Myopic Gittins Exploration Heuristic for PILCO

Rowan McAllister

August 5, 2015

Contents

1 Situation

The simulate function generates cumulative-cost *distributions* $\mathcal{C}_\theta$ (with derivatives) for any controller parameterisation $\theta \in \Theta$ we input,

$$\mathcal{C}_\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2) \quad \leftarrow \quad \texttt{simulate}(\theta) \tag{1}$$

Currently we do gradient decent, to select the optimal θ w.r.t. mean information only:

$$\theta^* \quad \leftarrow \quad \arg\min_\theta \left[\mu_\theta\right] \tag{2}$$

In addition to cumulative-cost mean information μ_θ, we also have:

- cumulative-cost uncertainty information σ_θ^2,
- the number of trials left n.

We possibly also know the equivalent 'observation noise' that a sample from a rollout would give us (by an ability to estimate the reduction in uncertainty that would result from a future rollout).

2 Task

How can we find a θ that is optimal, not just w.r.t. mean information, but w.r.t. all:

- μ_θ mean information,
- σ^2_θ uncertainty information,
- n number of trials left information.

In addition we *might* know:

- $\hat{\sigma}^2_{\theta,t+1}$ estimated uncertainty of the next time iteration (default 0), or
- $\sigma^2_y(\theta)$ observation noise as a function of θ (default 0 for all θ).

3 Approximate Solution: Finite-Horizon

The exact solution involves infinite belief lookahead (intractable). We therefore restrict our solution-space to myopic (one-step) lookahead solutions. Myopic lookahead is an approximation to infinite belief lookahead, which only considers what could plausibly be learned in the next time step, and thereafter assumes the agent will learn nothing further. If we can make the assumption that each Gaussian $\mathcal{C}_\theta$ is independent from all other $\{\mathcal{C}_{\theta'} : \theta' \in \Theta \neq \theta\}$, then we apply the Gittin's index solution.

3.1 Optimistic Approximation

We can make the optimistic assumption that our observation noise σ^2_y is zero. I.e. we are about to learn everything there is to learn about our Gaussian bandit in the next time step. With prior cumulative-cost distribution $\mathcal{C}^t_\theta$, and applying controller parameterised by θ, and observing the rollout data $\mathcal{D}^t$, the posterior cumulative-cost ($\mathcal{C}^{t+1}_\theta$) will collapse to a delta function:

$$\mathcal{C}^t_\theta \sim \mathcal{N}(\mu^t_\theta, \sigma^2_\theta) \tag{3}$$

$$\mathcal{D}^t \leftarrow \texttt{applyController}(\theta) \tag{4}$$

$$\mathcal{C}^{t+1}_\theta | \mathcal{D}^t \sim \mathcal{N}(\mu^{t+1}_\theta, 0) \tag{5}$$

This is to say, we assume that after just one rollout with parameter θ, we will know (with certainty) the cumulative-cost of a controller parameterised by θ. This is optimistic, since the system itself has non-negligible process noise and has non-negligible observations noise.

This optimism of 'learning everything about $\mathcal{C}_\theta$ with just one rollout' is known in the POMDP literature as a strict upper bound on the true value function, usually denoted to as V_{MDP} or Q_{MDP}. The value function in POMDPs is sublinear w.r.t. the belief simplex, whereas V_{MDP} is linear (hence an upper bound). In this case we have a belief over not the physical state (as per usual in POMDPs), but a belief over the latent value function. If indeed the agent can learn everything in one rollout, then the upper bound is equal to the true value function.

3.1.1 Computing Solution

The solution involves:

1. using a myopic look-ahead tree,
2. computing the Gittins index (exact except for finite-horizon approximation) of our look-ahead tree (which uses number of trials n information),
3. using gradient decent w.r.t. mean *and* uncertainty.

