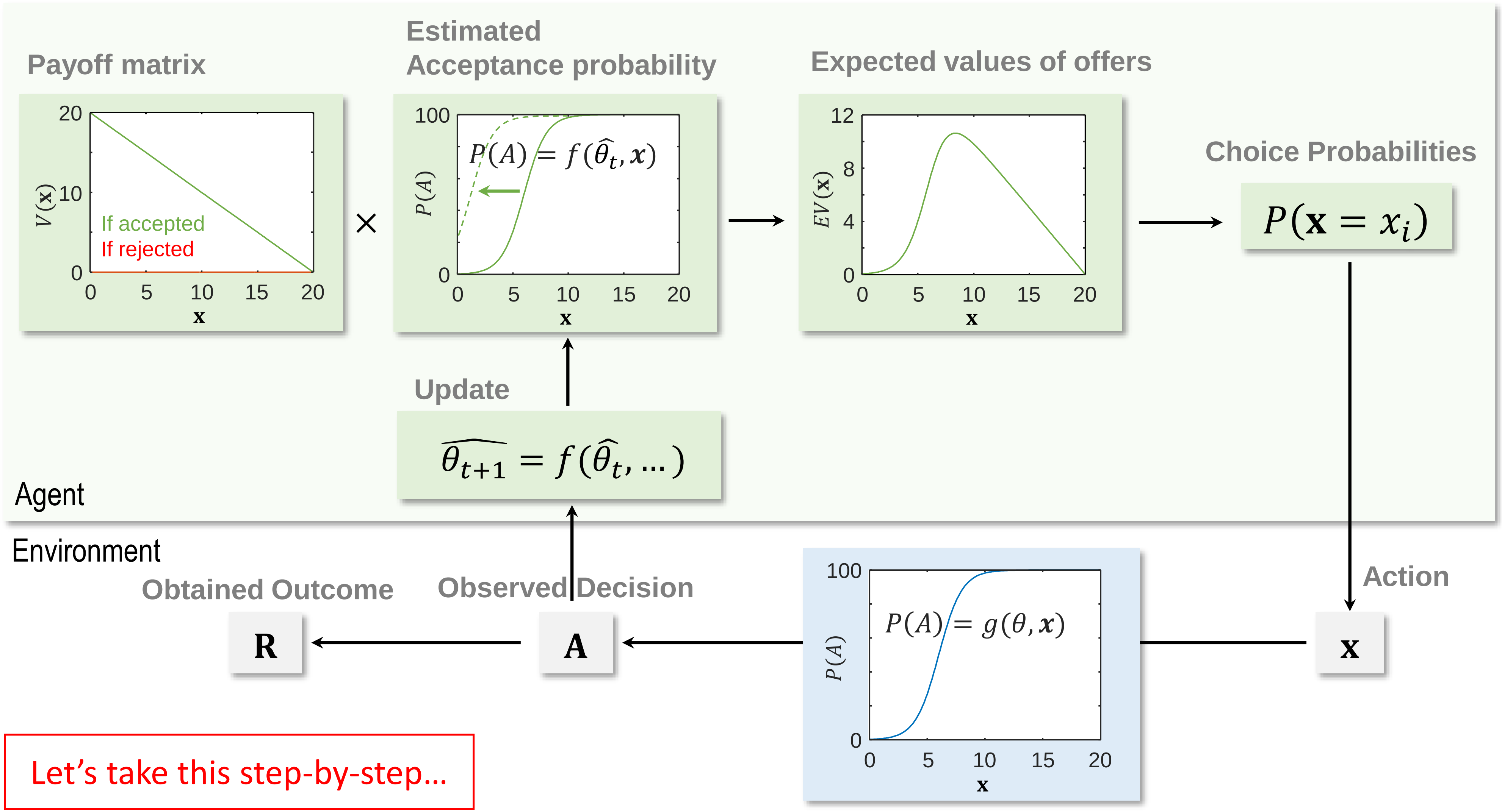
$$P(A_t) = f\left(\widehat{\theta_t}, \mathbf{x}_t\right)$$

