Modelling the Water Maze

Nadav Amir

1. Modeling the Rat-Environment Interaction

We model the interaction between the rat and the environment as a Linear Quadratic control system with Gaussian noise.

For this, we start by modelling the free dynamics (the uncontrolled movement of the rat) as a Brownian motion with a drift term representing the forces acting upon it by the environment.

$$dx(t) = Ax(t)dt + \Sigma^{\frac{1}{2}}dW$$

Where $x = \left[q_x, q_y, p_x, p_y\right]^T$ is the state vector (location and momentum coordinates in the x and y directions), $\Sigma^{\frac{1}{2}}dW$ is a Brownian motion with a covariance matrix Σ and A is the state transition matrix.

We assume that all rats have equal mass (taken as $m = 1$ for convenience) and are acted on by an isotropic central harmonic potential creating a restoring force towards the origin, as well as by a dissipative drag force proportional to the velocity (=momentum):

$$\vec{F} = -(kq_x + \gamma p_x)\hat{x} - \left(kq_y + \gamma p_y\right)\hat{y}$$

Here $k > 0$ is the spring constant, $\gamma > 0$ is the drag constant and $\hat{x}$ and $\hat{y}$ are unit vectors in the x and y directions respectively.

We can thus write the state transition matrix as:

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -k & 0 & -\gamma & 0 \\ 0 & -k & 0 & -\gamma \end{bmatrix}$$

These are the dynamics of a damped harmonic oscillator. When $\gamma^2 - 4k < 0$ the oscillator is said to be underdamped and its mean behavior is described by the equation $x(t) = ae^{-\frac{\gamma}{2}t}\cos(\omega_1 t + \phi)$ where $\omega_1 = \sqrt{k^2 - \frac{\gamma^2}{4}}$ is the oscillation frequency (a and ϕ are determined by the initial condition).

2. Optimal Control – the Linear Quadratic Regulator with Gaussian Noise

The actions of the rat are described using a control signal which is learnt during the experiment. The control vector expresses the forces applied by the rat in the x and y directions $u = \left[u_x, u_y\right]^T$. By controlling these forces the rat can create desirable trajectories (policies) which

can result in higher reward values (see below) than the uncontrolled movement. The controlled dynamics are given by:

$$dx(t) = \big(Ax(t) + Bu(t)\big)dt + \Sigma dW$$

Where the input matrix is:

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(And A, Σ as before).

The quadratic cost function which is to be minimized is given by:

$$J(x,u) = E\left(\int_0^\infty \big(x^T(t)Qx(t) + u^T(t)Ru(t) + 2x(t)^T Nu(t)\big)dt\right)$$

Where Q, R, N can be time dependent but for simplicity take them to be constant matrices. We have also used the infinite horizon approximation of the Linear Quadratic regulator meaning that we are only interested in the steady state optimal control and not in the transient dynamics. This assumption is usually reasonable in practice since solutions of the Riccatti equation converge quickly to their stationary values.

Since the desired target is not located at the origin $x(t)$ needs to be replaced in the cost functional by $\Delta x(t) = x(t) - x_T$, the difference between the state vector of the rat and that of the target. In optimal-control terms, this is a *tracking* problem (and not a *regulating* problem) since we want to drive $x(t)$ to a nonzero state.

The optimal control is given by the state feedback law:

$$u(t) = Kx(t)$$

Where $K = -R^{-1}(B^T S + N^T)$ and S is the solution of the Riccatti equation:

$$0 = A^T S + SA - (SB + N)R^{-1}(B^T S + N^T) + Q$$

The interaction between the optimal control force and the central harmonic force gives rise to trajectories which will tend to fluctuate around the origin (if k is very large relative to K) or around the optimal control target (when k is small relative to K).

3. Partial Observability – The Kalman Filter

We can assume that the rat only has access to a noisy and possibly partial observation of the state vector $x(t)$. This is modeled by introducing the observation vector:

$$y(t) = Cx(t) + \xi$$

Where $\xi \sim N(0, V)$ is the observation noise and $C \in \mathbb{R}^{p \times 4}$ (with $p \leq 4$) is the output matrix. The goal of the rat is now to find the optimal control based on the observations he has access to. The solution to this problem uses the Kalman Filter (also known as the LQE - Linear Quadratic Estimator) which states that the best estimator (in the sense of minimizing the mean square error) for the uncontrolled state vector $x(t)$, using only the observation vector $y(t)$, satisfies the following (deterministic) dynamics:

$$d\hat{x}(t) = \Big(A\hat{x}(t) + L(t)\big(y(t) - C\hat{x}(t)\big)\Big)\, dt$$

With the initial condition $\hat{x}(0) = E\big(x(0)\big)$.

