Multi-armed Bandits for Human-Machine Decision Making

Paul Reverdy

Aerospace and Mechanical Engineering
University of Arizona

Vaibhav Srivastava

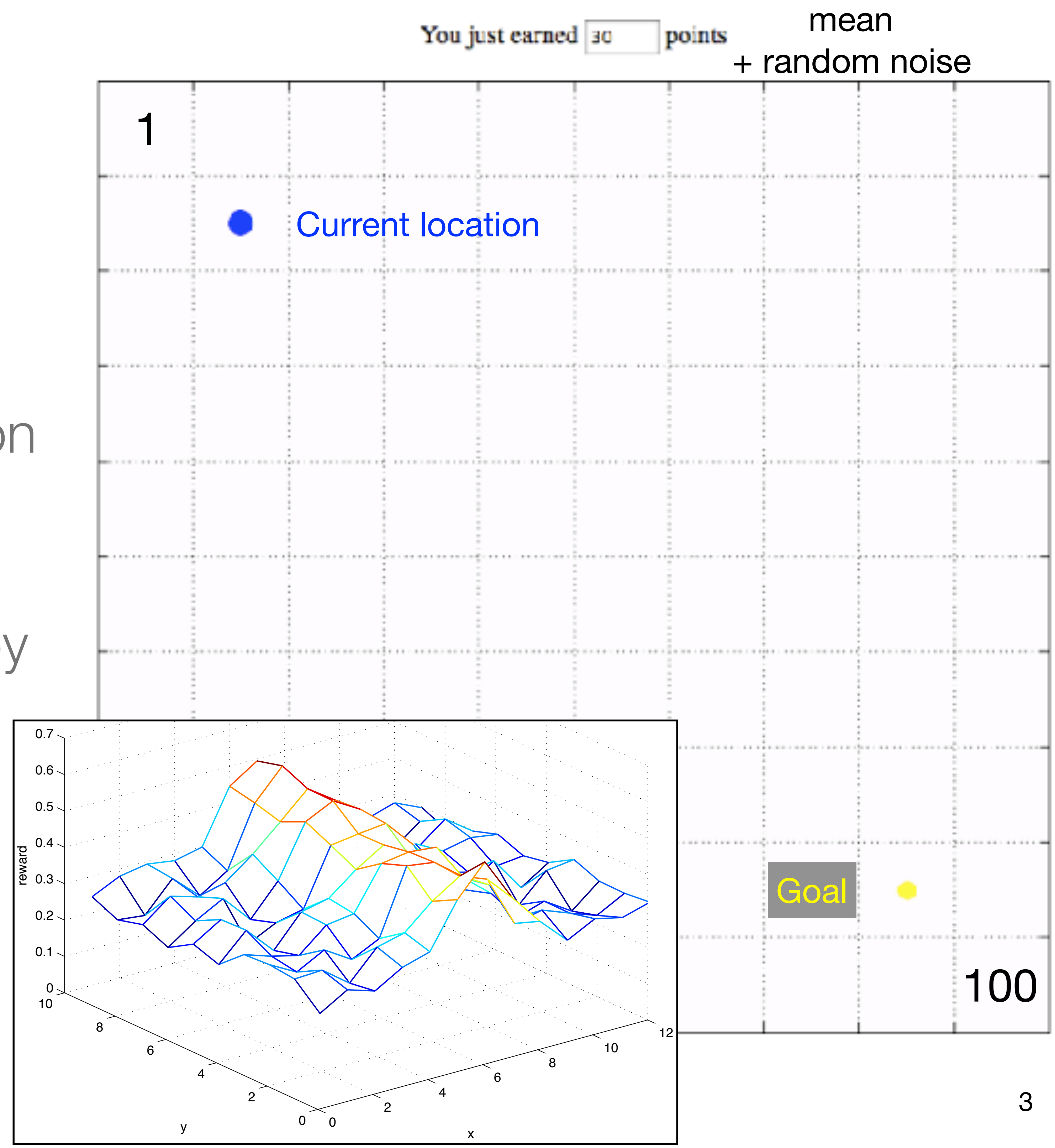
Electrical and Computer Engineering
Michigan State University

Decision-making under uncertainty

- Inference: the process of reaching conclusions from data
- *Active* Inference: can actively chose experiments to gather new data
- When such inference is used for control, results in the *explore-exploit* tradeoff: do I
    - Gather more information about the system?
    - Use existing information to maximize current performance?
- Many systems have significant structure which allows humans to achieve good performance. How to capture?

