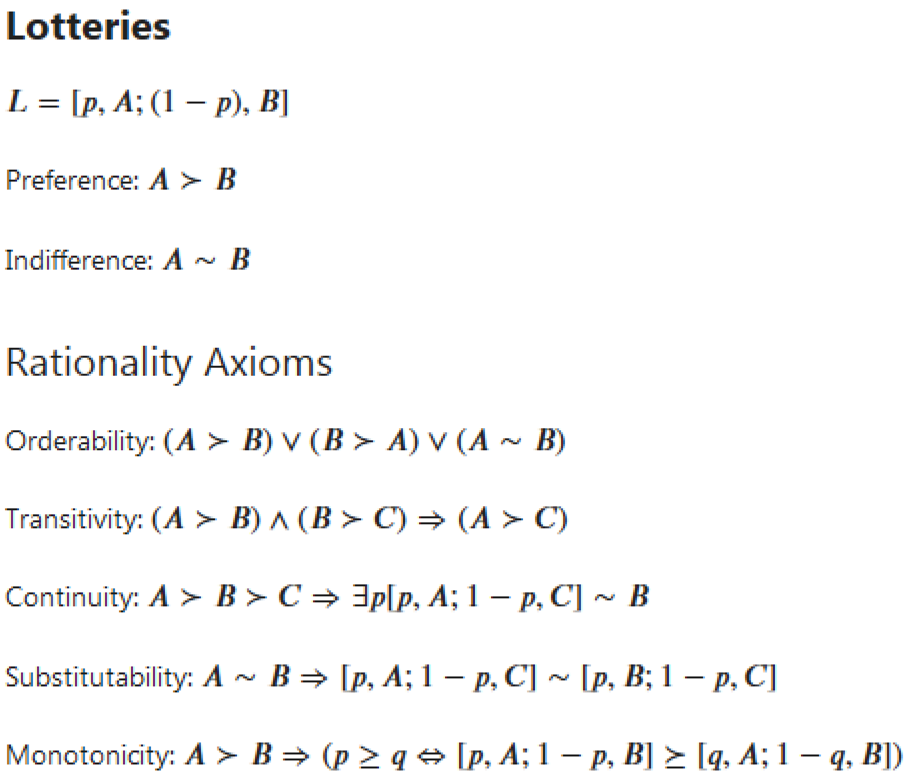
<u>Some Formulas You May or May Not Need:</u>



Lotteries

$L = [p, A; (1 - p), B]$

Preference: $A \succ B$

Indifference: $A \sim B$

Rationality Axioms

Orderability: $(A \succ B) \vee (B \succ A) \vee (A \sim B)$

Transitivity: $(A \succ B) \wedge (B \succ C) \Rightarrow (A \succ C)$

Continuity: $A \succ B \succ C \Rightarrow \exists p [p, A; 1 - p, C] \sim B$

Substitutability: $A \sim B \Rightarrow [p, A; 1 - p, C] \sim [p, B; 1 - p, C]$

Monotonicity: $A \succ B \Rightarrow (p \geq q \Leftrightarrow [p, A; 1 - p, B] \succsim [q, A; 1 - q, B])$

Bellman Equations

$V^*(s) = max_a Q^*(s,a)$

$Q^*(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$

$V^*(s) = max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$

Value Iteration

$V_{k+1}(s) \leftarrow max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$

Policy Evaluation

$V_0^\pi(s) = 0$

$V_{k+1}^\pi(s) \leftarrow \sum_{s'} T(s,\pi(s),s')[R(s,\pi(s),s') + \gamma V_k^\pi(s')]$

Policy Extraction

$\pi^*(s) = argmax_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$

Policy Improvement

$\pi_{i+1}(s) = argmax_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^{\pi_i}(s')]$

Reinforcement Learning

Temporal Difference Learning

Sample of V(s): $sample = R(s, \pi(s), s') + \gamma V^{\pi}(s')$

Update to V(s): $V^{\pi}(s) \leftarrow (1 - \alpha)V^{\pi}(s) + (\alpha)sample$

Same Update, Rewritten: $V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha(sample - V^{\pi}(s))$

Q-Learning

$Q_{k+1}(s, a) \leftarrow \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma max_{a'} Q_k(s', a')]$

$\text{transition} = (s, a, r, s')$

$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + (\alpha)[r + \gamma max_{a'} Q(s', a')]$

Approximate Q-Learning

$V(s) = w_1 f_1(s) + w_2 f_2(s) + \ldots + w_n f_n(s)$

$Q(s, a) = w_1 f_1(s, a) + w_2 f_2(s, a) + \ldots + w_n f_n(s, a)$

$\text{transition} = (s, a, r, s')$

$\text{difference} = [r + \gamma max_{a'} Q(s', a')] - Q(s, a)$

$Q(s, a) \leftarrow Q(s, a) + \alpha[\text{difference}]$

$w_i \leftarrow w_i + \alpha[\text{difference}] f_i(s, a)$