Here $L(t)$ is known as the Kalman Gain, and is given by $L(t) = P(t)C^T V$ where $P(t)$ is the solution of the (forward) matrix differential Riccatti equation:

$$\dot{P}(t) = AP(t) + P(t)A^T - P(t)C^T V^{-1} C P(t) + \Sigma$$

With the boundary condition:

$$P(0) = E(x(0)x(0)^T)$$

Due to the independence of the optimal control and optimal estimation problems we can now combine the two solutions to obtain the optimal control of the partially observed system using the "certainty equivalence" principle which states that the optimal control signal of the original system is simply the Linear Quadratic Regulator optimal control applied to the deterministic state estimator system, i.e.:

$$d\hat{x}(t) = \Big(A\hat{x}(t) + Bu(t) + L(t)\big(y(t) - C\hat{x}(t)\big)\Big)\, dt$$

$$u(t) = K(t)\hat{x}$$

Where K is the control gain as before.

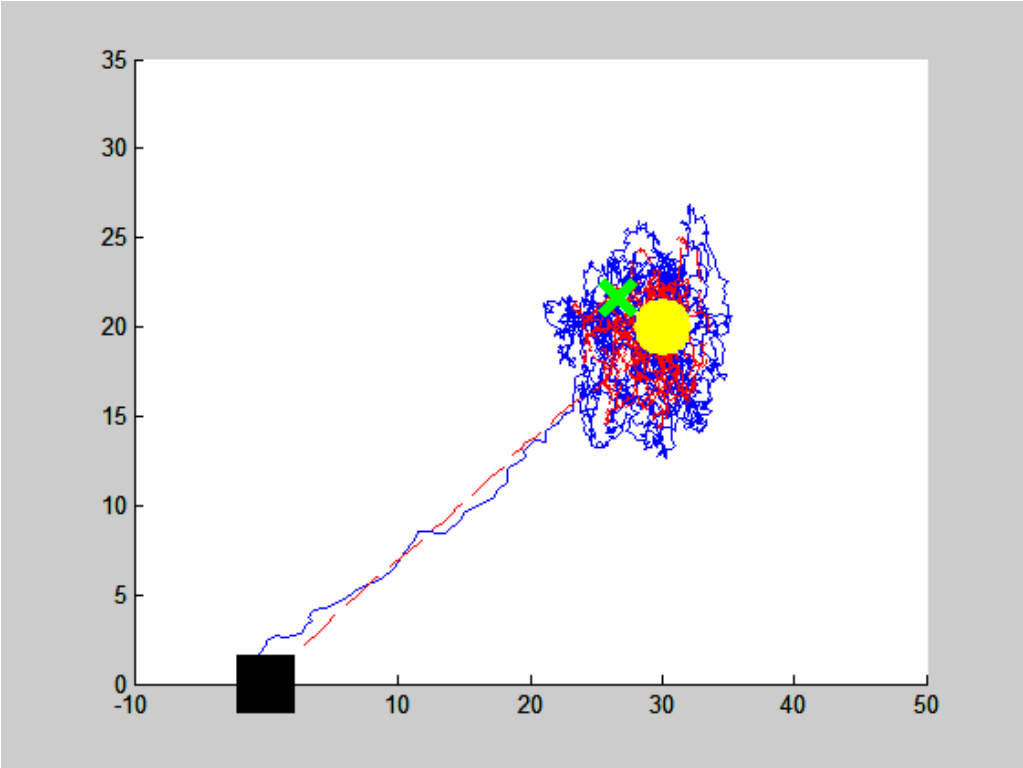
4. Running the Simulation

We initially take the output matrix to be $C = I_{4\times4}$, which corresponds to the case in which the rat has access to a noisy observation of the full state vector (location and velocity). We also set the parameters as follows:

$$\begin{aligned} \nu &= 0.043 \\ k &= 0.02 \\ \Sigma &= I_{4\times4} \\ V &= 2 \cdot I_{2\times2} \\ Q &= I_{4\times4} \\ R &= 2 \cdot I_{2\times2} \end{aligned}$$