At each trial, s/he observes a receiver's decision to an offer, and updates his current parameter estimate $\widehat{\theta_t}$, so as to ultimately make the offer x that which offers the best trade-off between self-interest and the (unknown) fairness norm in the considered population

Outline

1. **Computational framework**
2. **Experimental framework (behavior)**
3. **Results: Experiment 1**
4. **Results: Experiment 2**

Computational Framework: General

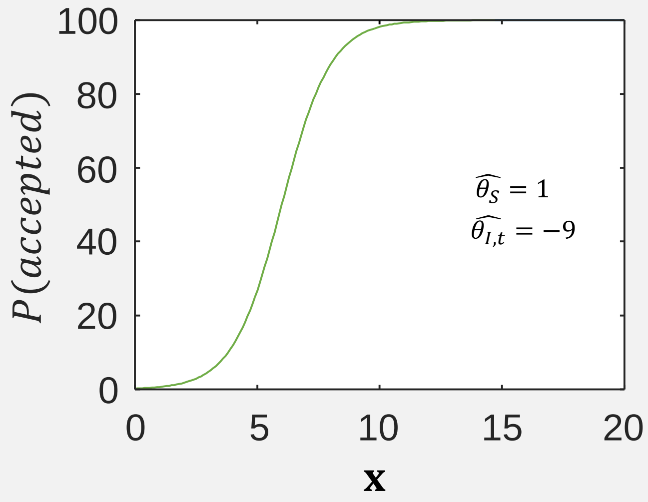


Let's take this step-by-step…

Estimated acceptance probability

Functional form: logistic function

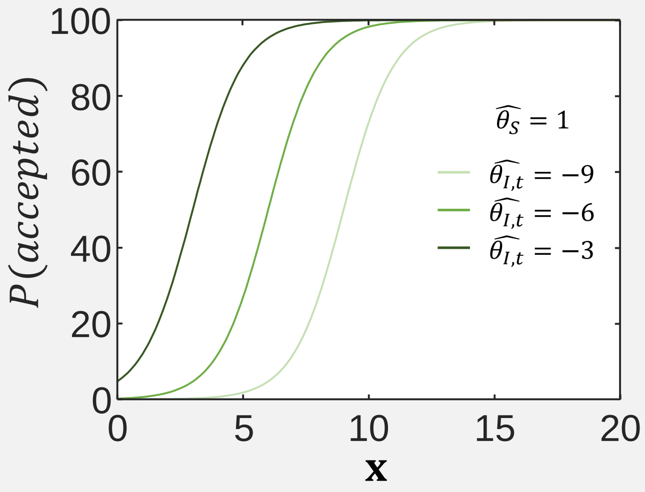
$$P_t(A) = \frac{1}{1 + \exp(-(\widehat{\theta_{I,t}} + \widehat{\theta_S}\mathbf{x}))}$$



Fixed and variable parameters

$\widehat{\theta_{I,t}}$: intercept at time t

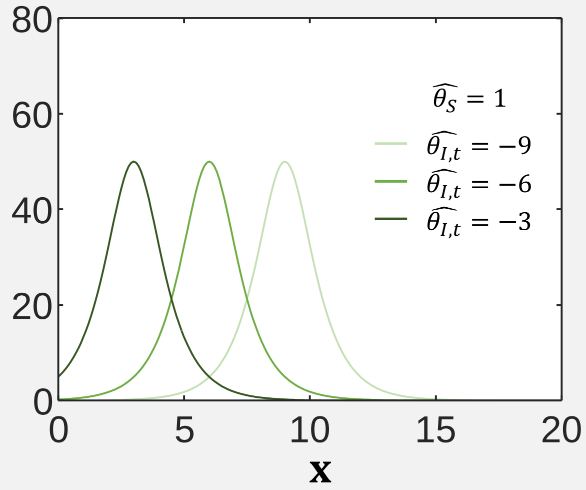
$\widehat{\theta_S}$: slope; fixed



Interpretation (1)

Cumulative distribution function of the logistic distribution (~ normal distribution with heavier tails)

$$f(x, \widehat{\theta_{I,t}}, \widehat{\theta_S}) = \frac{\exp(-(\widehat{\theta_{I,t}} + \widehat{\theta_S}\mathbf{x}))}{\frac{1}{\widehat{\theta_S}}(1 + \exp(-(\widehat{\theta_{I,t}} + \widehat{\theta_S}\mathbf{x}))^2}$$



Interpretation (2)

Norms ~ what can be expected to be accepted in a population distribution.

