Optimal Control as Probabilistic Inference Review

Chenggang Liu

May 21, 2022

1 Introduction

This note is intended to be a reference for personal learning and implementation. It covers literatures on path integral method, optimal control (reinforcement learning) as a probabilistic inference.

2 Linear solvable Markov decision problems

[Todorov, 2006] studies a class of MDP problems:

- The controlled transition probabilities

$$p_{ij}(u) = \bar{p}_{ij}\exp(u_j)$$

- The one step cost

$$\begin{aligned} l(i,u) &= q(i) + r(i,u) \\ &= q(i) + \mathrm{KL}(p_i(u)||\bar{p}_i) \end{aligned}$$

- Bellman equation

$$v(i) = \min_{u \in \mathcal{U}}\{l(i,u) + \sum_j p_{ij}v(j)\}$$

2.1 Important conclusions

- Optimally-controlled transition probabilities:

$$p^*_{ij} = \frac{\bar{p}_{ij} z(j)}{\sum_k \bar{p}_{ik} z(k)}$$

  where $z := \exp(-v(i))$ is celled 'state desirability'.

- z can be obtained by solving a linear Eigenvalue problem:

$$\mathbf{z} = G\bar{P}\mathbf{z}$$

- Z-learning

$$\hat{z}(i_k) \leftarrow (1-\alpha)\hat{z}(i_k) + \alpha_k e^{-q_k} \hat{z}(j_k)$$

2.2 Experiments

- Shortest-path problem `https://youtu.be/NOSQHOqbYLw`
- Z-learning `https://youtu.be/KyqfCMNdO2s`
- Source code `https://github.com/cgliu/z-learning.git`

3 Linear theory for control of non-linear stochastic systems

How to solve optimal deterministic control problems in the absence of noise:

- PMP: Pontryagin Minimization Principle
- HJB: Hamilton-Jacobi-Bellman equation

How to solve optimal stochastic control problems in the presence of noise:

- PMP: difficult to solve
- Stochastic HJB: the curse of dimensionality

[Kappen, 2005] studied a restrict class of optimal stochastic control problem.

- Dynamics

  $$dx = (b(x,t) + u)dt + dw$$

  where dw is a Wiener process with $< dw_i, dw_j >= v_{ij}dt$ and ν_{ij} is independent of x, u, and t.

- Minimize the following cost function:

  $$C(x,u,t) = \mathbb{E}[\phi(x(t_f) + \int_t^{t_f} \big[d\tau \frac{1}{2} u^\top Ru + q(x,\tau)\big]$$

  where $q(x,t)$ is a state dependent potential function.

- Important conclusions:

  $$\psi(x_0, t) = \int [\mathrm{d}\xi]_{x_0} \big(- \frac{1}{\lambda} S(\xi)\big)$$

  where $V(x,t) = -\log \psi(x,t)$, $\int [\mathrm{d}x]_{x_0}$ means an integral over all paths x that state at x_0 and

  $$S(\xi) = \phi(x_f) + \int_t^{t_f} d\tau \big(\frac{1}{2} u^\top Ru + q(x,\tau)\big)$$

Note:

- I have changed the symbols to be more consistent with other papers.
- The dynamics is fully actuated, where all state variables can be changed by u or the noise. [Theodorou et al., 2010] extended the dynamics, which can be under-actuated and only controlled states have noise.

4 Path Integral for robot control

Path integral method was further developed by [Theodorou et al., 2010], which studies a class of stochastic optimal control problem, where

- System dynamics:

  $$\dot{x} = f(x,t) + G(x)(u + \varepsilon) \tag{1}$$

  where ε is Gaussian noise with variance Σ_ε.

  Note:

- The noise term has to be in the control or the directly controlled state, otherwise, the method doesn't apply.
- It is sometime referred as 'linear in control'.

- Immediate cost function:

$$r_t(x,u) = q_t(x) + \frac{1}{2}u^\top Ru \tag{2}$$

  Note:

  - The immediate cost can be split into a state cost and a control dependent cost.
  - It is sometimes referred as 'quadratic'.

- Finite horizon cost function:

$$R(\tau) = \Phi(t_N) + \int_{t_i}^{t_N} r_t dt$$

  where Φ is the terminal cost function.