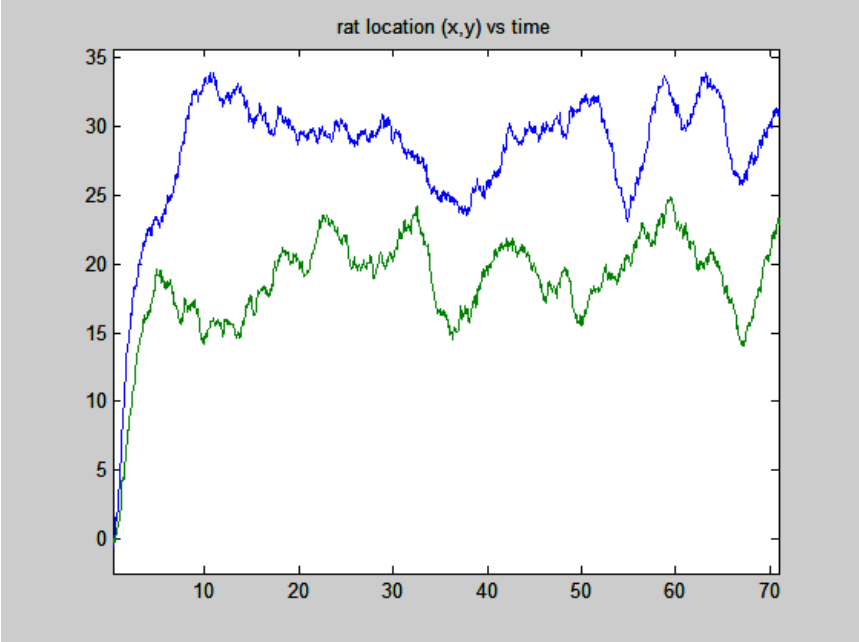
$$F = diag(0.1,0.1,0.5,1)$$
$$E(x(0)x(0)^T) = I_{4\times4}$$
$$E\big(x(0)\big) = 0$$
$$x_T = [30,20,0,0]^T$$
$$t_f = 200$$
$$dt = 0.05\ (simulation\ timestep)$$

The resulting trajectories look like this:

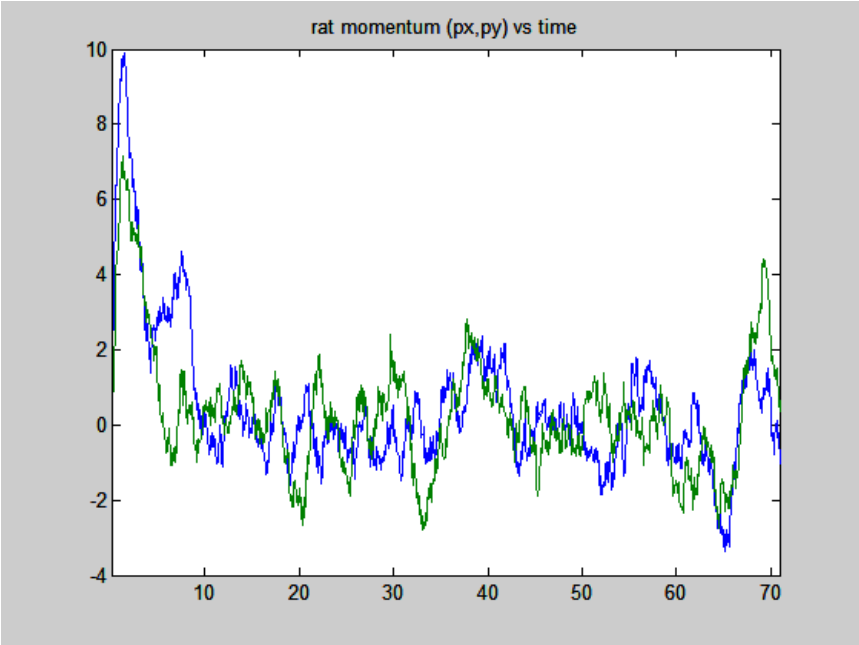


The black square at the origin represents the center of the harmonic restoring potential. The rat starts somewhere near the origin and his movement is represented by the blue trace. The red trace represents the Kalman Filter estimate of the trajectory. The yellow circle represents the location of the IA_1 and the green X is the final location of the rat.

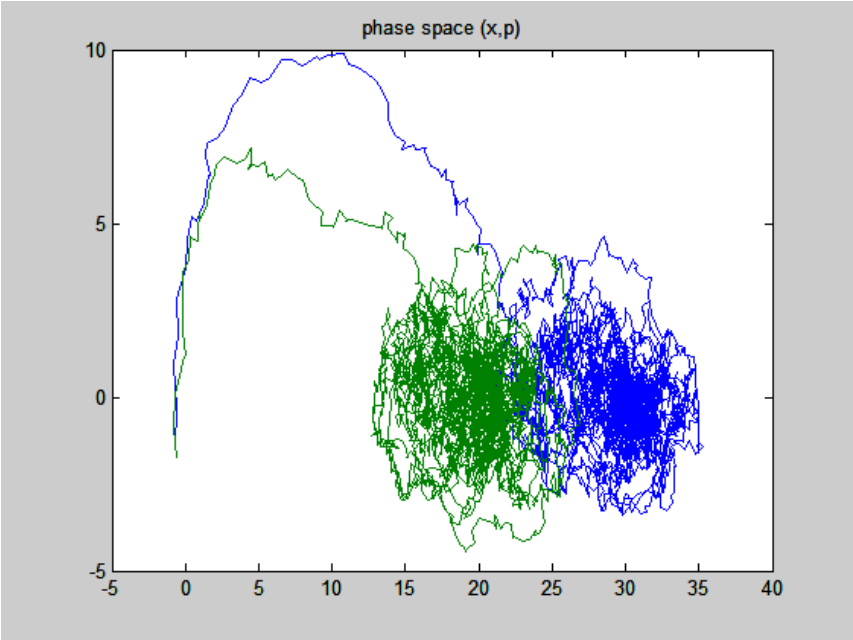
The rat (x, y) location vs. time is given below (blue: x, green: y, only first 70 seconds shown):



