1. $A = \{a_{0:L-1}\}$

$b_k = p(x_k | a_{0:k-1}, z_{1:k})$

$p(x_k | x_{k-1}, a_{k-1})$; $p(z|x)$

$r(a,b) = \int_x r(x,a) b(x) dx$

(a)

$$J(b_k, a_{k:k+L-1}) = \mathbb{E}_{z_{k+1:k+L}} \left\{ \sum_{l=0}^{L-1} r(b_{k+l}, a_{k+l}) + r_T(b_k) \right\}$$

(b) $J(b_0, a_{0:L-1}) = \mathbb{E}_{z_{1:L}} \left\{ \sum_{l=0}^{L-1} r(b_l, a_l) + r_T(b_l) \right\} =$

$= \mathbb{E}\{r(b_0, a_0)\} + \mathbb{E}_{z_{1:L}} \left\{ \sum_{l=1}^{L-1} r(b_l, a_l) + r_T(b_l) \right\} =$

linearity

$= r(b_0, a_0) + \int p(z_{1:L} | b_0, a_{0:L-1}) \left[\sum_{l=1}^{L-1} r(b_l, a_l) + \right.$

sufficient statistics

$\left. + r(b_L) \right] dz_{1:L} =$

$$= r(b_0, a_0) + \int_{z_{1:L}} p(z_1 | b_0, a_0) p(z_{2:L} | b_0, a_{1:L-1}, z_1) \left[\sum_{l=1}^{L-1} (b_l, a_l) + \right.$$

chain rule + indep.

$$\left. + r(b_L) \right] dz_{1:L} =$$

$$= r(b_0, a_0) + \int_{z_1} \int_{z_{2:L}} p(z_1 | b_0, a_0) p(z_{2:L} | b_0, a_{1:L-1}, z_1) \cdot$$

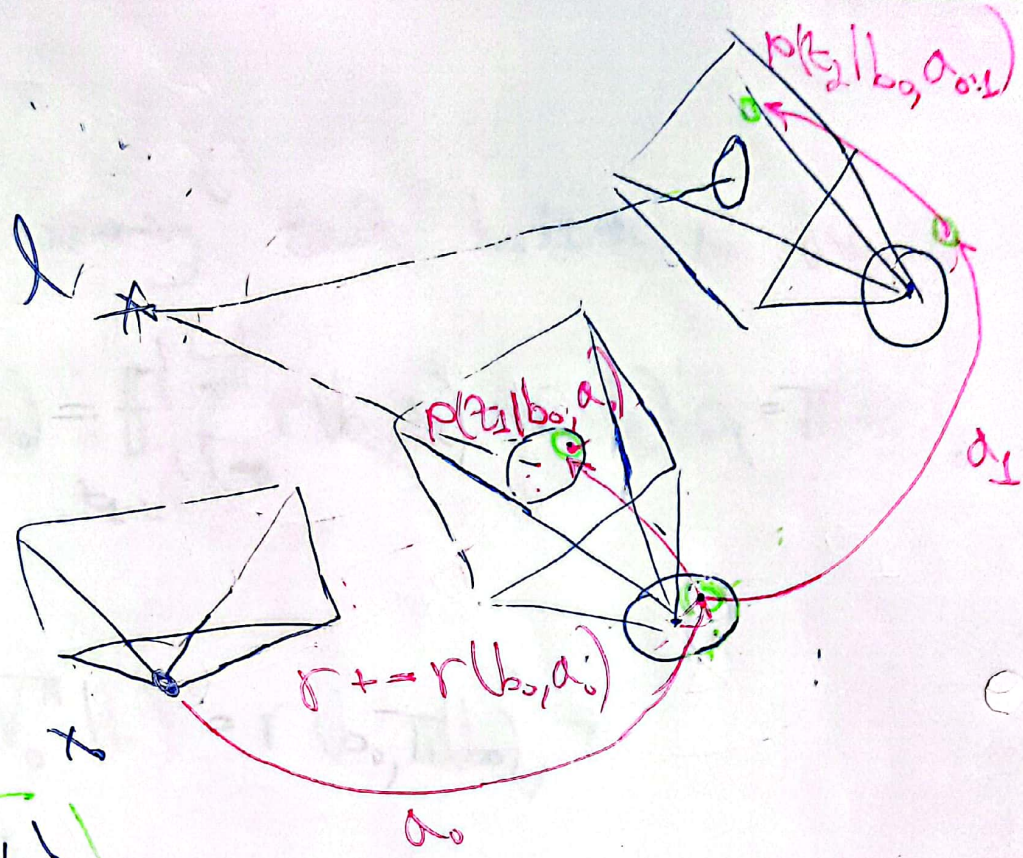
$$\cdot \left[\sum_{l=1}^{L-1} (b_l, a_l) + r(b_L) \right] dz_{2:L} dz_1 =$$

