1 Discrete Choice (Static)

A decision maker in state s must choose an action a from a finite set $A(s) = \{a_1, \ldots, a_A\}$.

The payoff (or reward) is $r(s, a)$ and is subject to a "taste" or independent random perturbations $\xi(a)$ which is observed by the decision maker (but not by the modeler).

Let $\xi \triangleq (\xi(a_1), \ldots, \xi(a_A)) \in \mathbb{R}^A$. The decision maker's optimal decision is:

$$\pi^*(s, \xi) = \arg\max_{a \in A(s)} [r(s, a) + \xi(a)]$$

The *conditional choice probability* is defined as:

$$\Pr(a|\, s) = \int \mathbf{1}_{\{a=\pi^*(s,\xi)\}} f(\xi|\, s) d\xi$$

Assume $\xi(a)$ is Gumbel distributed:

$$f_{\xi(a)}(x) = e^{-e^{-\frac{x-\mu}{\sigma}}} = \exp(-\exp(\frac{x-\mu}{\sigma}))$$

with $E[\xi(a)] = \mu + \sigma\gamma$ and $Var[\xi(a)] = \sigma^2 \frac{\pi^2}{6}$ and γ is Euler's constant. For this particular choice,

$$\begin{aligned}
\Pr(a|\, s) &= \int \mathbf{1}_{\{a=\pi^*(s,\xi)\}} f(\xi|\, s) d\xi \\
&= \frac{\exp(\frac{r(s,a)}{\sigma})}{\sum\limits_{a' \in A(s)} \exp(\frac{r(s,a')}{\sigma})}
\end{aligned}$$

This is known as *multinomial logit* model in discrete choice literature.

1.1 Estimation

Suppose data is of the form of state-action pairs $\{(s_i, a_i)\}_{i=1}^N$. Given a parametric model of $r_\theta(s, a)$ the maximum likelihood estimator $\hat{\theta}$ is the solution to

$$\hat{\theta} \in \arg\max \log(\prod_{i=1}^N \frac{\exp(\frac{1}{\sigma} r_\theta(s_i, a_i))}{\sum\limits_{a \in A(s_i)} \exp(\frac{1}{\sigma} r_\theta(s_i, a))})$$

2 Dynamic Discrete Choice Model (Rust 1987)

In addition to earning a reward $r(s, a)$ the state transitions according to state s' according to $P(s'|\, s, a)$. With discount factor the decision maker solves the problem

$$\max_\pi E[\sum_{t \geq 0} \beta^t (r(s_t, a_t) + \xi_t(a_t)]$$

with $\{\boldsymbol{\xi}_t : t \geq 0\}$ i.i.d and Gumbel. As shown in Rust (1987) the Bellman equation is of the form

$$V(s, \epsilon) = \max_{a \in A} \{r(s, a) + \xi(a) + \beta E_{s' \sim P(\cdot|s,a), \xi' \sim G}[V(s', \xi')]\}$$

With $V(s) \triangleq E_{\xi \sim G}[V(s,\xi)]$ the optimal policy is of the form:

$$\pi(a|s) = \frac{\exp \frac{1}{\sigma} Q(s,a)}{\sum_{a' \in A} \exp \frac{1}{\sigma} Q_\theta(s,a')}$$

and the *soft* Bellman equation is:

$$V(s,\epsilon) = \max_{a \in A}[Q(s,a) + \xi(a)]$$

where

$$Q(s,a) = r(s,a) + \beta \sum\nolimits_{s' \in S} P(s'|s,a)V(s')$$

and

$$V(s') = \gamma + \log(\sum\nolimits_{a' \in A} \sigma \exp(\frac{Q(s,a')}{\sigma})$$

where $\gamma > 0$ is Euler's constant. From now on assume $\sigma = 1$.

Given a set of expert trajectories $D = \{(s_{i,t}, a_{i,t}) | \, i \in \mathcal{I}, t \in \mathcal{T}\}$ and a parameterized r_θ find the value θ that maximizes the log-likelihood function

$$\log \ell(\theta) = \sum_{i \in \mathcal{I}} \sum_{t \in \mathcal{T}} \log \pi_\theta(a_{i,t}|s_{i,t}) + \sum_{i \in \mathcal{I}} \sum_{t \in \mathcal{T}} P(\, s_{i,t+1} | \, s_{i,t}, a_{i,t})$$

assuming the policies are of the form

$$\pi_\theta(a|s) = \frac{\exp Q_\theta(s,a)}{\sum_{a' \in A} \exp Q_\theta(s,a')}$$

The log-likelihood function is

$$\log \ell(\theta) = \sum_{i=1}^{N} \sum_{t>0} \log \pi_\theta(a_{i,t}|s_{i,t}) + \sum_{i=1}^{N} \sum_{t>0} P(\, s_{i,t+1} | \, s_{i,t}, a_{i,t})$$

Assuming P is known, we want to find θ that maximizes $\log \ell(\theta)$.