Grid task: abstraction of spatial search

- Study human behavior in spatial search tasks
- Discretize space
- Earn points based on location (unknown to subject a priori)
- Subject's goal: earn points by navigating through the grid (i.e., find peak quickly)
- Restricted movement or allow jumping in space



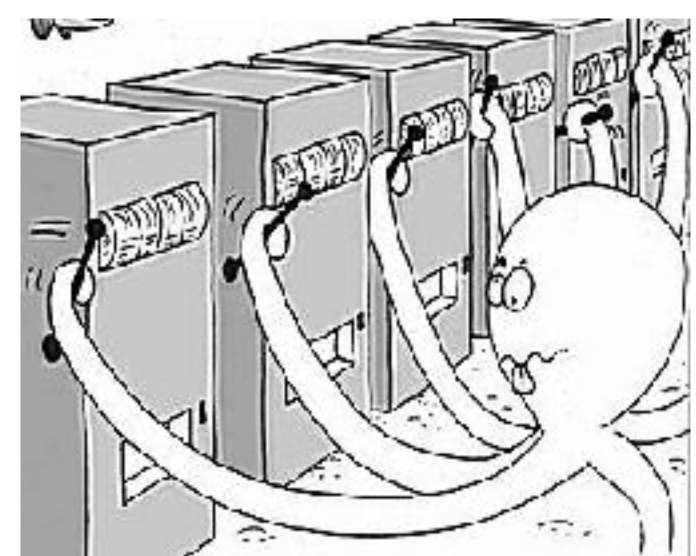
Spatial multi-armed bandit task

The multi-armed bandit problem

- A canonical representation of the *explore-exploit* tradeoff
- N options (arms), indexed by i
- Each arm has an associated distribution $p_i(r)$ with mean m_i (unknown)
- For each sequential decision time $t \in \{1, \ldots, T\}$, pick arm i_t, receive reward $r_t \sim p_{i_t}(r)$
- Objective: maximize cumulative expected reward

$$\max_{\{i_t\}} J, \quad J = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$

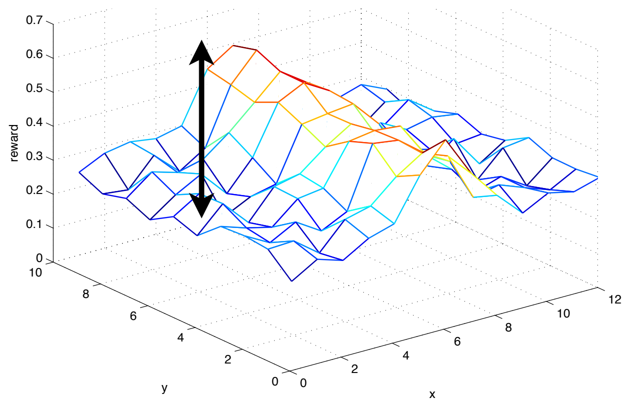
Sequential decisions



research.microsoft.com

Regret

- Bounds on optimal performance more easily formulated in terms of *regret*:
- Define $m_* = \max_i m_i$ and $R_t = m_* - m_{i_t}$, expected regret at time t



- Objective: minimize cumulative expected regret (analytical quantity)

Omniscient optimal

Mean value of decisions made

$$J_R = \sum_{t=1}^{T} R_t = T m_* - \sum_{t=1}^{T} m_{i_t}$$

Sum over decisions

$$= \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_i^T]$$

Sum over options

$\Delta_i = m_* - m_i$: Expected regret

n_i^T : Number of times option i chosen

Bounds on optimal performance

- A fundamental result of Lai and Robbins (1985) shows

$$p_i = \mathcal{N}(m_i, \sigma_s^2)$$
$$p_{i^*} = \mathcal{N}(m_{i^*}, \sigma_s^2)$$

$$\mathbb{E}\left[n_i^T\right] \geq \left(\frac{1}{D(p_i||p_{i^*})} + o(1)\right) \log T$$