Note we are minimising cumulative-costs here (rather than the usual bandit setting of maximising cumulative-rewards). OK, so we treat C_θ^t as a bandit, and we bring in a fixed Gittins bandit which offers a constant cumulative-cost λ per trial. For illustrative simplicity, let us initially assume our prior distribution C_θ^t is a standard Gaussian $c \sim \mathcal{N}(0,1)$. The Gittins index is the cost of the fixed-bandit that makes us ambivalent between choosing the uncertain bandit and the fixed bandit, i.e.:

$$\underbrace{C_{MDP}(\theta)}_{\text{cost of trying uncertain bandit}} = \underbrace{n\lambda}_{\text{cost of staying with fixed bandit}} \tag{6}$$

We now solve for λ. Let $\phi(\cdot)$ be the standard normal distribution, $\Phi(\cdot)$ its cumulative distribution function:

$$\begin{aligned}
n\lambda &= C_{MDP}(\theta) && (7)\\
&= \int_{-\infty}^{\infty} \phi(c)\min(\lambda, c)\,dc && (8)\\
&= \int_{-\infty}^{\lambda} \phi(c)\min(\lambda, c)\,dc + \int_{\lambda}^{\infty} \phi(c)\min(\lambda, c)\,dc && (9)\\
&= \int_{-\infty}^{\lambda} n\phi(c)c\,dc + \int_{\lambda}^{\infty} \phi(c)\big(c + \lambda(n-1)\big)\,dc && (10)\\
&= (n-1)\cdot\int_{-\infty}^{\lambda} \phi(c)c\,dc + \lambda(n-1)\cdot\int_{\lambda}^{\infty} \phi(c)dc && (11)\\
&= -(n-1)\phi(\lambda) + \lambda(n-1)(1-\Phi(\lambda)) && (12)\\
\lambda &= -(n-1)(\phi(\lambda) + \lambda\Phi(\lambda)) && (13)
\end{aligned}$$

Solve for λ using Newton's method:

$$\begin{aligned}
g_i = 0 &= (n-1)(\phi(\lambda) + \lambda\Phi(\lambda)) + \lambda && (14)\\
g_i' = \frac{dg_i}{d\lambda} &= (n-1)(-\lambda\phi(\lambda) + \Phi(\lambda) + \lambda\phi(\lambda)) + 1 && (15)\\
&= (n-1)\Phi(\lambda) + 1 && (16)\\
\lambda_0 &\leftarrow 0 && (17)\\
\lambda_{i+1} &\leftarrow \lambda_i - \frac{g_i}{g_i'} && (18)
\end{aligned}$$

This converges within 5-6 iterations.

3.1.2 Using Solution

OK great, we have this λ value, which is a function of the number of trails left n. How do we select θ^*?

$$\begin{aligned} \mathcal{C}_\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2) &\leftarrow \texttt{simulate}(\theta) && (19)\\ \theta^* &\leftarrow \arg\min_\theta \left[\mu_\theta - \lambda(n, \sigma_\theta^2)\right] && (20) \end{aligned}$$

Note how eq (20) equation differs with our current approach in PILCO, eq (2). Not only does eq (20) utilise mean information μ_θ, as does eq (20), but also uncertainty information σ_θ^2, and horizon information (number of trails remaining) n (via function $\lambda(n)$). Also note there are no free parameters!

3.2 Pessimistic Approximation

In this section we discuss the pessimistic approximation, which is more general that the optimistic approximation (since we can always revert to the optimistic approximation by setting observation noise to zero in this section).

