Optimal policies for Bayesian olfactory search in a turbulent environment

R. A. Heinonen[1], L. Biferale[1], A. Celani[2], and M. Vergassola[3]

[1]Dept. Physics and INFN, University of Rome, "Tor Vergata"
[2]The Abdus Salam International Center for Theoretical Physics
[3]Dept. Physics, Ecole Normale Supérieure

Supported by the European Research Council under grant No. 882340

Introduction: searching for an odor source in a turbulent environment

- Insects often need find source (usually upwind) of an odor or other cue advected by the atmosphere
- E.g. mosquito looking for human drawn by CO_2 and odors; moth looking for mate drawn by pheromones
- Source may be ~ 100 m away(!)

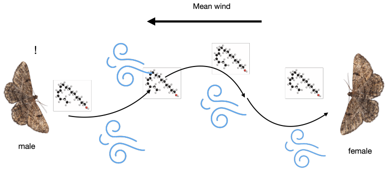


Figure Artist's conception of a moth searching for a mate via pheromone cues.

Introduction: searching for an odor source in a turbulent environment

- Classical search strategy is chemotaxis, i.e. just go up the concentration gradient
- But: (far from source) turbulence mixes cue into patches/plumes over background of very small concentration $\implies$ insect only detects the cue **intermittently**. Gradient estimation is unfeasible

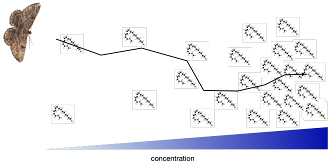


Figure Artist's conception of chemotaxis strategy.

Figure A turbulent environment leads to a patchy odor landscape with intermittent detections.

Intermittent concentration signal

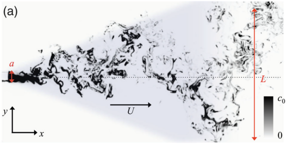


Figure Concentration field from jet flow experiment [Villermaux and Innocenti, 1999]. Fig taken from [Celani et al., 2014]

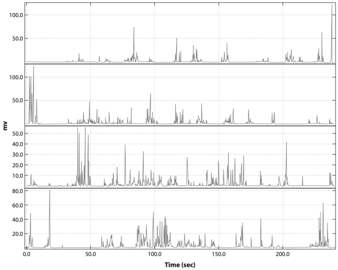


Figure Time series from experiment showing concentration signal 50 m from a propylene source over 16 minutes. From [Yee et al., 1993]

Basic motivation

How to search when cue detection is intermittent? What kind of strategies work well? We can write down heuristics, but what is the *optimal* strategy?

Model search problem

- Agent makes observation — detection or nondetection, then moves
- Try to reach source in as few Δt as possible — give reward γ^T for reaching source in T steps ($0 < \gamma < 1$)
- Key physics input is $p(\mathrm{obs}|\mathbf{s}), \mathbf{r} - \mathbf{r}_0$. Spatial dependence of concentration statistics in turbulent environment? (c.f. [Celani et al., 2014])

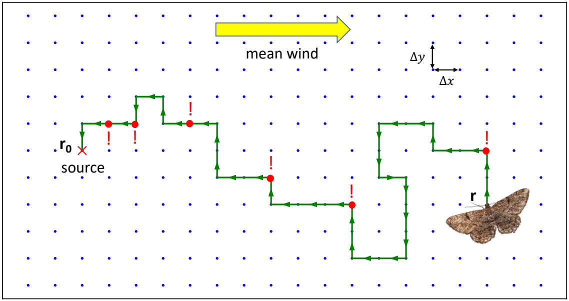
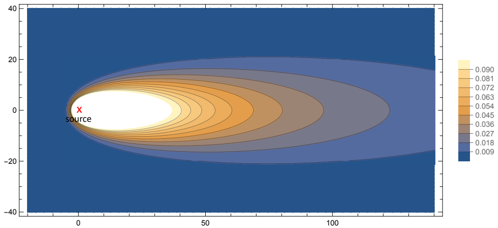