Horizon

$$D(p_i||p_{i^*}) = \frac{\Delta_i^2}{2\sigma_s^2}$$

Kullback-Liebler divergence

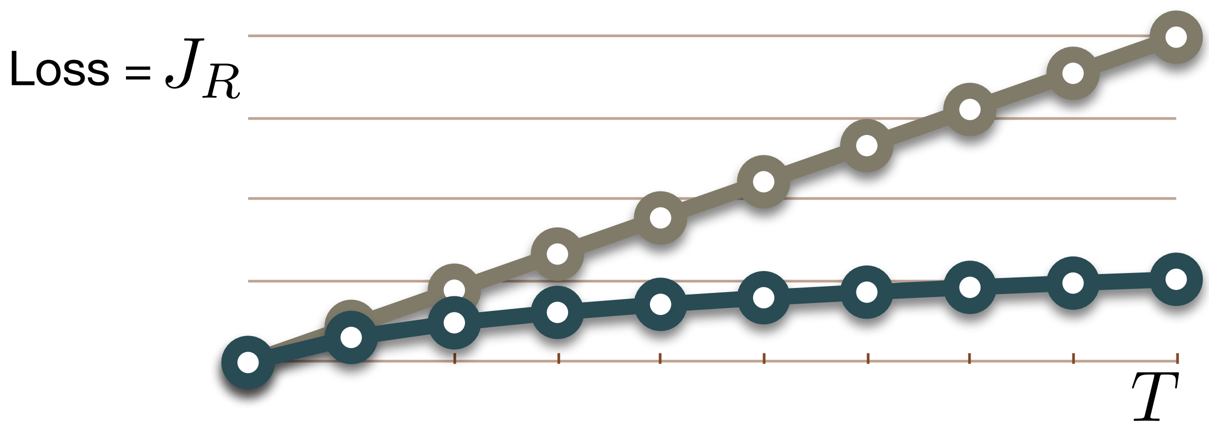
- So regret grows at least logarithmically in time:

$$J_R(T) \geq \mathcal{C} \log T$$

- Lai-Robbins is an asymptotic result; the literature seeks uniform bounds (in T)

- Uniform logarithmic regret is considered optimal

$$J_R(T) < \mathcal{C}' \log T \qquad \mathcal{C}' \,,\ \mathcal{C} \quad \text{differ by a constant factor}$$