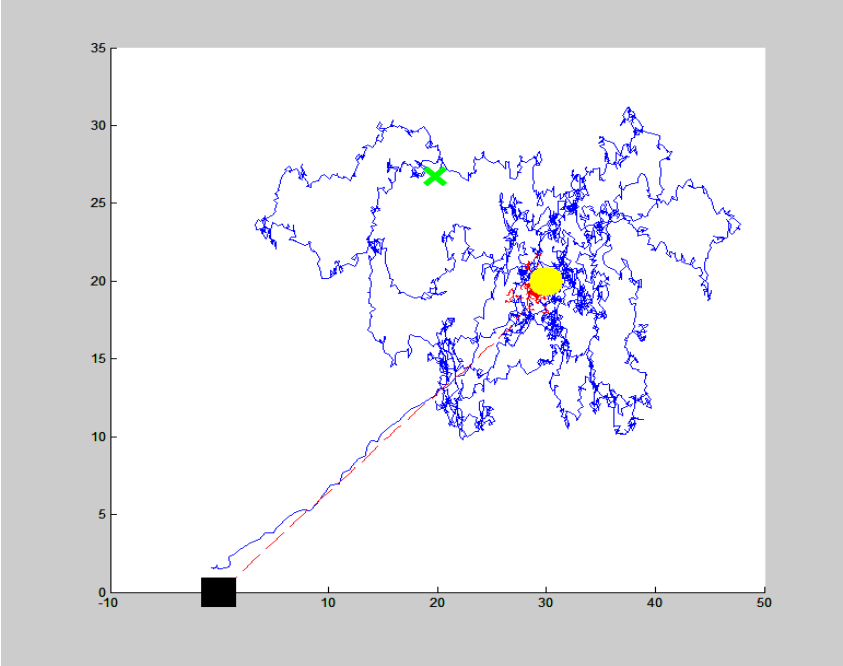
And the momentum vs time (blue: x, green: y, first 70 seconds):



Finally, the phase space (location vs momentum):