Figure In our setup, agent lives on the gridworld (blue points) and tries to find the source (red x). Grid is large, 81×41

Detection likelihood model



Advection-diffusion eq.

$$\partial_t c + \underbrace{V}_{\text{mean wind}} \partial_x c = \underbrace{D\nabla^2 c}_{\text{turb. diffusion}} + \underbrace{R\delta(\mathbf{x})}_{\text{point source}} - \underbrace{c/\tau}_{\text{turb. mixing time}},$$

stationary solution + $4\pi aDc$ detections/time $\implies$ detection rate

$$h = \frac{aR}{|\mathbf{x}|} \exp\left(\frac{Vx}{2D} - \frac{|\mathbf{x}|}{\lambda}\right), \; p(\mathrm{obs}|\mathbf{x}) = 1 - e^{-h\Delta t}$$

Capturing the information

- At timestep t, agent has history $(a_1, o_1, a_2, o_2, \ldots, a_{t-1}, o_t)$. What does this say about source location?
- If agent knows $p(o|\mathbf{s})$ (and system is Markovian), information can be stored in a *belief* b over $\mathbf{s}$
- Update b after each observation using Bayes' theorem

$$b(s')_{o,a} = p(o|s') \sum_{s} b(s) p(s'|s, a)/Z$$

- This describes a partially observable Markov decision process (POMDP) — state not accessible to agent, only observations

Optimal policy: Bellman equation

- Define value function $V_\pi(b)$ as total expected reward $\mathbb{E}[\gamma^T]$ under π, conditioned on b. Optimal value function satisfies Bellman equation

$$V^*(b) = \max_{a \in A} \left[\underbrace{\sum_{s \in S} R(s,a)b(s)}_{\substack{\text{immediate expected}\\ \text{reward}}} + \gamma \underbrace{\sum_{o \in O} p(o|b,a)V^*(b_{o,a})}_{\text{future expected rewards}} \right]$$

- Partial observability makes solution computationally hard — belief simplex very large (dimension $|S| - 1$). "Curse of dimensionality"
- Need approximation methods. We use "Perseus" algorithm [Spaan and Vlassis, 2005, Shani et al., 2006], coupled with potential reward shaping [Ng et al., 1999]

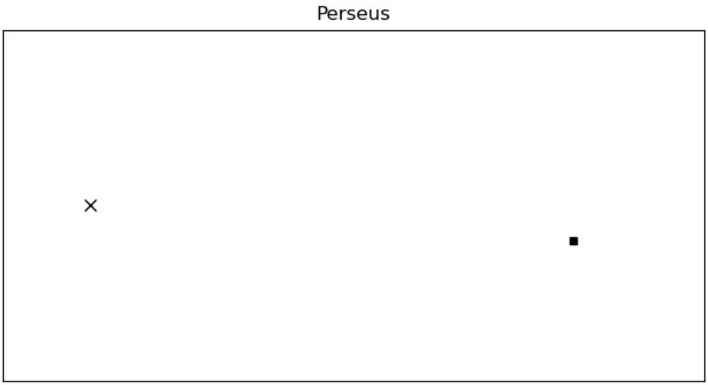
Reward shaping

- Search problem suffers from reward sparsity — $R(s,a)$ is zero for almost all state-action pairs. Slow to propagate to beliefs localized far from the source
- However one can show that adding a function of the form

$$F(s,a) = -\phi(s) + \gamma \sum_{s'} p(s'|s,a)\phi(s')$$

  to reward does not change the optimal policy
- **Good choice solves reward sparsity issue and can accelerate convergence!** E.g. $\phi(s) \propto D(s)$ is good try for search problem — yields small reward for moving closer towards source