Let our current prior at time t is $\mathcal{C}_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$ (dropping the inherent θ index to simplify notation). We therefore anticipate, according to our prior, that we will receive an observation distributed as

$$y_t \sim \mathcal{N}(\mathcal{C}_t, \sigma_y^2) \tag{21}$$

Thus, our new posterior will be

$$\begin{aligned} \mathcal{C}_{t+1} &\sim \mathcal{N}(\mu_{t+1}, \sigma_{t+1}^2), \text{ where} && (22)\\ \sigma_{t+1}^2 &= (\sigma_t^{-2} + \sigma_y^{-2})^{-1} && (23)\\ \mu_{t+1} &= \sigma_{t+1}^2(\sigma_t^{-2}\mu_t + \sigma_y^{-2}y_t) && (24) \end{aligned}$$

We now make the pessimistic assumption that the next time iteration is the only thing we will ever learn (according to the normal observation noise) about this particular parameterisation θ. I.e. the mean function will be our expected reward payout from $\mathcal{C}_\theta$ all the way up until the horizon.

OK so what will the approximate Gittins Index look like now? First let us note:

$$\begin{aligned} y_t &\sim \mathcal{N}(\mathcal{C}_t, \sigma_y^2) && (25)\\ &\sim \mathcal{N}(\mu_t, \sigma_t^2 + \sigma_y^2) && (26)\\ \mu_{t+1} &= \sigma_{t+1}^2(\sigma_t^{-2}\mu_t + \sigma_y^{-2}y_t) && (27)\\ &\sim \mathcal{N}(\sigma_{t+1}^2(\sigma_t^{-2}\mu_t + \sigma_y^{-2}\mu_t),\ \sigma_{t+1}^4\sigma_y^{-4}(\sigma_t^2 + \sigma_y^2)) && (28)\\ &\sim \mathcal{N}(\mu_t,\ \underbrace{\sigma_t^2 \cdot \tfrac{\sigma_t^2}{(\sigma_t^2+\sigma_y^2)}}_{s^2}) && (29) \end{aligned}$$

For simple notation let $\mu = \mu_{t+1}$. Since the Gittins index can be decomposed into a mean and additive function of variance, then for these calculation we centre $\mathcal{C}_t$, i.e. $\mu_t = 0$ (and will un-centre it later). OK now the approximate

pessimistic Gittins index:

$$
\begin{aligned}
n\lambda &= \int_{-\infty}^{\infty} \mathcal{N}(\mu; 0, s^2) \min(\lambda, \mu) d\mu && (30)\\
&= \int_{-\infty}^{\lambda} \mathcal{N}(\mu; 0, s^2) \min(\lambda, \mu) d\mu + \int_{\lambda}^{\infty} \mathcal{N}(\mu; 0, s^2) \min(\lambda, \mu) d\mu && (31)\\
&= \int_{-\infty}^{\lambda} y + (n-1)\mathcal{N}(\mu; 0, s^2)\mu \, d\mu + \int_{\lambda}^{\infty} \mathcal{N}(\mu; 0, s^2)\big(y + \lambda(n-1)\big) d\mu \\
&= (n-1) \cdot \int_{-\infty}^{\lambda} \mathcal{N}(\mu; 0, s^2)\mu \, d\mu + \lambda(n-1) \cdot \int_{\lambda}^{\infty} \mathcal{N}(\mu; 0, s^2) d\mu \\
&= (n-1)\left[-s\phi(\tfrac{\lambda}{s}) + \lambda(1 - \Phi(\tfrac{\lambda}{s}))\right] && (32)\\
\lambda &= -(n-1)(s\phi(\tfrac{\lambda}{s}) + \lambda\Phi(\tfrac{\lambda}{s})) && (33)
\end{aligned}
$$

Solve for λ using Newton's method:

$$
\begin{aligned}
g_i = 0 &= (n-1)(s\phi(\tfrac{\lambda}{s}) + \lambda\Phi(\tfrac{\lambda}{s})) + \lambda && (34)\\
g_i' = \frac{dg_i}{d\lambda} &= (n-1)(-\tfrac{\lambda}{s}\phi(\tfrac{\lambda}{s}) + \Phi(\tfrac{\lambda}{s}) + \tfrac{\lambda}{s}\phi(\tfrac{\lambda}{s})) + 1 && (35)\\
&= (n-1)\Phi(\tfrac{\lambda}{s}) + 1 && (36)\\
\lambda_0 &\leftarrow 0 && (37)\\
\lambda_{i+1} &\leftarrow \lambda_i - \frac{g_i}{g_i'} && (38)
\end{aligned}
$$

3.2.1 Example Values of $-\lambda$ for Various Inputs

Note n is the number of trials remaining (including the current trial we are considering), and σ_y^2/σ_t^2 is the ratio of the observation noise from observing a rollout and the prior variance.