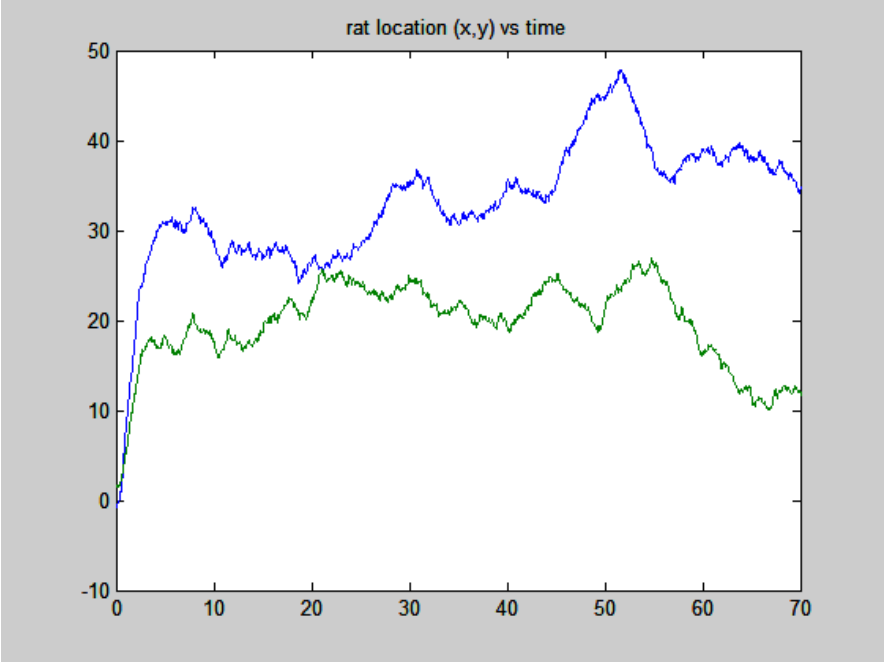


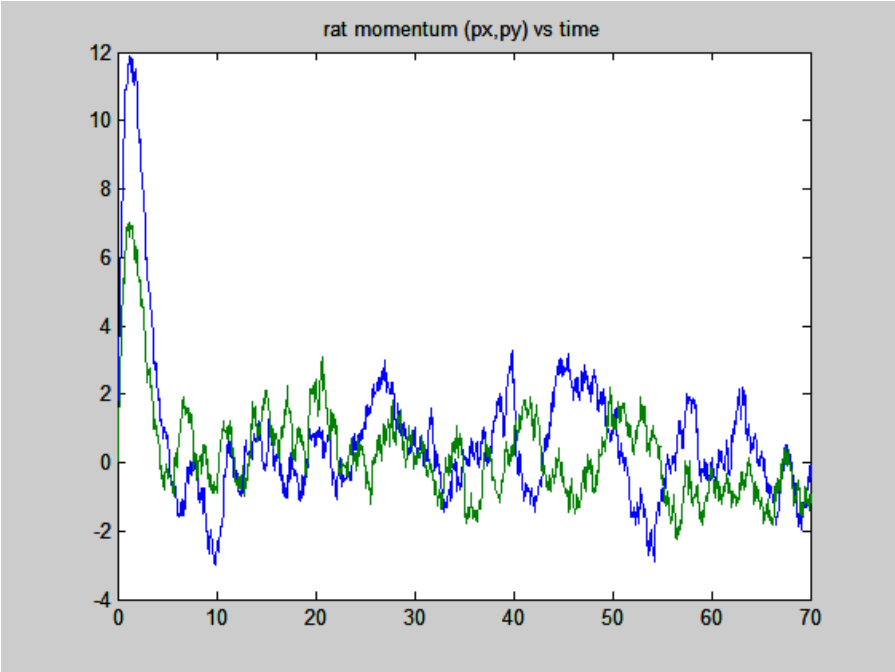
If we now set $C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, meaning that the rat has access only to (a noisy version of) its velocity, but not to its location we get the following trajectories:



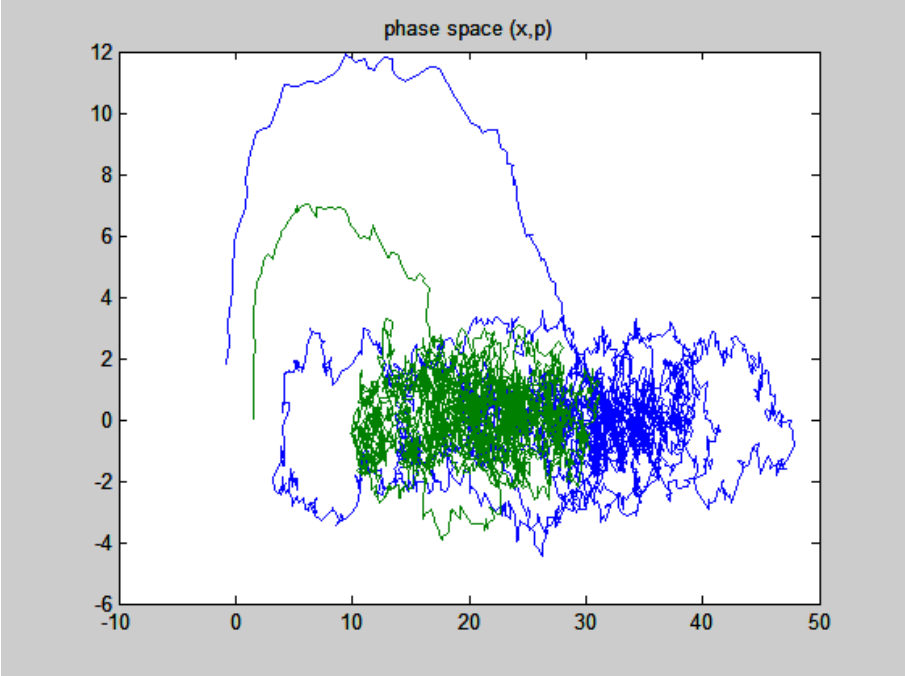
We can already see that the actual location of the rat is further away from the IA even though its estimated location is quite close to it.

The location and momentum vs time:





And the phase diagram:



5. Learning the Cost Function: Inverse Optimal Control

Here we shall be interested in the so called "inverse optimal control" problem: estimating the cost function which best describes a given set of empirical trajectories. For simplicity we also assume full observation, i.e. $C = I$ and no observation noise. Thus, following [1], we can express the cost functional in the following canonical form:

$$J(x,u) = \int_{t_0}^{\infty} \left(\Delta x^T(t) K^T K \Delta x(t) + u^T(t)u(t) - 2\Delta x^T(t) K^T u(t)\right) dt$$
$$= \int_{t_0}^{\infty} \|u(t) - K\Delta x(t)\|^2 dt$$

And it can be seen that K is precisely the optimal gain associated with this cost functional. The state cost we are minimizing is given by a quadratic form of $\Delta x(t) = x(t) - x_T$, weighed by Q, where x_T is the location of the target (the platform in the Morris Water Maze case). This formulation yields an optimal control signal of the form $u^*(t) = K\Delta x(t)$, where the gain matrix K is constant under our infinite horizon assumption.

