Data simulation

Reference: Reinforcement learning design for cancer clinical trials Yufan Zhao, Michael R. Kosorok, and Donglin Zeng

A system of ODE model:

W_t : patient wellness (toxicity)

M_t : tumor size

D_t : chemotherapy agent dose (dose is the action A_t here)

$$\bar{W}_t = a_1 max(M_t, M_0) + b_1(D_t - d_1)$$

$$\bar{M}_t = [a_2 min(W_t, W_0) - b_2(D_t - d_2)] \times 1(M_t > 0)$$

where time (month) $t = 0, 1, \cdots, T-1, T = 6$. $\bar{W}_t$ and $\bar{M}_t$ are the transition functions. These changing rate yields a piece-wise linear model over time. $a_1 = 0.1, a_2 = 0.15, b_1 = 1.2, b_2 = 1.2, d_1 = 0.5$ and $d_2 = 0.5$.

$$W_{t+1} = W_t + \bar{W}_t$$

$$M_{t+1} = M_t + \bar{M}_t$$

$M_0 \sim Uniform(0, 2), W_0 \sim Uniform(0, 2)\ D_0 \sim Uniform(0.5, 1), D_t \sim Uniform(0.5, 1), t = 1, \cdots, 5$

The survival indicator:

Time interval $(t-1, t], t = 1, 2, \cdots, 6$

Log of the hazard function

$$log\lambda(t) = \mu_0 + \mu_1 W_t + \mu_2 M_t$$

where $\mu_1 = \mu_2 = 1$

$$\lambda(t) = exp\{\mu_0 + \mu_1 W_t + \mu_2 M_t\}$$

The cumulative hazard function

$$\triangle\Lambda(t) = \int_{t-1}^{t} \lambda(s)\, ds$$

$$= \int_{t-1}^{t} exp\{\mu_0 + \mu_1 W_t + \mu_2 M_t\}\, ds$$

$$= exp\{\mu_0 + \mu_1 W_t + \mu_2 M_t\}$$

The survival function

$$\triangle F(t) = exp[-\triangle\Lambda(t)]$$

$$= exp[-exp\{\mu_0 + \mu_1 W_t + \mu_2 M_t\}]$$

The random event of death, F = 1,

$$F \sim \text{Bernoulli}(p)$$

$p = 1 - \triangle F(t) = 1 - exp[-exp\{\mu_0 + \mu_1 W_t + \mu_2 M_t\}]$

Rewards

$r_t = R_t(s_t, a_t, s_{t+1}), R_t = R_{t1} + R_{t2} + R_{t3}$

$R_{t1} = -60, \text{if patient dies, o.w. } 0$

$R_{t2} = 5, \text{if } W_{t+1} - W_t \leq -0.5; -5 \text{ if } W_{t+1} - W_t \geq -0.5; 0 \text{ o.w}$

$R_{t3} = 15, \text{if} M_{t+1} = 0; 5 \text{if } M_{t+1} - M_t \leq -0.5, \text{ but } M_{t+1} \neq 0; -5, \text{if } M_{t+1} - M_t \geq 0.5; 0, \text{ o.w.}$

Overall

The trajectories / training data generated according to the ODE model

$$\{S_{0i}, A_{0i}, R_{0i}, S_{1i}, \cdots, S_{5i}, A_{5i}, R_{5i}, S_{6i}\}_{i=1}^{N}$$

where action is the does level $A_t = D_t$, $S_t = (M_t, W_t, F_t)$, discount factor $\gamma = 0.8$.

Least-square policy iteration

LSPI

Linear architectures, where Q is approximated by a linear parametric combination of k basis functions (features

ϕ_j):

$$\widehat{Q}^{\pi}(s, a; \omega) = \sum_{j=1}^{k} \phi_j(s, a)\widehat{\omega}_j$$

LSQ step in LSPI: Least-Squares Fixed-Point Approximation

1. Force the approximate Q function to be a fixed point under the Bellman operator: $T\widehat{Q}^{\pi} \approx \widehat{Q}^{\pi}$.

   Bellman operator T: $TQ^{\pi}(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s, a, s') \sum_{a' \in \mathcal{A}} \pi(a', s')Q(s', a')$. Bellman residual minimizing approximation is another choice.

2. A sample (s, a, r, s') contributes to the approximation:

$$\widehat{A} \leftarrow \widehat{A} + \phi(s, a)\{\phi(s, a)^{\intercal} - \gamma\phi(s', \pi(s'))^{\intercal}\},$$

$$\widehat{b} \leftarrow \widehat{b} + \phi(s, a)r$$

3. Solve the linear system for ω^{π},

$$A\omega^{\pi} = b$$

LSPI is completed by choosing the policy $\pi(s') = max_a \widehat{Q}(s', a; \widehat{\omega})$, here $A_t = D_t$ is continuous.