Formal but intuitive definition of fairness norm !!

Expected value

Expected value of an offer

Proposer makes an offer x and gets

- $M - \mathbf{x}$ if accepted
- $\mathbf{0}$ if rejected

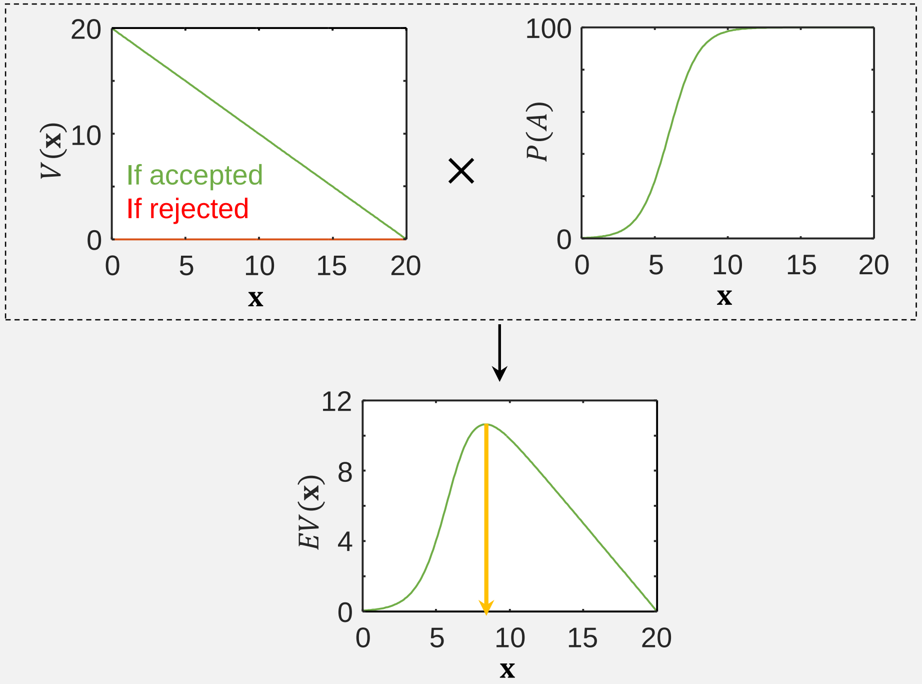
Therefore, $\forall x \in [0\ M]$,

- $EV(x) = (M - x)P(A = 1) + 0 * P(A = 0)$
- $EV(x) = (M - x)P(A = 1)$

With $P_t(A) = \frac{1}{1+\exp(-(\widehat{\theta_{I,t}}+\widehat{\theta_S}\mathbf{x}))}$, (see preceding slide) we get

- $EV(x) = \frac{M-x}{1+\exp(-(\widehat{\theta_{I,t}}+\widehat{\theta_S}x))}$

Expected values along the offer-space



The Proposer makes the offer x_{max} which (softly) maximize the expected value $EV(\mathbf{x})$

Here $x_{max} \sim 8.50$

Intermediary: Deriving optimal policy

A bit of high-school math

- We want to find x that maximize the expected payoff, i.e. argmax(EV(x))

$$\forall x \in [0\ M], EV(x) = \frac{M - x}{1 + \exp(-(\widehat{\theta_{I,t}} + \widehat{\theta_S} x))}$$