Since we are dealing with discrete time data, sampled at intervals of $\Delta t = 0.2s$, we need to sample the continuous time model appropriately (see [2]):

$$x(t_{k+1}) = Fx(t_k) + Gu(t_k) + \tilde{w}(t_k)$$

Where:

$$F = e^{A\Delta t}$$

$$G = A^{-1}\left(e^{A\Delta t} - I\right)B$$

$$\widetilde{w} \sim N\left(0, \Sigma_{\Delta T}^{Free}\right), \qquad \Sigma_{\Delta T}^{Free} = \int_0^{\Delta t} e^{As}\Sigma e^{A^T s} dt$$

For the LQR case, $u(t_k) = K(x(t_k) - x_T)$ and in this case the continuous dynamics can be expressed as:

$$dx(t) = \left((A + BK)x(t) - BKx_T\right)dt + \Sigma^{\frac{1}{2}} dW$$

And the appropriate sampled dynamics is given by:

$$x(t_{k+1}) = F_{LQR}x(t_k) - G_{LQR}Kx_T + \widetilde{w}_{LQR}(t_k)$$

With:

$$F_{LQR} = e^{(A+BK)\Delta t}$$

$$G_{LQR} = (A + BK)^{-1}\left(e^{(A+BK)\Delta t} - I\right)B$$

$$\widetilde{w}_{LQR} \sim N\left(0, \Sigma_{\Delta T}^{LQR}\right), \qquad \Sigma_{\Delta t}^{LQR} = \int_0^{\Delta t} e^{(A+BK)s}\Sigma e^{(A+BK)^T s} dt$$

Note that $\Sigma_{\Delta T}^{Free}$ and $\Sigma_{\Delta T}^{LQR}$ are the solutions to the Lyapunov equations:

$$AX + XA^T + \Sigma - e^{A\Delta t}\Sigma e^{A^T \Delta t} = 0$$

and:

$$(A + BK)X + X(A + BK)^T + \Sigma - e^{(A+BK)\Delta t}\Sigma e^{(A+BK)^T \Delta t} = 0$$

respectively.

We define the empirical innovation at time t_k for the free dynamics ($u \equiv 0$) and LQR model as:

$$\hat{\epsilon}^{Free}(t_k) = x(t_k) - Fx(t_{k-1})$$

$$\hat{\epsilon}^{LQR}(t_k) = x(t_k) - \left(F_{LQR}x(t_{k-1}) - G_{LQR}Kx_T\right)$$

The free dynamics parameters, $\theta_{Free} = \left[k, \gamma, \sigma_q^2, \sigma_p^2\right]$ can now be estimated by maximizing the likelihood

$$L(\theta_{Free}) = \frac{1}{2}\sum_{k=2}^{N}\left(\hat{\epsilon}^{Free}(t_k)^T {\Sigma_{\Delta t}^{Free}}^{-1}\hat{\epsilon}^{Free}(t_k) + \log\left(\left|\Sigma_{\Delta t}^{Free}\right|\right)\right)$$

Where $\hat{\epsilon}^{Free}$ and $\Sigma_{\Delta t}^{Free}$ depend on θ_{Free}.

Similarly, the feedback gain matrix $\theta_{LQR} = K$ is estimated by using the free model parameters estimated above and maximizing the likelihood:

$$L\left(\theta_{LQR}\right) = \frac{1}{2}\sum_{k=2}^{N}\left(\hat{\epsilon}^{LQR}(t_k)^T {\Sigma_{\Delta t}^{LQR}}^{-1}\hat{\epsilon}^{LQR}(t_k) + \log\left(\left|\Sigma_{\Delta t}^{LQR}\right|\right)\right)$$

Note however that when minimizing the likelihood for K, care should be taken to make sure that it is a stabilizing matrix, i.e. that all eigenvalues of $A + BK$ have negative real parts.

6. **<u>Measuring the Value-Information Tradeoff</u>**

The empirical mean and covariance of the free model innovations can be calculated as:

$$\hat{\mu}^{Free} = \frac{1}{N-1}\sum_{k=2}^{N}\epsilon^{Free}(t_k)$$

$$\hat{\Sigma}^{Free} = \frac{1}{N-2}\sum_{k=2}^{N}(x(t_k) - \hat{\mu}^{Free})^T(x(t_k) - \hat{\mu}^{Free}),$$

Where N is the length of the empirical trial.