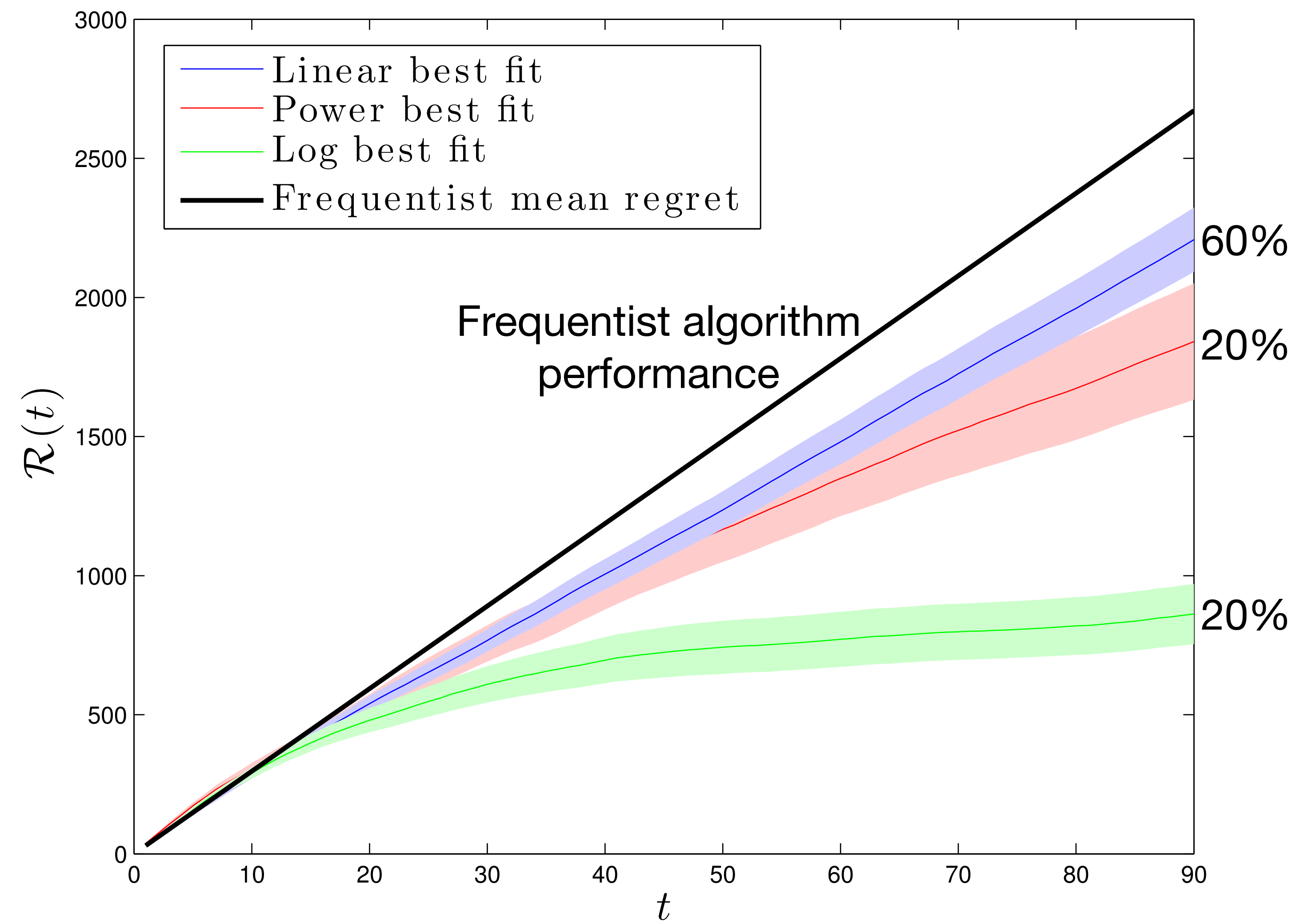


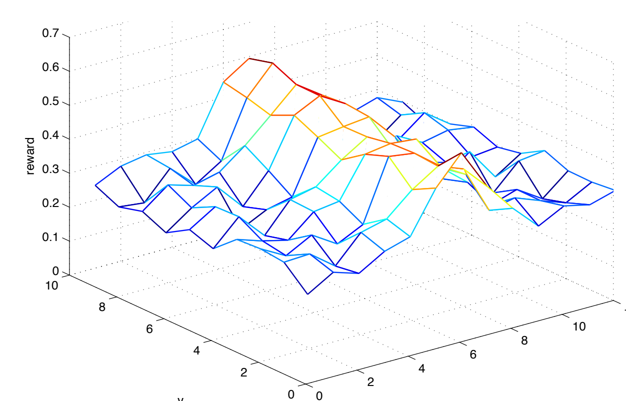
Observed human performance phenotypes

- Data from grid task; short horizon
- Fit models to observed regret:

$$\mathcal{R}(t) = a + bt$$
$$\mathcal{R}(t) = at^b$$
$$\mathcal{R}(t) = a + b\log t$$

- This set of models captures most observed performance
- Some people display logarithmic regret: “optimal” performance!
- Can we capture these three classes in a model?