	σ_y^2/σ_t^2				
n	0	0.1	1	10	∞
1	0	0	0	0	0
2	0.2760	0.2632	0.1952	0.0832	0
3	0.4363	0.4160	0.3085	0.1316	0
4	0.5492	0.5236	0.3883	0.1656	0
5	0.6360	0.6064	0.4497	0.1918	0
6	0.7065	0.6736	0.4996	0.2130	0
7	0.7658	0.7301	0.5415	0.2309	0
8	0.8168	0.7788	0.5776	0.2463	0
9	0.8616	0.8215	0.6092	0.2598	0
10	0.9015	0.8595	0.6374	0.2718	0

3.2.2 Derivatives

How would λ change for small changes in σ_t^2? To answer this, we first define λ as a function of s, i.e. $\lambda(s)$ (noting that s is the only variable that is directly a function of σ_t^2), and determine how λ would respond to changes in s such that

the constraint $g = 0$ is still satisfied. I.e.:

$$
\begin{aligned}
g = 0 &= (n-1)(s\phi(\tfrac{\lambda}{s}) + \lambda\Phi(\tfrac{\lambda}{s})) + \lambda && (39)\\
\therefore \frac{dg}{ds} = 0 &= (n-1)(\phi(\tfrac{\lambda}{s}) + \frac{d\lambda}{ds}\Phi(\tfrac{\lambda}{s})) + \frac{d\lambda}{ds} && (40)\\
\therefore \frac{d\lambda}{ds} &= -\frac{\phi(\frac{\lambda}{s})}{\Phi(\frac{\lambda}{s}) + 1/(n-1)} && (41)
\end{aligned}
$$

Now note:

$$
\begin{aligned}
\frac{d\lambda}{d\sigma_t^2} &= \frac{d\lambda}{ds}\cdot\frac{ds}{d\sigma_t^2} && (42)\\
&= \frac{d\lambda}{ds}\cdot\Big[\frac{\sqrt{\sigma_t^2+\sigma_y^2} - s/2}{\sigma_t^2+\sigma_y^2}\Big] && (43)
\end{aligned}
$$

4 Approximate Solution: Infinite-Horizon

Similar to solution in Sec. 3.2, but with infinite horizon and discounting γ.

$$
\begin{aligned}
\frac{\lambda}{1-\gamma} &= \int_{-\infty}^{\lambda} \mathcal{N}(\mu; 0, s^2)\min(\lambda,\mu)d\mu + \int_{\lambda}^{\infty}\mathcal{N}(\mu;0,s^2)\min(\lambda,\mu)d\mu && (44)\\
&= \int_{-\infty}^{\lambda} y + \mathcal{N}(\mu;0,s^2)\frac{\mu\gamma}{1-\gamma}\,d\mu + \int_{\lambda}^{\infty}\mathcal{N}(\mu;0,s^2)\big(y + \frac{\lambda\gamma}{1-\gamma}\big)d\mu\\
&= -\frac{s\gamma}{1-\gamma}\cdot\phi(\tfrac{\lambda}{s}) + \frac{\lambda\gamma}{1-\gamma}\cdot(1-\Phi(\tfrac{\lambda}{s})) && (45)\\
\lambda &= -\frac{\gamma}{1-\gamma}\cdot(s\phi(\tfrac{\lambda}{s}) + \lambda\Phi(\tfrac{\lambda}{s})) && (46)
\end{aligned}
$$