We now approximate the empirical distribution of the free innovations as a Gaussian with the empirical mean ($\hat{\mu}^{Free}$) and covariance ($\hat{\Sigma}^{Free}$ respectively):

$$\hat{P}_{Free}\left(\epsilon(t_k)\right) \sim \mathcal{N}\left(\hat{\mu}^{Free}, \hat{\Sigma}^{Free}\right)$$

To measure the information gain between the free dynamics and the empirical (learnt) distributions we calculate the DKL between the empirical and the free transition distribution:

$$S = D_{KL}\left(\hat{P}_{Free} \| P_{Free}\right) = \left(\hat{\mu}^{Free^T}{\Sigma_{\Delta t}^{Free}}^{-1}\hat{\mu}^{Free} + trace\left(\Sigma_{\Delta t}^{-1}\hat{\Sigma}^{Free}\right) - \log\left(\frac{\left|\hat{\Sigma}^{Free}\right|}{|\Sigma_{\Delta t}|}\right) - n\right)$$

Where n is the dimensionality of the state vector ($n = 4$ in our case). Note that since we have collapsed the empirical innovation distribution to a single Gaussian, the expression above corresponds to the mean information gain in a single time-step. To obtain the total information gain we need to multiply by the length of the trial: $S_{Total} = NS$.

In order to compute the value of an empirical trajectory, we need to somehow estimate the empirical control signal. This may be done by subtracting the LQR model innovation from the

free model innovation, since the former can be thought of as an estimate of the noise whereas the latter can be considered an estimate of the combined noise and control. In other words, if $\hat{u}(t_k)$ is the empirical control at time t_k we have:

$$\hat{\epsilon}^{LQR}(t_k) \approx \hat{\widetilde{w}}(t_k)$$

$$\hat{\epsilon}^{Free}(t_k) \approx G\hat{u}(t_k) + \hat{\widetilde{w}}(t_k)$$

And thus:

$$\hat{u}(t_k) \approx (G^T G)^{-1} G^T \left(\hat{\epsilon}^{Free}(t_k) - \hat{\epsilon}^{LQR}(t_k)\right)$$

Explicitly we have:

$$\begin{aligned}\hat{\epsilon}^{Free}(t_k) - \hat{\epsilon}^{LQR}(t_k) &= \left(F_{LQR}x(t_{k-1}) - G_{LQR}Kx_T\right) - Fx(t_{k-1}) \\ &= \left(F_{LQR} - F\right)x(t_{k-1}) - G_{LQR}Kx_T \\ &= e^{A\Delta t}\left(e^{BK\Delta t} - I\right)x(t_{k-1}) - (A+BK)^{-1}\left(e^{(A+BK)\Delta t} - I\right)BKx_T \\ &\approx \left((I + A\Delta t)BK\Delta t\right)x(t_{k-1}) - (BK\Delta t)x_T \approx BK\Delta t(x(t_{k-1}) - x_T)\end{aligned}$$

And since $G \approx B\Delta t$ we have:

$$\hat{u}(t_k) \approx (G^T G)^{-1} G^T \left(\hat{\epsilon}^{Free}(t_k) - \hat{\epsilon}^{LQR}(t_k)\right) \approx K\Delta x(t_{k-1})$$

So that to first order in Δt, the estimated control is equal to the optimal control.

To properly estimate the empirical cost we need to obtain a sampled version of the continuous cost function (as we did for the continuous dynamics.) This can be done by computing the following matrix exponential:

$$\exp\left(\begin{pmatrix} -A^T & 0 & Q & N \\ -B^T & 0 & N^T & R \\ 0 & 0 & A & B \\ 0 & 0 & 0 & 0 \end{pmatrix}\Delta t\right) \equiv \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ 0 & \Phi_{22} \end{pmatrix}$$

And it can be shown [3] that the sampled equivalents of the cost matrices are given by:

$$\begin{pmatrix} Q_{\Delta t} & N_{\Delta t} \\ N_{\Delta t}^T & R_{\Delta t} \end{pmatrix} = \Phi_{22}^T \Phi_{12}$$

In our case, we have $Q = K^T K, R = I, N = -K^T$ and after finding the suitable discretized versions of Q, R, N we get:

$$\hat{J} \approx \sum_k \left(\Delta x^T(t_k) Q_{\Delta t} \Delta x(t_k) + \hat{u}^T(t_k) R_{\Delta t} \hat{u}(t_k) - 2\Delta x^T(t_k) N_{\Delta t} \hat{u}(t_k)\right)$$

The empirical value is now simply defined as the negative of the empirical cost: $V = -\hat{J}$

The empirical Value-Information plot is now computed by calculating the (S, V) pair for each trajectory.

7. Computing the Optimal Value-Information Curve

For comparing the empirical Value-Information points with the theoretical limits, we must solve the problem of the LQR under information constraints. For this we define the following Hamiltonian (see [4]), using here the shorthand x_k to denote $x(t_k)$:

$$H_k = S_k - \beta V_k + \lambda(Fx_k + Gu_k)$$

Where:

$$S_k = D_{KL}\big(P_\beta(x_{k+1}|x_k) \| P_{Free}(x_{k+1}|x_k)\big)$$

Is the KL distance between the controlled transition distribution:

$$P_\beta(x_{k+1}|x_k) \sim N(Fx_k + Gu_k, \Sigma)$$

And the free dynamics transition distribution:

$$P_{Free}(x_{k+1}|x_k) \sim N(Fx_k, \Sigma_{\text{Free}})$$

F and Σ_{Free} are assumed to be known (from the free dynamics) and we wish to find u and Σ .