- Value function

$$V(x_{t_i}) = \min_{u_t} \mathbb{E}_\tau[R(\tau)]$$

  the expectation of is taken over all possible trajectories, τ, starting at x_{ti}

- The stochastic HJB equation is:

$$\partial_t V_t = \min_u \big(r_t + (\nabla_x V_t)^\top (f + Gu) + \frac{1}{2} Tr(\nabla_{xx} V_t) G_t \Sigma_\varepsilon G_t^\top)\big) \tag{3}$$

  which is a diffusion process.

- The Hamiltonian for the stochastic process:

$$H := r_t + (\nabla_x V_t)^\top (f + Gu) + \frac{1}{2} Tr(\nabla_{xx} V_t) G_t \Sigma_\varepsilon G_t^\top)$$

  compared with deterministic process, the difference is the red term, which is from the noise.

– The optimal control u^* is given by (set the gradient of Hamiltonian w.r.t. control to zero):

$$u^* = -R^{-1}G_t^\top(\nabla_x V_t)$$

substitute it into the stochastic HJB Eq. (3), and then use **an exponential transformation**:

$$V_t = -\lambda \log \Psi_t$$

where λ is a scalar.

Furthermore, set R to be inverse proportional to the noise variance as $\lambda R^{-1} = \Sigma_\varepsilon$, so that

$$\lambda G_t R^{-1} G^\top = G_t \Sigma_\varepsilon G_t^\top = \Sigma(x_t) := \Sigma_t$$

we get

$$-\partial_t \Psi_t = -\frac{1}{\lambda} q_t \Psi_t + f^\top(\nabla_x \Psi_t) + \frac{1}{2} Tr\big((\nabla_{xx}\Psi_t) G_t \Sigma_\varepsilon G_t^\top\big)$$

with boundary condition: $\Psi_{t_N} = \exp(-\frac{1}{\lambda}\Phi_N)$. This partial differential equation (PDE) corresponds to so called Chapman Kolmogorov PDE. One step further, we can apply Feynman-Kac theorem to get one of the major conclusion:

$$\Psi_{ti} = \mathbb{E}_{\tau_i}\big[\exp(-\frac{\Phi_N + \int_{t_i}^{t_N} q_t dt}{\lambda})\big] \tag{4}$$

where $\tau_i := (x_{t_i}, \ldots, x_{t_N})$ is a sample path starting at state x_{t_i}.

Note:

1. Since $V_t = -\lambda \log(\Psi_t)$. If we can get Ψ_t, we can get V_t and thus solve optimal control problem. To get the value function, we don't need to solve the HJB but We can approximate it using forward path integral!
2. We have replace the control with optimal control, so we don't need to solve it, explicitly.

3. It is still hard to solve Eq. (4), but we can get its approximation by sampling.
4. Regarding the simplification $\lambda R^{-1} = \Sigma_\varepsilon$, it couples the control cost with the system dynamics. This assumption transforms the Gaussian probability for state transitions into a quadratic command cost.
5. In [Sutton and Barto, 2018], λ is referred as temperature. High temperatures cause the actions to be all (nearly) equiprobable. Low temperatures cause a greater difference in selection probability for actions that differ in their value estimates.

4.1 Special case:

For fully actuated system:

- Optimal control at every time step t_i:

$$u^*_{t_i} = \int P(\tau_i) u(\tau_i) d\tau_i$$

- Probability of a trajectory:

$$p(\tau_i) = \frac{\exp(-\frac{1}{\lambda}\tilde{S}(\tau_i))}{\int \exp(-\frac{1}{\lambda}\tilde{S}(\tau_i)) d\tau_i}$$

For systems that can be partitioned into directly actuated part and non-directly actuated:

$$\begin{pmatrix} x^m \\ x^c \end{pmatrix} = \begin{pmatrix} f^m(x) \\ f^c(x) \end{pmatrix} + \begin{pmatrix} 0 \\ G^c \end{pmatrix} (u + \varepsilon)$$

When G^c is square and state independent, the optimal control is given by (refer to eq:23):

$$u^*_{t_i} = \frac{\int \exp(-\frac{1}{\lambda}\tilde{S}(\tau_i)) \varepsilon_{t_i} d\tau}{\int \exp(-\frac{1}{\lambda}\tilde{S}(\tau_i)) d\tau} \tag{5}$$

where, for many systems,

$$\tilde{S}(\tau_i) = \Phi_{t_N} + \int_{t_i}^{t_N} r_t dt \tag{6}$$

Note: Eq. (6) has been simplified for specific systems and is different from what in Table 1 of the original paper. For derivation, refer to 8.1

For other more general cases and PIˆ2 (policy improvement with path integrals) method, please refer to the paper.