$$= r(b_0, a_0) + \int_{z_1} p(z_1 | b_0, a_0) \int_{z_{2:L}} p(z_{2:L} | b_0, a_{1:L-1}, z_1) \cdot$$

indep.

$$\cdot \left[\sum_{l=1}^{L-1} (b_l, a_l) + r(b_L) \right] dz_{2:L} dz_1 =$$

$$= r(b_0, a_0) + \int_{z_1} p(z_1 | b_0, a_0) J(b_1 | z_1, a_{1:L-1}) dz_1 =$$



$b_k = \{x_k\}$

$$J(b_0, a_{0:1}) = r(b_0, a_0) + \int_{z_1} p(z_1|b_0, a_0) \cdot J(b_1|z_1, a_1) \, dz_1$$

$\propto$ probability of objective fcn

value of objective fcn

$$p(b_1|z_1) = \eta p(b_0) p(b_1|b_0, a_0) p(z_1|b_1)$$

belief propagation

(c).

(i) assuming finite horizon, no discount γ:

$$V_0^{\pi}(b_0) = \mathop{\mathbb{E}}_{z_{1:L}}\left\{\sum_{l=0}^{L-1} r(b_l, a_l) + r_T(b_L) \,\middle|\, a_l = \pi(b_l)\right\}$$

(ii) $$V_0^{\pi}(b_0) = r(b_0, \pi(b_0)) +$$

$$+ \mathbb{E}\left\{\sum_{l=1}^{L-1} (r_{b_l}, a_l) + r_T(b_L) \,\middle|\, a_l = \pi(b_l)\right\} =$$

$$= r(b_0, \pi(b_0)) + \int_{z_{1:L}} p(z_{1:L} | b_0, \pi(b_{0:L-1}) \left[\sum_{l=1}^{L-1} r(b_l, \pi(b_l)) + \right.$$

$$\left. + r(b_L)\right] dz_{1:L} =$$

$$= r(b_0, \pi(b_0)) + \int_{z_{1:L}} p(z_1 | b_0, \pi(b_0)) p(z_{2:L} | b_0, \pi(b_{0:L-1}, z_1) \cdot$$

$$\cdot \left[\sum_{l=1}^{L-1} r(b_l, \pi(b_l)) + r(b_L)\right] dz_{L:L} =$$

$$= r(b_0, \pi(b_0)) + \int_{z_1} p(z_1 | b_0, \pi(b_0)) \int_{z_{2:L}} p(z_{2:L} | b_0, \pi(b_{0:L-1}, z_1)) \cdot$$

$$\cdot \left[\sum_{l=1}^{L-1} r(b_l, \pi(b_l)) + r(b_L)\right] dz_{2:L} dz_1 =$$

$$= r(b_0, \pi(b_0)) + \int_{z_1} p(z_1 | b_0, \pi(b_0)) V_1^{\pi}(b_1) dz_1 .$$

(iii) in the objective fun formulation J, we assume a set of known actions.

The value fun V^{π} uses a policy $\pi(b)$ to create actions. As such, actions are generated online, and π can be optimized online.