The value constraint term is given as usual by:

$$V_k = -\frac{1}{2}\big(x_k^T Q x_k + u_k^T R u_k^T\big)$$

Since all distributions are Gaussian, we can calculate the DKL explicitly to obtain:

$$\begin{aligned} S_k &= D_{KL}\big(P_\beta(x_{k+1}|x_k) \| P_{Free}(x_{k+1}|x_k)\big) \\ &= \frac{1}{2}\left(\big(\mu_\beta - \mu_{Free}\big)^T \Sigma_{\text{Free}}^{-1}\big(\mu_\beta - \mu_{Free}\big) + trace\big(\Sigma_{\text{Free}}^{-1}\Sigma\big) - \log\left(\frac{|\Sigma|}{|\Sigma_{\text{Free}}|}\right) - n\right) \end{aligned}$$

Where n is the dimensionality of the state vector and $\mu_\beta - \mu_{Free} = (Fx_k + Gu_k) - Fx_k = Gu_k$

It can now be seen that the DKL depends on the mean control signal only through the mean difference term: $\frac{1}{2}u_k^T G^T \Sigma_{\text{Free}}^{-1} G u_k$.

Thus the information-constrained Hamiltonian for u_k (the mean value of the optimal control signal at time t_k) can be written as:

$$H_k = \frac{1}{2} u_k^T G^T \Sigma_{\text{Free}}^{-1} G u_k + \frac{\beta}{2}\left(x_k^T Q x_k + u_k^T R u_k^T\right) + \lambda(F x_k + G u_k)$$

Or after rearranging the terms:

$$H_k = -\frac{1}{2}\beta(x_k^T Q x_k + u_k^T R(\beta) u_k^T) + \lambda(F x_k + G u_k)$$

$$R(\beta) = R + \frac{G^T \Sigma_{Free}^{-1} G}{\beta}$$

And this can now be treated as a standard Hamiltonian of a new set of LQR problems in which the cost matrices depend on β. This results in a new set of Riccatti equations, one for each value of β:

$$0 = F^T S(\beta) + S(\beta) F - S(\beta) G R(\beta)^{-1} G S(\beta) + Q(\beta)$$

And solving these equations, using standard techniques, gives us the mean optimal information-constrained control for each value of β:

$$u(\beta) = K(\beta) x$$

Where:

$$K(\beta) = -R(\beta)^{-1} G^T S(\beta)$$

To find the steady-state covariance of the optimal solution, we can use the fact that the steady-state covariance of the controlled system must satisfy the following Lyapunov equation:

$$\big(F + GK(\beta)\big)\Sigma(\beta) + \Sigma(\beta)\big(F + GK(\beta)\big)^T + \Sigma_{\text{Free}} = 0$$

And again, using standard techniques this equation can be solved to find $\Sigma(\beta)$ for each β.

8. Extension to Multiple Targets

We assume that the watermaze has two platforms, P_0 which is a fake platform – sinking when the agent reaches it and P_1 which is the real platform. We model the policy of the mouse at any point in the state space as:

$$u(x, \beta) = I_0(x) u_0(x, \beta) + I_1(x) u_1(x, \beta)$$

Where

$$u_i(x, \beta) = K_i(\beta)(x - x_i)$$

And $I_i(x)$ are indicator functions with the form:

$$I_1(x) = \begin{cases} 1, & |x - x_1| < \gamma |x - x_0| \\ 0, & otherwise \end{cases}$$

And

$$I_0(x) = 1 - I_1(x)$$

The confidence parameter γ represents the knowledge of the agent regarding the identity of the true platform. If $\gamma \gg 1$, then the policy is almost always optimal, i.e. directed towards the true platform: $u(x, \beta) = u_1(x, \beta)$, whereas $\gamma \ll 1$ the policy is almost always directed to the wrong platform. For $\gamma \approx 1$, the preferred platform is the closer one.

Bibliography

[1] F. a. R. F. Nori, "Linear optimal control problems and quadratic cost functions estimation.," in *Linear optimal control problems and quadratic cost functions estimation.*, 2004.

[2] L. a. A. W. Ljung, "Issues in sampling and estimating continuous-time models with stochastic disturbances," *Automatica ,* vol. 46, no. 5, pp. 925-931, 2010.

[3] G. F. Franklin, J. D. Powell and M. L. Workman, Digital Control of Dynamic Systems, Menlo Park, CA: Addison-Wesley, 1997.

[4] A. E. H. Y.-C. Bryson, Applied Optimal Control, Optimization, Estimation, and Control., New York-London-Sydney-Toronto: John Wiley & Sons, 1979.