2.1 Nested Fixed Point Algorithm

- (*Outer-loop*) For $\theta^k$and V^k set the model

$$\pi_\theta^k(a|s) = \frac{\exp(r_\theta(s,a) + \beta \sum_{s' \in S} P(s'|s,a)V^k(s'))}{\sum_{a' \in A} \exp(r_\theta(s,a') + \beta \sum_{s' \in S} P(s'|s,a')V^k(s'))}$$

  and solve

$$\theta^{k+1} = \arg\max_{\boldsymbol{\theta}} \log(\prod_{t=0}^{T} \pi_\theta^k(a|s) P(\, s_{t+1} | \, s_t, a_t))$$

- (*Inner-loop*) For fixed θ^{k+1} compute V^{k+1} as

$$\begin{aligned} V^{k+1}(s') &= \gamma + \log(\sum\nolimits_{a' \in A} \sigma \exp(\frac{Q^{k+1}(s,a')}{\sigma}) \\ Q^{k+1}(s,a) &= r_{\theta^{k+1}}(s,a) + \beta \sum\nolimits_{s' \in S} P(s'|s,a)V^{k+1}(s') \end{aligned}$$

2.2 Relationship to Inverse Reinforcement Learning

A series of papers in IRL use an entropy-augmented reward:

$$\max_\pi E[\sum_{t\geq 0} \beta^t (r(s_t, a_t) + \sigma H_\pi(s_t))]$$

where

$$H_\pi(s) = -\sum\nolimits_{a\in A} \log \pi_\theta(a|s)\pi_\theta(a|s)$$

is the entropy of $\pi_\theta(a|s)$. It is shown in Haarnoja et al. (2017) that the optimal policy is of the form

$$\pi(a|s) = \frac{\exp \frac{1}{\sigma} Q(s,a)}{\sum_{a'\in A} \exp \frac{1}{\sigma} Q_\theta(s,a')}$$

2.3 Recursive Estimation (Aguirregaviria and Mira 2002)

We need to compute the gradient:

$$\begin{aligned}
\nabla_\theta \log \pi_\theta(a|s) &= \nabla_\theta \log(\frac{\exp Q_\theta(s,a)}{\sum_{a'\in A} \exp Q_\theta(s,a')}) \\
&= \nabla_\theta Q_\theta(s,a) - \nabla_\theta \log \sum_{a'} \exp Q_\theta(s,a') \\
&= \nabla_\theta Q_\theta(s,a) - \nabla_\theta V_\theta(s) \\
&= \nabla_\theta Q_\theta(s,a) - \sum_{a'} \pi_\theta(a'|s) \nabla_\theta Q_\theta(s,a').
\end{aligned}$$

To compute $\nabla_\theta Q_\theta$ we need to approximate Q_θ. To this end, we write

$$Q_\theta(s,a) = h_\theta(s,a) + g_\theta(s,a)$$

where h_θ and g_θ satisfy the *soft* Bellman equation:

$$h_\theta(s,a) = r_\theta(s,a) + \beta \sum_{s'} \sum_{a'} P(s'|s,a)\pi_\theta(a'|s) h_\theta(s',a') \tag{1}$$

and

$$g_\theta(s,a) = \beta \sum_{s'} \sum_{a'} P(s'|s,a)\pi_\theta(a'|s)[\gamma - \log \pi_\theta(a'|s') + \beta g_\theta(s',a')] \tag{2}$$

This is because if π_θ is the optimal policy for rewards r_θ it holds that:

$$E[\epsilon(a)|a \sim \pi_\theta(a|s)] = \gamma - \log \pi_\theta(a|s)$$

When the data is tabular, equations (1) and (2) can be solved by matrix inversion:

$$h_\theta = (I - \beta P \circ \pi_\theta)^{-1} r_\theta$$

and

$$g_\theta = (I - \beta P \circ \pi_\theta)^{-1} \beta P \circ \pi_\theta (\gamma - \log \pi_\theta)$$

The method proposed by Aguirregabiria and Mira (2002). At every iteration $k > 0$:

- Given $\hat{\pi}^k$ define

$$h_\theta^k := (I - \beta P \circ \hat{\pi}^k)^{-1} r_\theta \qquad g^k := (I - \beta P \circ \hat{\pi}^k)^{-1} \beta P \circ \hat{\pi}^k (\gamma - \log \hat{\pi}^k)$$

- Posit a model

$$\pi_\theta^k(a|s) = \frac{\exp(h_\theta^k(s,a) + g^k(s,a))}{\sum_{a'} \exp(h_\theta^k(s,a') + g^k(s,a'))}$$

and pseudolikelihood

$$\log \ell_k(\theta) := \sum_{i=1}^{N} \sum_{t>0} \log \pi_\theta^k(a_{i,t}|s_{i,t})$$

- Model update

$$\theta^{k+1} = \arg\max_{\theta \in \Theta} \log \ell_k(\theta)$$

and $\hat{\pi}^{k+1} = \pi_\theta^k\big|_{\theta=\theta^{k+1}}$

References

[1] "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher", John Rust (1987) Econometrica Vol. 55, No. 5

[2] "Reinforcement learning with deep energy-based policies", Tuomas Haarnoja, Haoran Tang, Pieter Abbeel and Sergey Levine, (2017) ICML'17: Proceedings of the 34th International Conference on Machine Learning pp. 1352-1361

[3] "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models" Victor Aguirregabiria and Pedro Mira (2002) Econometrica, Vol. 70, No. 4, pp. 1519-1543