Solve for λ using Newton's method:

$$
\begin{aligned}
g_i = 0 &= \gamma\cdot(s\phi(\tfrac{\lambda}{s}) + \lambda\Phi(\tfrac{\lambda}{s})) + \lambda(1-\gamma) && (47)\\
&= \gamma\cdot(s\phi(\tfrac{\lambda}{s}) + \lambda\Phi(\tfrac{\lambda}{s}) - \lambda) + \lambda && (48)\\
g_i' = \frac{dg_i}{d\lambda} &= \gamma\cdot(-\tfrac{\lambda}{s}\phi(\tfrac{\lambda}{s}) + \Phi(\tfrac{\lambda}{s}) + \tfrac{\lambda}{s}\phi(\tfrac{\lambda}{s}) - 1) + 1 && (49)\\
&= \gamma\cdot(\Phi(\tfrac{\lambda}{s}) - 1) + 1 && (50)\\
\lambda_0 &\leftarrow 0 && (51)\\
\lambda_{i+1} &\leftarrow \lambda_i - \frac{g_i}{g_i'} && (52)
\end{aligned}
$$

4.1 Derivatives

Similar to Sec. 3.2.2, we have:

$$
\begin{aligned}
g = 0 &= \gamma\cdot(s\phi(\tfrac{\lambda}{s}) + \lambda\Phi(\tfrac{\lambda}{s})) + \lambda(1-\gamma) && (53)\\
\therefore \frac{dg}{ds} = 0 &= \gamma\cdot(\phi(\tfrac{\lambda}{s}) + \frac{d\lambda}{ds}\Phi(\tfrac{\lambda}{s})) + \frac{d\lambda}{ds}(1-\gamma) && (54)\\
\therefore \frac{d\lambda}{ds} &= -\frac{\phi(\frac{\lambda}{s})}{\Phi(\frac{\lambda}{s}) - 1 + 1/\gamma} && (55)
\end{aligned}
$$

Now note:

$$\frac{d\lambda}{d\sigma_t^2} = \frac{d\lambda}{ds} \cdot \frac{ds}{d\sigma_t^2} \tag{56}$$

$$= \frac{d\lambda}{ds} \cdot \Big[\frac{\sqrt{\sigma_t^2 + \sigma_y^2} - s/2}{\sigma_t^2 + \sigma_y^2}\Big] \tag{57}$$

4.2 Bounds

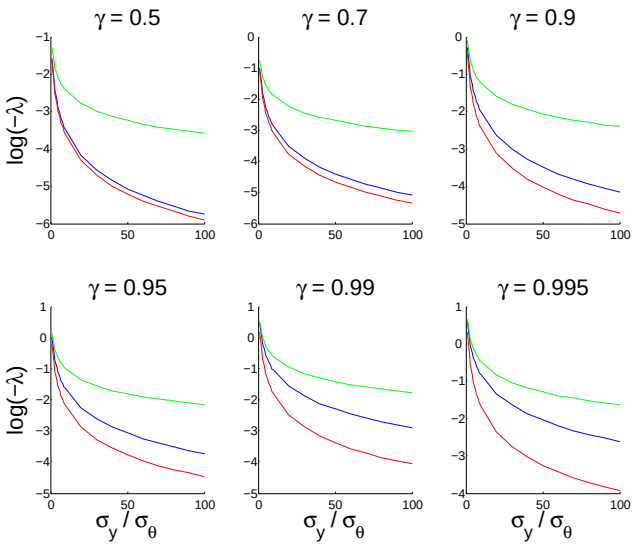

Figure 1: Bounds: The blue line is the exact Gittins index for infinite horizon. The Gittins index is upper bounded by the green line - which is the optimistic approximation to infinite horizon case. The Gittins index is also lower bounded by the red line - which is the pessimistic approximation to infinite horizon case. We set $\sigma_y = 1$.