- To find the maximum, we need to find where $\frac{dEV(x)}{dx} = 0$;

$$\frac{dEV(x)}{dx} = \frac{-1 \times \left(1 + \exp\left(-(\widehat{\theta_{I,t}} + \widehat{\theta_S} x)\right)\right) - (M - x) \times \left(-b \exp\left(-(\widehat{\theta_{I,t}} + \widehat{\theta_S} x)\right)\right)}{\left(1 + \exp\left(-(\widehat{\theta_{I,t}} + \widehat{\theta_S} \mathrm{x})\right)\right)^2} = 0$$

$$\Leftrightarrow \quad -1 \times \left(1 + \exp\left(-(\widehat{\theta_{I,t}} + \widehat{\theta_S} x)\right)\right) - (M - x) \times \left(-\widehat{\theta_S} \exp\left(-(\widehat{\theta_{I,t}} + \widehat{\theta_S} x)\right)\right) = 0$$

$$\Leftrightarrow \quad \exp\left(-(\widehat{\theta_{I,t}} + \widehat{\theta_S} \mathrm{x})\right) \times \left(\widehat{\theta_S}(M - x) - 1\right) = 1$$

$$\Leftrightarrow \quad -(\widehat{\theta_{I,t}} + \widehat{\theta_S} \mathrm{x}) + \log\left(\widehat{\theta_S}(M - x) - 1\right) = 0$$

$$\Leftrightarrow \quad \widehat{\theta_S} M - \widehat{\theta_S} x - 1 + \log\left(\widehat{\theta_S} M - \widehat{\theta_S} x - 1\right) = \widehat{\theta_S} M + \widehat{\theta_{I,t}} - 1$$

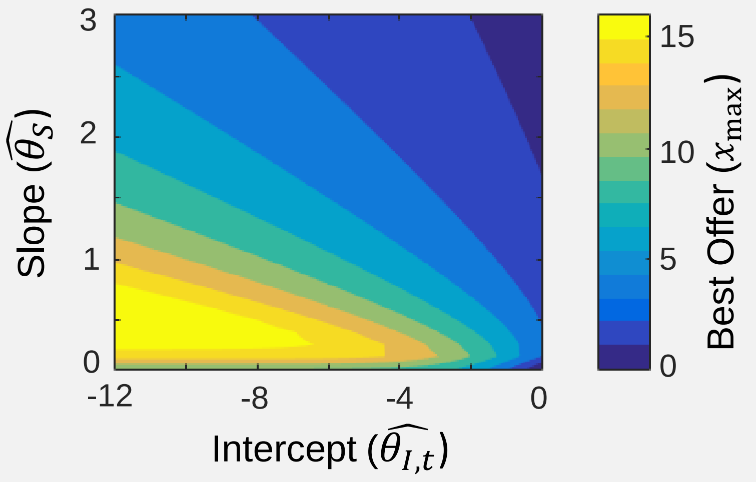
Solution

Solution from the Lambert W function:

$$\widehat{\theta_S} M - \widehat{\theta_S} x - 1 = \mathrm{W}\left(\exp\left(\widehat{\theta_S} M + \widehat{\theta_{I,t}} - 1\right)\right)$$

$$x_{max} = \frac{\widehat{\theta_S} M - 1 - W\left(\left(\exp\left(\widehat{\theta_S} M + \widehat{\theta_{I,t}} - 1\right)\right)\right)}{\widehat{\theta_S}}$$

(This is consistent with numerical approximation)