Sample trajectories using Perseus



Performance of Perseus policies vs. heuristics

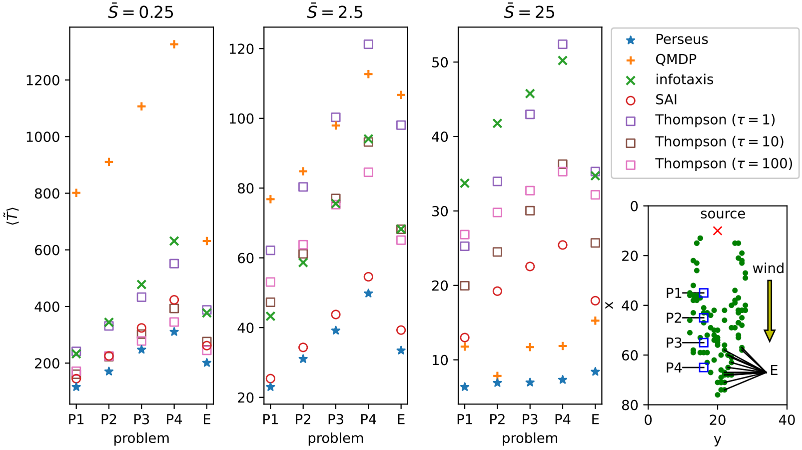


Figure Excess mean arrival times $\langle \tilde{T} \rangle = \langle T \rangle - \langle T_{MDP} \rangle$ for test problems. $\bar{S} = a\Delta t R/\Delta x$ is nondimensional emission rate

Conclusion

1. Have cast search problem as POMDP, solved for near-optimal policy for broad range of emission rates on a large grid
2. Near-optimal policy outperforms all heuristics — supremacy requires shaping the reward
3. Ongoing work: how do the policies perform in a "real" turbulent flow (DNS)?

References I

Celani, A., Villermaux, E., and Vergassola, M. (2014).
Odor landscapes in turbulent environments.
Physical Review X, 4(4):041015.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998).
Planning and acting in partially observable stochastic domains.
Artificial intelligence, 101(1-2):99–134.

Loisy, A. and Eloy, C. (2021).
Searching for a source without gradients: how good is infotaxis and how to beat it.
arXiv preprint arXiv:2112.10861.

Ng, A. Y., Harada, D., and Russell, S. (1999).
Policy invariance under reward transformations: Theory and application to reward shaping.
In *Icml*, volume 99, pages 278–287.

Shani, G., Brafman, R. I., and Shimony, S. E. (2006).
Prioritizing point-based pomdp solvers.
In *European Conference on Machine Learning*, pages 389–400. Springer.

Spaan, M. T. and Vlassis, N. (2005).
Perseus: Randomized point-based value iteration for pomdps.
Journal of artificial intelligence research, 24:195–220.

Vergassola, M., Villermaux, E., and Shraiman, B. I. (2007).
'Infotaxis' as a strategy for searching without gradients.
Nature, 445(7126):406–409.

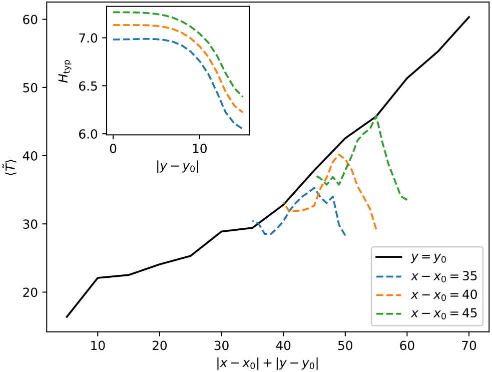
References II

Villermaux, E. and Innocenti, C. (1999).
On the geometry of turbulent mixing.
Journal of Fluid Mechanics, 393:123–147.

Yee, E., Kosteniuk, P., Chandler, G., Biltoft, C., and Bowers, J. (1993).
Statistical characteristics of concentration fluctuations in dispersing plumes in the atmospheric surface layer.
Boundary-Layer Meteorology, 65(1):69–109.