5 INFORCE algorithm

- run simulator with π_θ to collect ξ^1
- $\nabla_\theta J = \frac{1}{N}\left[R(\xi^i)\sum_t \nabla_\theta \log \pi(a_t^i|s_t^i)\right]$
- $\theta_{new} = \theta_{old} + \alpha \nabla_\theta J$

6 Optimal control as a graphical model inference problem

[Kappen et al., 2012] established the link between optimal control and probabilistic inference in a clear way.

The optimal control is to minimize the following KL divergence:

$$\begin{aligned} C(x^0, p) &= D_{KL}(p||\psi) \\ \psi(\tau) &= q(\tau)\exp(-\sum_{t=0}^{T} C^x(x,t)) \end{aligned}$$

where p is the probability of controlled trajectory, q is the probability of uncontrolled trajectory, $\tau = x^{0:T}$, $S(\tau) = \sum_{t=0}^{T} C^x(x,t)$, and C^x is the state dependent cost.

Because

$$D_{KL}(p||\psi) = \int_\tau p\log(\frac{p}{q\exp(-S)})d\tau = \int_\tau p[\log(\frac{p}{q}) + S]d\tau \tag{7}$$

we can rewrite the cost function as:

$$\hat{R}(x^t, u^t, x^{t+1}, t) = \log(\frac{p^t(x^{t+1}|x^t, u^t)}{q^t(x^{t+1}|x^t)}) + C^x(x^t, t) \quad t = 0, \ldots, T-1$$

and

$$\hat{R}(x^T, u^T, x^{T+1}, T) = C^x(x^T, T)$$

The result of this KL minimization yields the "Boltzman distribution"

$$p(\tau) \quad = \quad \frac{1}{Z(x^0)}\psi(\tau)$$

and the optimal cost:

$$C(x^0, p^*) = -\log(Z(x^0)),$$

where $Z(x^0) = \sum_\tau \psi(\tau)$ is a normalization constant.

The optimal control in the current state x^0 is given by

$$p(x^1|x^0) = \sum_{x^{2:T}} p(x^{1:T}|x^0)$$

7 Reinforcement learning and control as probabilistic inference

[Levine, 2018] uses a graphical model to model the relationship between state, action, the next state, and rewards. It formulates the probability for s and a to be optimal as exp(r(s,a)), This leads to a very natural posterior distribution over actions when condition on $O_t = 1$ for all $t \in 1, , T$

$$p(\tau|Q_{1:T}) \propto [p(s_1)\prod_{t=1}^{T} p(s_{t+1}|s_t, a_t)] \exp(\sum_{t=1}^{T} r(s_t, a_t)) \tag{8}$$

The probability of observing a given trajectory is given by the product between its probability to occur according to the dynamics (the term in square brackets on the last line), and the exponential of the total reward along that trajectory.

With such formulation, the objective is to minimize the KL divergence between:

$$D_{KL}(\hat{p}||p(\tau)) = -\mathbb{E}_{\hat{p}}[\sum_{t=1:T} r(s,a)] - \mathcal{H}(\hat{p})$$

and thus it is to maximize

$$-D_{KL}(\hat{p}||p(\tau)) = \mathbb{E}_{\hat{p}}[\sum_{t=1:T} r(s,a)] + \mathcal{H}(\hat{p}) \tag{9}$$

This type of control objective is sometimes referred to as maximum entropy reinforcement learning or maximum entropy control.

If we define

$$q := \exp(-\sum_{t=0:T} C^u), \tag{10}$$

where C^u is the control dependent cost. Eq. (7) becomes

$$\begin{aligned} -D_{KL}(p||\psi) &= -\int_\tau p\log(p)d\tau - \mathbb{E}_p[\sum_{t=0:T}(C^u + C^x)] \\ &= \mathcal{H}(p) + \mathbb{E}[\sum_{t=1:T} r(s,a)] \end{aligned} \tag{11}$$

As we can see that (7) is a special form of (9), where the reward can be partitioned into a state dependent term and a control dependent term.

Eq (10) shows the probability of uncontrolled trajectory is a function of 'control' cost, which sounds wired. But because the noise term is in the control, you can think it as 'noise' cost.