Basis function

Basis function construction

$$1$$

$$exp[-c_m(M - medianM)^2]$$

$$exp[-c_w(W - medianW)^2]$$

$$d$$

$$d^2$$

$$d * exp[-c_m(M - medianM)^2]$$

$$d * exp[-c_w(W - medianW)^2])$$

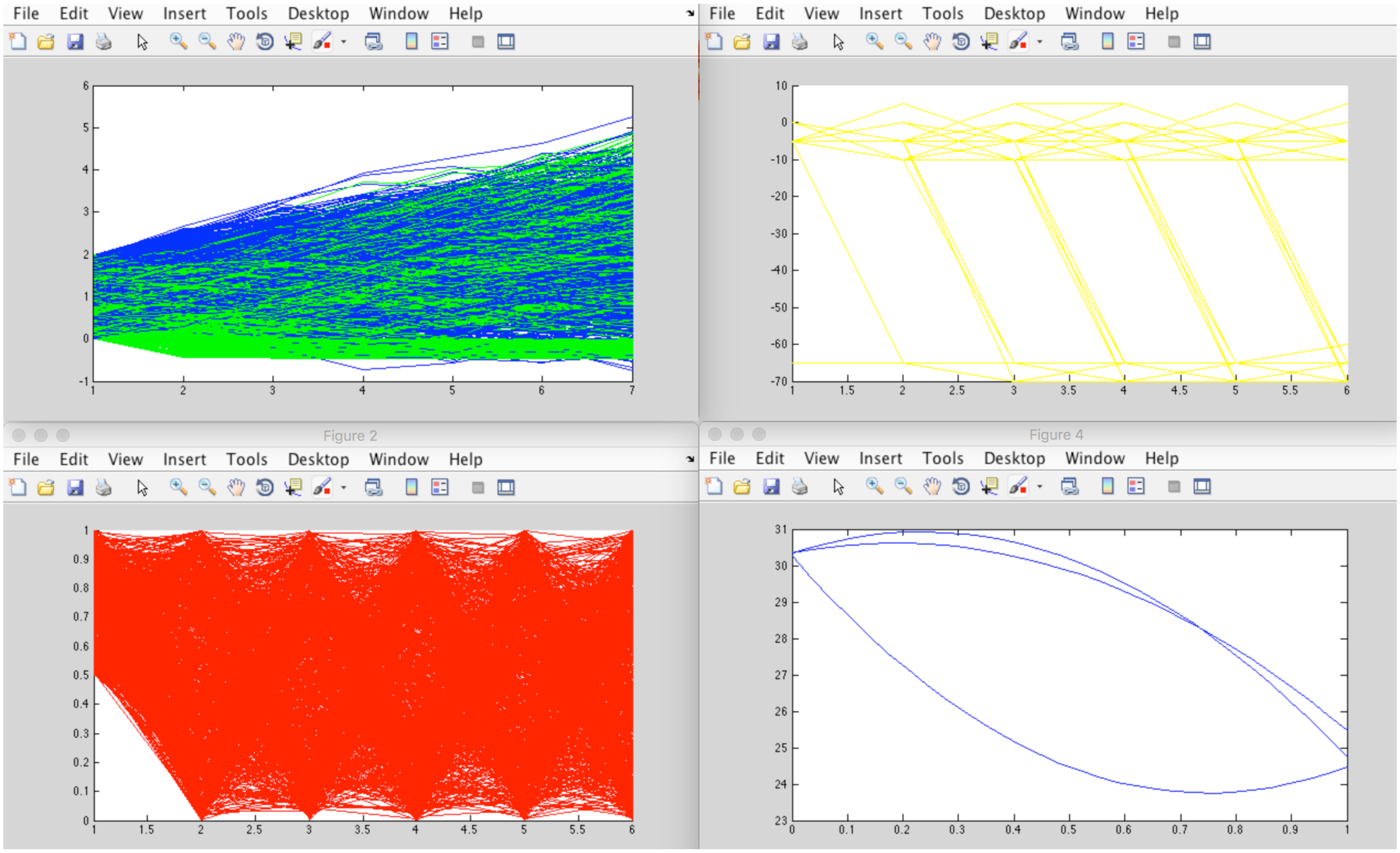
$$d^2 * exp[-c_m(M - medianM)^2]$$

$$d^2 * exp[-c_w(W - medianW)^2])$$

Action A_t = Dose level $D_t \in [0, 1]$

```
function phi = feature ( m, w, d, f, med_m, med_w, k)
    cnst_m = 0.05;
    cnst_w = 0.025;
    if ( f == 1 )
        % absorb state
        phi = zeros(k, 1);
    else
        phi = [ 1;...
                exp( -cnst_m * (m - med_m)^2); ...
                exp( -cnst_w * (w - med_w)^2); ...
                d; ...
                d^2; ...
                d * exp( -cnst_m * (m - med_m)^2); ...
                d * exp( -cnst_w * (w - med_w)^2);...
                d^2 * exp( -cnst_m * (m - med_m)^2); ...
                d^2 * exp( -cnst_w * (w - med_w)^2) ];
    end
end
```

Plots



Left top : Wellness (blue) / Tumor Size (green) vs. time

Left bottom: Dose vs. time

Right top: Reward vs. time

Right bottom: Estimated Q-val vs. D dose (M and W: mean, 25th quantile, 75 quantile, F=0)