Problem difficulty dependence on starting position

- Immediate application — how hard is problem starting from different positions (measured by $\langle T \rangle - \langle T_{MDP} \rangle$)?
- Anisotropic — starting further downwind generally harder than further crosswind. Related to casting?



Searching with an imperfect model

What happens when parameters used for inference and training are incorrect? Now infotaxis much better than Perseus

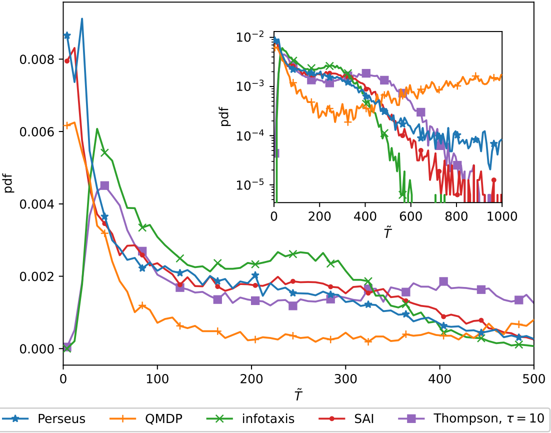


Figure Excess arrival time pdfs in $R = 5$ environment for the start point (45,-4), when the searcher's model is imperfect. Here $D \to 2D$, $V \to V/2$

Convergence of Perseus

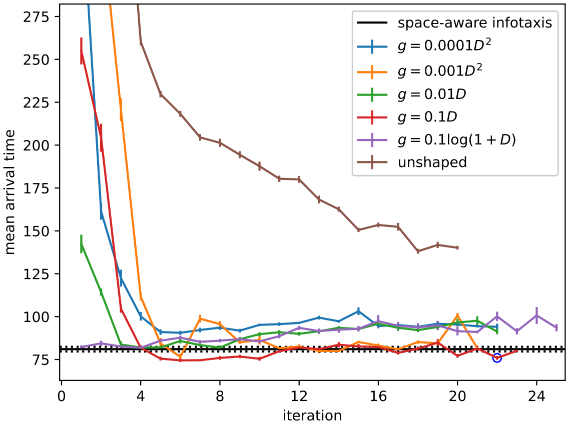


Figure Mean arrival time of Perseus policy (over ensemble of 100 start points) as function of iteration, for several shaping functions. Here $D = V = a = 1, \tau = 150, R = 5, \gamma = 0.98$ is empirically found to produce the best policy for these parameters. g is the shaping function

Perseus algorithm sketch

1. Collect large ($\sim 10^4$) sample of typical beliefs $\mathcal{B}$ by exploring with a heuristic policy
2. Assume piecewise linear and convex (PWLC) form for V^*:

$$V^*(b) = \max_{\alpha \in \mathcal{A}} b \cdot \alpha,$$

   $\mathcal{A}$ a collection of hyperplanes
3. Use Bellman equation on $b \in \mathcal{B}$ to iteratively generate α and associated actions. Old α used to approximate V^* in next iteration

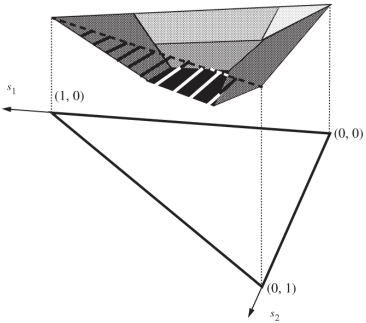


Figure PWLC value function for $|S| = 3$. High-information beliefs are located towards the corner of the simplex. From [Kaelbling et al., 1998]