8 Appendix

8.1 Appendix A

$\tilde{S}$ is defined as:

$$\tilde{s} = \Phi_{t_N} + \sum_{j=i}^{N-1} q_{tj}dt + \frac{1}{2}\sum_{j=i}^{N-1} ||\frac{x^c_{t_{j+1}} - x^c_{t_j}}{dt} - f^c_{t_j}||^2_{H^{-1}_{t_j}}dt + \frac{1}{2}\sum_{j=i}^{N-1}\log|H_j| \tag{12}$$

and

$$H_{t_j} := G^c_{t_j} R^{-1} {G^c_{t_j}}^\top$$

For many systems, when $dt \to 0$,

$$\frac{x^c_{t_{j+1}} - x^c_{t_j}}{dt} - f^c_{t_j} \to G^c u$$

$$\begin{aligned} ||\frac{x^c_{t_{j+1}} - x^c_{t_j}}{dt} - f^c_{t_j}||^2_{H^{-1}_{t_j}} &\to (G^c u)^\top (G^c R^{-1} G^{c\top})^{-1} G^c u \\ &\to u^\top R u \end{aligned}$$

For some system, when $G_t^{c\top} = G_t^{c-1}$ and $dt \to 0$

$$\tilde{s} = \Phi_{t_N} + \int_{t_i}^{t_N} r_t dt + C$$

where C is a constant and it can then be canceled in Eq. (5).

8.2 Diffusion process

a diffusion equation is usually written as:

$$\partial_t \Phi(r,t) = \nabla \cdot [D(\Phi, r) \nabla \Phi(r,t)]$$

where $\Phi(r,t)$ is the density of the diffusing material at location r and time t and $D(\phi, r)$ is the collective diffusion coefficient [1]. Diffusion equation shows us how the diffusion speed depends on the density of the material.

8.3 Laplace approximation

`https://james-brennan.github.io/posts/laplace_approximation/` `https://bookdown.org/rdpeng/advstatcomp/laplace-approximation.html`

Simply put the Laplace approximation entails finding a Gaussian approximation to a continuous probability density.

Key notes:

- Technically, it works for functions that are in the class of L2,

  $$\int g^2(x) dx < \infty$$

- We are interested in approximating a distribution:

  $$p(z) = \frac{1}{Z} f(z)$$

  where Z is the normalization coefficient. Especially we are interested in its posterior, which is in general computationally intractable.

[1] `https://en.wikipedia.org/wiki/Diffusion_equation`

- log trick + its second-order Taylor expansion at a peak. In details: To calculate $\int f(z)dx$, do 'log' trick, $f(z) = \exp(\ln(f(z))$, and make a Taylor expansion of it centered on the peak z_0:

  $$\ln(f(z)) \approx \ln(f(z_0)) - \frac{1}{2}(A)(z - z_0)^2$$

  where

  $$\mathbb{A} = -\frac{d^2}{dz^2} \ln f(z)|_{z=z_0}$$

  Thus:

  $$f(z) \approx f(z_0) \exp(-\frac{1}{2}(A)(z - z_0)^2) = f(z_0)(\frac{A}{2\pi})^{-1/2} q(z)$$

  where $q(z)$ is distribution PDF $\mathcal{N}(z|z_0, A^{-1})$

  $$\ln(f(z)) \approx \ln(f(z_0) - \frac{1}{2}(z - z_0)^\top \mathbb{H}(z - z_0)$$

  where H is the Hessian matrix, the matrix of second-order partial derivatives which describes the local curvature of $ln(f)$

8.4 Posterior mode

The maximum of a distribution is called 'mode', the peaks in a distribution.

An alternative estimate to the posterior mode is the posterior mean. It is given by E (|s), whenever it exists. If the posterior distribution of is symmetric about its mode, and the expectation exists, then the posterior mean is the same as the posterior mode, but otherwise these estimates will be different.

References

[Kappen, 2005] Kappen, H. J. (2005). Linear theory for control of nonlinear stochastic systems. *Phys. Rev. Lett.*, 95:200201.

[Kappen et al., 2012] Kappen, H. J., Gómez, V., and Opper, M. (2012). Optimal control as a graphical model inference problem. *Machine Learning*, 87(2):159–182.

[Levine, 2018] Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

[Theodorou et al., 2010] Theodorou, E., Buchli, J., and Schaal, S. (2010). A generalized path integral control approach to reinforcement learning. *J. Mach. Learn. Res.*, 11:3137–3181.

[Todorov, 2006] Todorov, E. (2006). Linearly-solvable markov decision problems. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*.