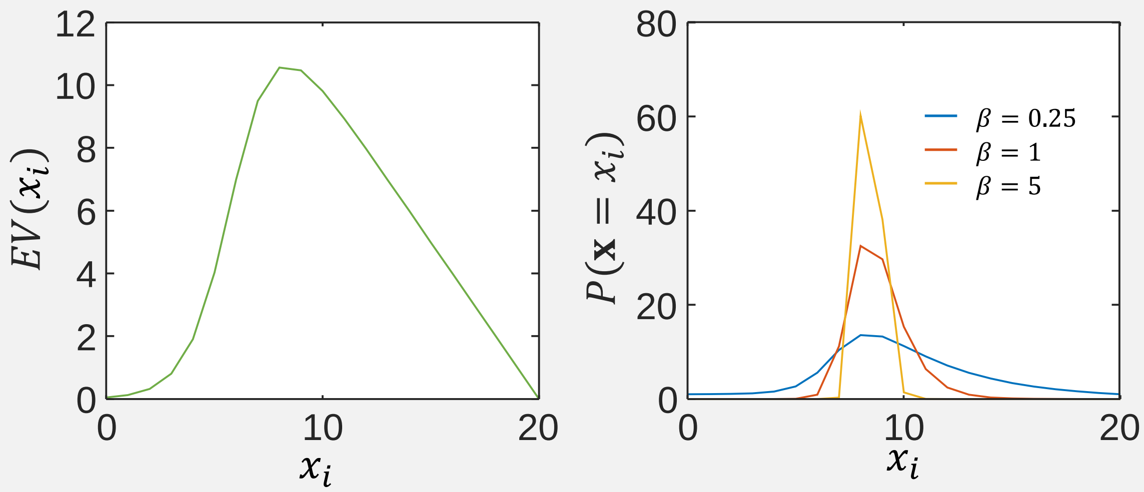
Choice function

Functional form: softmax function

multinomial logistic; transform expected values in offer probabilities

$$P(\mathbf{x} = x_i) = \frac{\exp(\beta \times EV(x_i))}{\sum_{j=1}^{M} \exp(\beta \times EV(x_j))}$$

β: temperature parameter



Interpretation

Decision (inverse) noise

Also captures

- Exploration/exploitation
- Soft/hard-maximizers

Updated rule

Update rule: delta rule

Observed Decision

$$\mathbf{A}_t = \begin{cases} 1 \text{ if accepted} \\ 0 \text{ if rejected} \end{cases}$$

Estimated probability of acceptance

$$P(\mathbf{A}_t | \mathbf{x}_t, \widehat{\theta}_t)$$

Choice prediction error

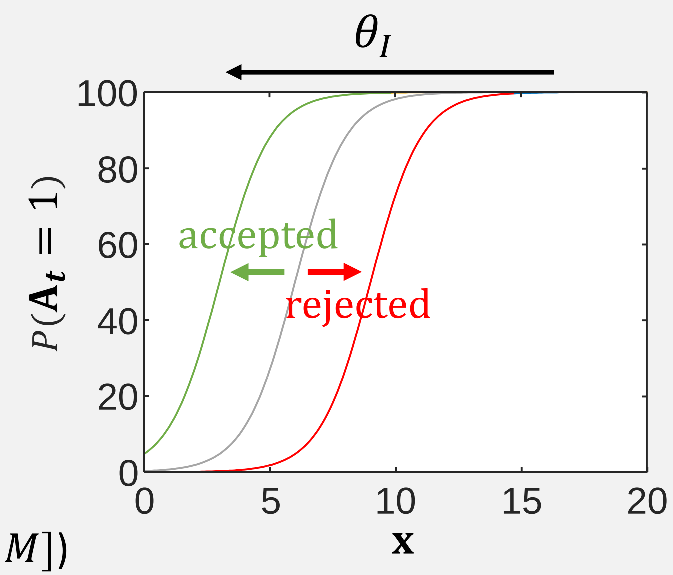
$$\delta_t = \mathbf{A}_t - P(\mathbf{A}_t | \mathbf{x}_t, \widehat{\theta}_t)$$

Update of the acceptance probability function parameters (intercept)

$$\widehat{\theta_{I,t+1}} = \widehat{\theta_{I,t}} + \alpha M \delta_t$$

α is learning rate

M is the initial endowment ($\mathbf{x} \in [0 - M]$)