Bellman error convergence

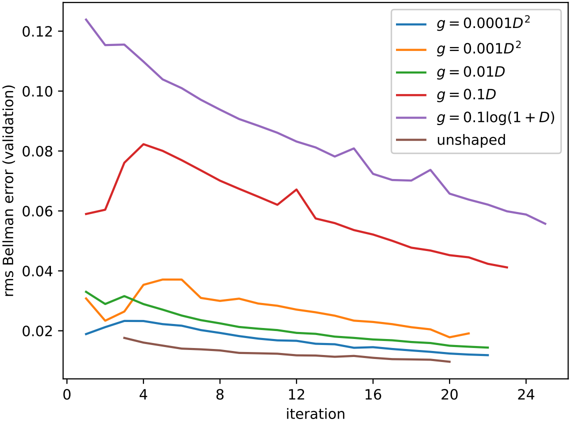
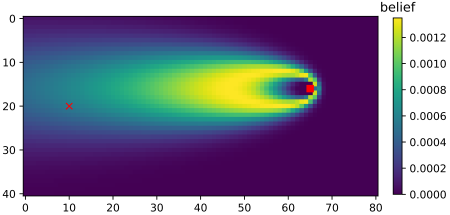


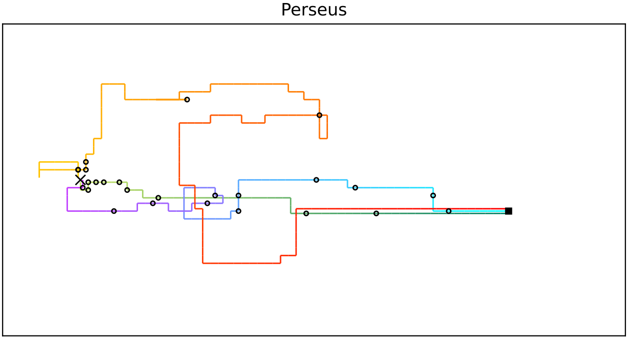
Figure rms Bellman error for beliefs encountered during testing, as function of iteration, for $R = 5$

Initial belief

- Uniform prior not realistic — real insects generally do not begin searching until they get a detection
- Forcing detection at $t = 0$ leads to strong initial bias towards the source being very near
- Instead, we let agent wait in place and update belief until it gets a detection (up to 1000 timesteps). Thus initial belief is stochastic



Sample trajectories (Perseus)

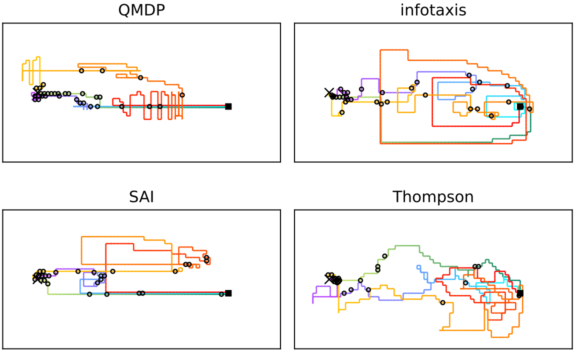


Heuristic strategies

Now we need a policy $\pi : b \mapsto a$. First try: use a hard-wired heuristic

- QMDP: take action which essentially minimizes the expected distance to the source. Exploitative (greedy)
- Infotaxis [Vergassola et al., 2007]: take action maximizing the expected gain in information (negative entropy) $I = \sum_s b(s) \log b(s)$. Explorative (less greedy)
- Space-aware infotaxis [Loisy and Eloy, 2021]: take action minimizing a function with contributions from both the distance and the entropy
- Thompson sampling: sample a point $\mathbf{r}^*$ from b, move for τ timesteps towards $\mathbf{r}^*$, repeat.

Sample trajectories (heuristics)



- QMDP: take action which essentially minimizes the expected distance to the source. Exploitative (greedy)
- Infotaxis [Vergassola et al., 2007]: take action maximizing the expected gain in information (negative entropy) $I = \sum_s b(s) \log b(s)$. Explorative (less greedy)
- Space-aware infotaxis [Loisy and Eloy, 2021]: take action minimizing a function with contributions from both the distance and the entropy
- Thompson sampling: sample a point $\mathbf{r}^*$ from b, move for τ timesteps towards $\mathbf{r}^*$, repeat.