Reverdy et al. *Proc. IEEE* (2014)

The Upper Credible Limit Algorithm (UCL)

Covariance belief: smoothness e.g., length scale λ

- Prior belief $\mathbf{m} \sim \mathcal{N}(\boldsymbol{\mu_0}, \Sigma_0)$ Update: Kalman filter, no dynamics

Mean reward values

Mean belief

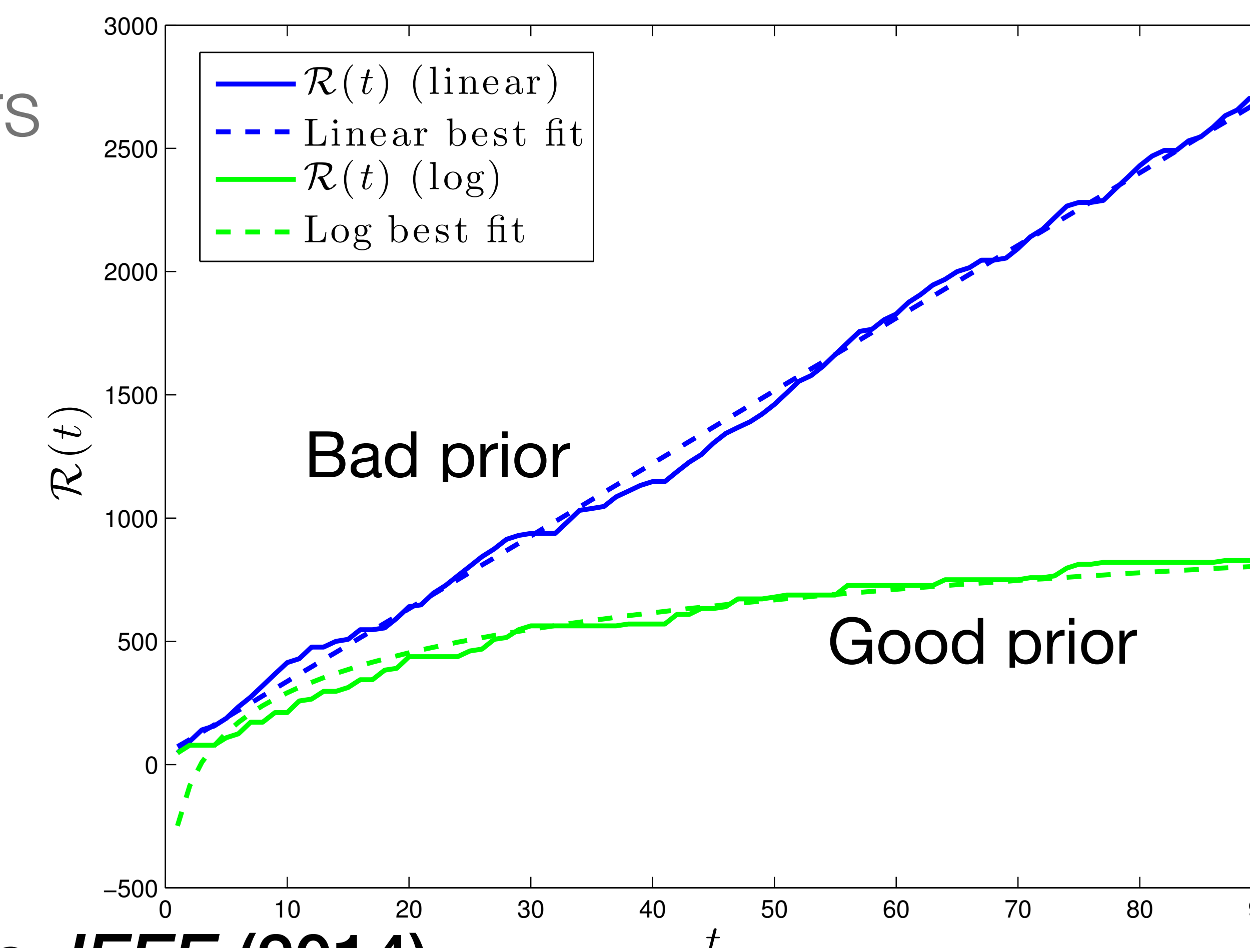
ΔI_i^t Info gain

A Ambiguity bonus: value of information

- Heuristic $$Q_i^t = \mu_i^t + \underbrace{\sigma_i^t \Phi^{-1}(1 - \alpha_t)}_{C_i^t}$$

C_i^t Uncertainty

- For $\alpha_t = 1/(\sqrt{2\pi e}t)$, achieve logarithmic regret for good priors
- And linear regret for bad priors
- Prior quality depends on accuracy and certainty