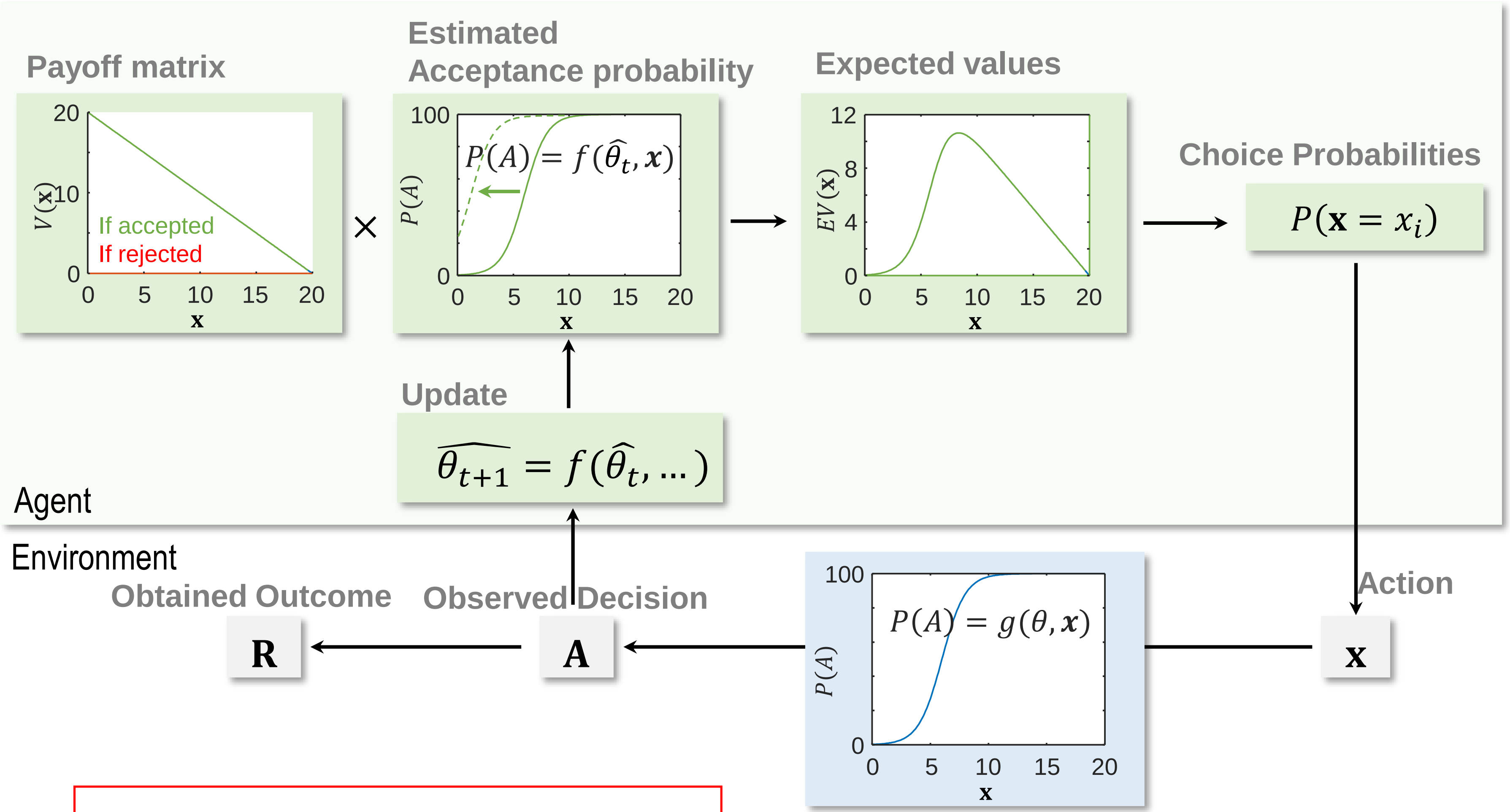
Variations

Asymmetric learning

- α^+: learning rate after accepted
- α^-: learning rate after rejected

Choice / reward prediction errors

Computational Framework: Reminder



Properties of the "environment"

- θ_I: intercept of g
- θ_S: slope of g

Free parameters

- $\widehat{\theta_{I,0}}$: initial ($t = 0$) intercept of f
- $\widehat{\theta_S}$: slope of f
- β: temperature of the choice function
- α, or $\alpha^+$and α^- : learning rate(s)

Now we understand everything ☺

Computational Framework: Generative QC (1)