Single start point arrival time statistics, $R = 0.5$

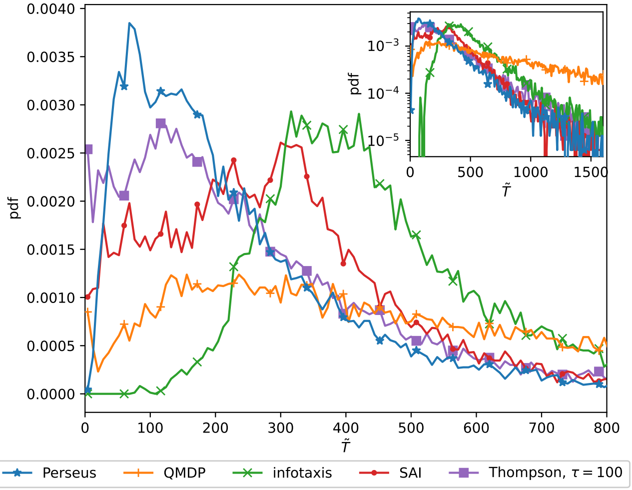


Figure Excess arrival time pdfs in $R = 0.5$ environment for the start point (45,-4), for Perseus and some heuristic policies.

Single start point arrival time statistics, $R = 5$

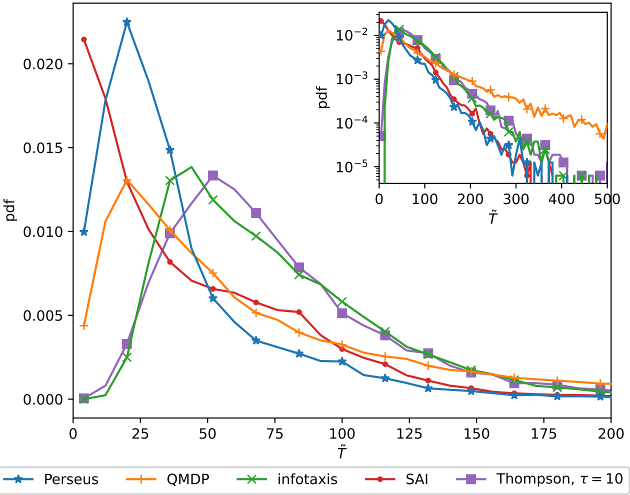


Figure Excess arrival time pdfs in $R = 5$ environment for the start point (45,-4), for Perseus and some heuristic policies.

Single start point arrival time statistics, $R = 50$

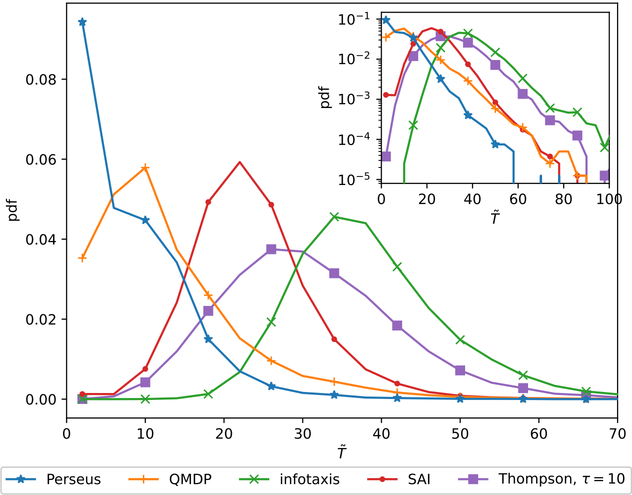


Figure Excess arrival time pdfs in $R = 50$ environment for the start point (45,-4), for Perseus and some heuristic policies.