Reverdy et al. *Proc. IEEE* (2014)

Stochastic UCL

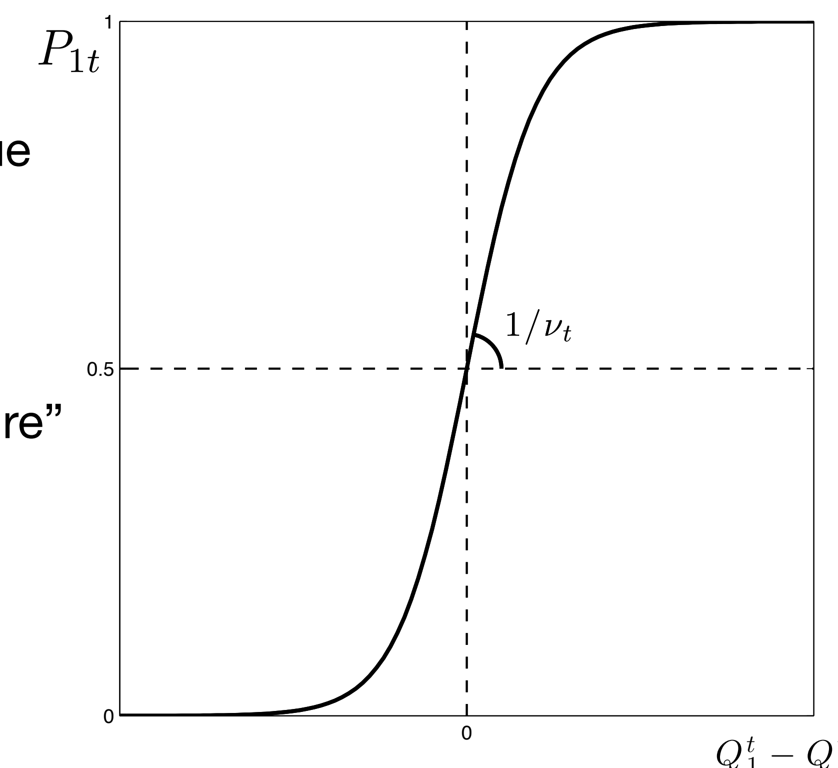
- Human decision making is stochastic, so extend UCL to stochastic policies
- Use Boltzmann/softmax action selection

Heuristic value

$$P_{it} = \frac{\exp(Q_i^t/\upsilon_t)}{\sum_{j=1}^{N} \exp(Q_j^t/\upsilon_t)}$$

Selection probability

"Temperature"



- Use dynamic temperature parameter

$$\upsilon_t = \frac{\Delta Q_{\min}^t D}{2 \log t}$$

where $\Delta Q_{\min}^t = \min_{i \neq j} |Q_i^t - Q_j^t|$ is the minimum gap between heuristic values, $D > 0$

- Stochastic UCL achieves logarithmic regret with a slightly larger constant
- But gains potential robustness to wrong priors

Reverdy et al. *Proc. IEEE* (2014)

Parameter estimation for UCL

- Have a model; need an observer
- Stochastic UCL defines a maximum likelihood estimator; requires solving hard non-convex optimization problem
- If the heuristic is a linear function of the unknown parameters, we get a generalized linear model (GLM)

$$P_{it} = \frac{\exp(\theta^T \mathbf{x}_i^t)}{\sum_{j=1}^{N} \exp(\theta^T \mathbf{x}_j^t)}$$

- Reduces to convex problem$\Rightarrow$ estimators with provable convergence
- Can be applied to stochastic UCL via linearization