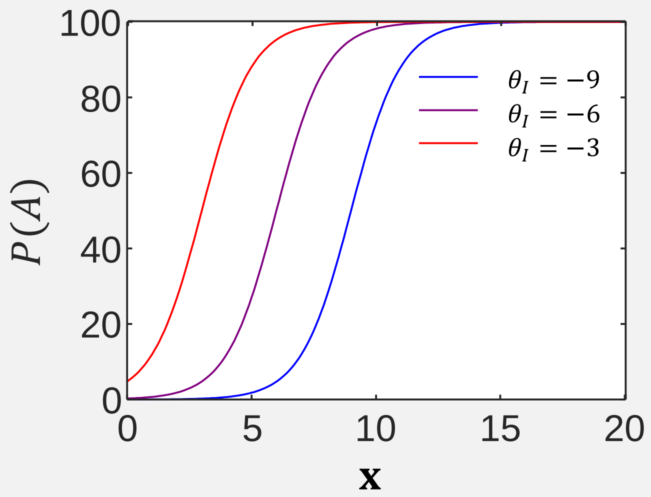
Generative QC:

Can the model generate the behavior of interest?

- Can it learn, from various priors, true receiver acceptability function parameters?
- Can it learn to ultimately make optimal offers, when faced with different receivers?

Environment: 3 receiver populations, with different acceptance probability function parameters

$$\theta_S = 1; \theta_I = \{-9; -6; -3)$$



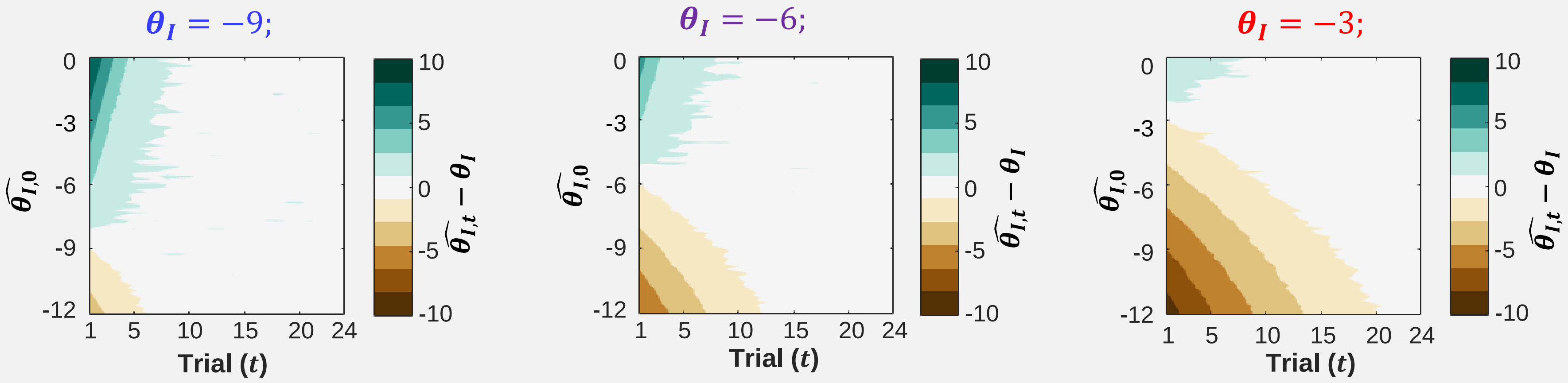
Simulations:

Assume

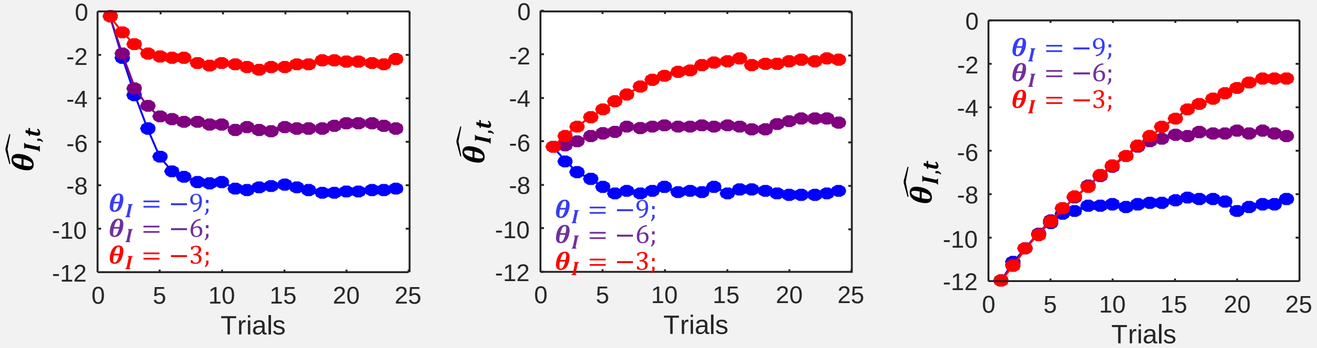
- $\widehat{\theta_S} = \theta_S = 1$; (~ experimental data – see later)
- Various levels of $\widehat{\theta_{I,0}}$
- $\beta = 5$ (~ experimental data – see later)
- $\alpha^+ = 0.25$ (~ experimental data – see later)
- $\alpha^- = 0.10$ (~ experimental data – see later)

Computational Framework: Generative QC (2)

Learning converges to the correct intercept whatever the initial belief

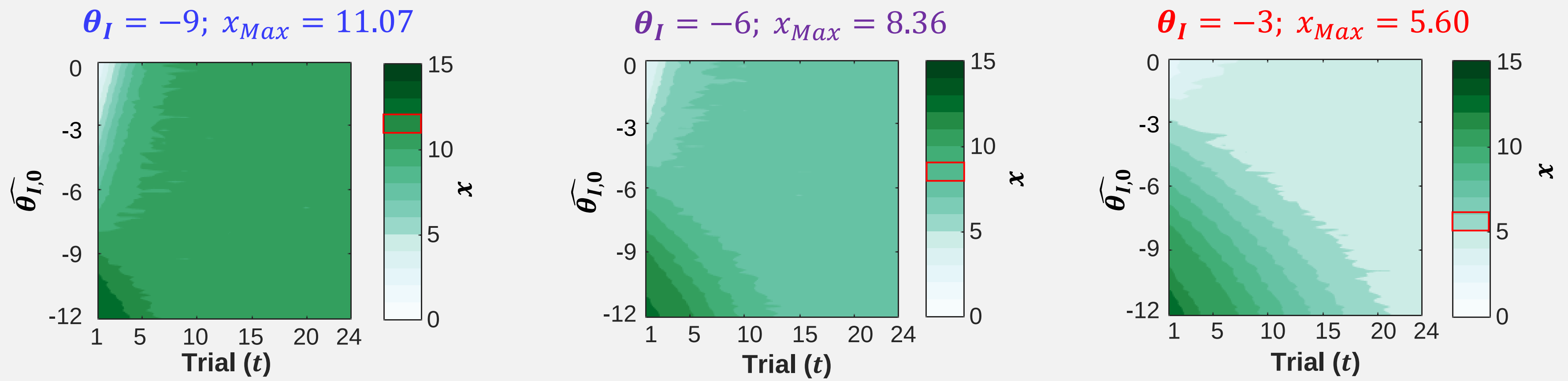


Another view at the learning dynamics



Computational Framework: Generative QC (3)

Learning therefore converges to the optimal offer whatever the initial belief



Computational Framework: Recovery QC

Recovery QC:

Can we identify different versions of the model (e.g. whether an individual uses symmetric vs asymmetric learning or choice vs reward prediction-errors)?

Can we recover the true value of free-parameters?

Estimation scheme:

Model comparison scheme:

Computational Framework: Recovery QC

Simulations: 4 models

Computational Framework: Conclusion

- Comprehensive framework;
- Produce behavior of interest
- Estimable

=> allow fine-tuned hypothesis testing, and inferences about underlying computations

Experimental Framework

- Creating different populations, with different norms