Reverdy and Leonard, TASE 2016

Parameter estimates

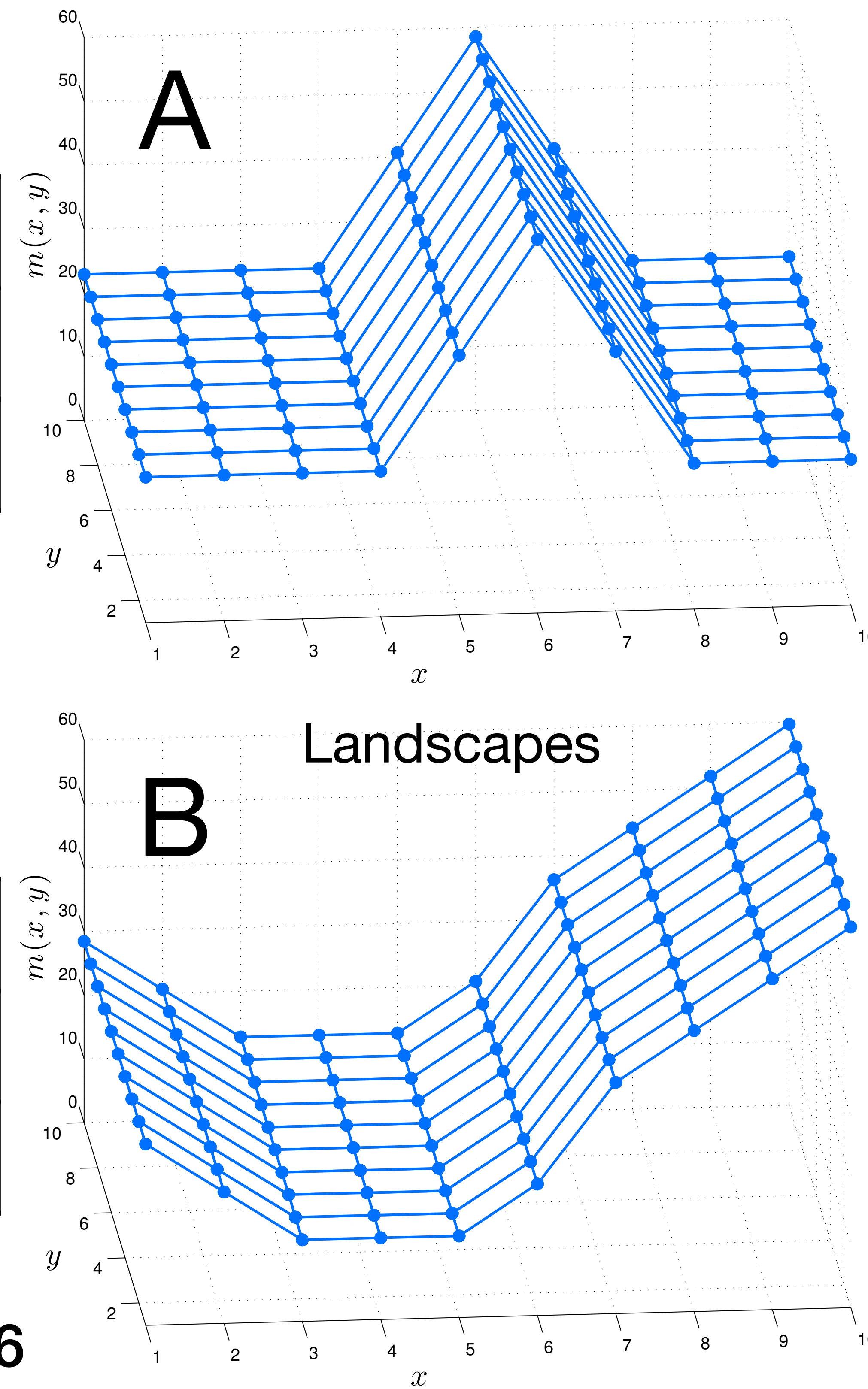
- Data from subjects with high performance
- Use GLM-based estimator
- Find statistically-significant difference between parameters for different landscapes
- Evidence for adapted strategies/priors

ν	25.5
μ_0	25.3
σ_0^2	3.32E+05

53 subjects

ν	29.5
μ_0	6.08
σ_0^2	3.35E+05

17 subjects



PR and Leonard, TASE 2016

Implications for Human-Machine Inference

- Some people (“experts”) are really good at inference, probably due to good priors
- Developed tools to learn these priors from behavior
- Algorithms can use priors to make automated decisions
- Ready to be leveraged to build human-machine active-inference and control systems

Thank you!

preverdy@email.arizona.edu
http://www.paulreverdy.com/