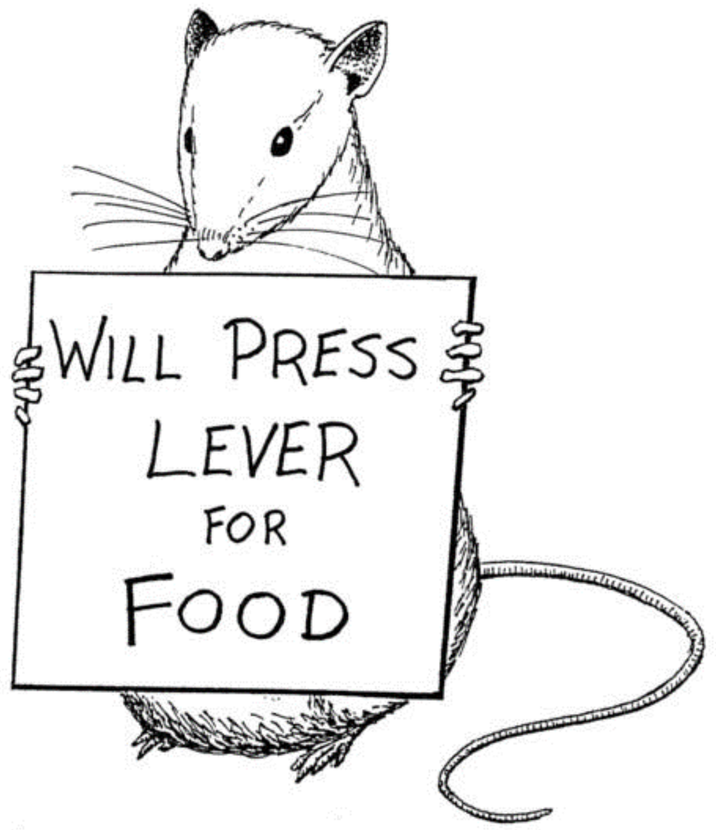
CS 277: Control and Reinforcement Learning

Winter 2021

Lecture 15: Control as Inference

Roy Fox
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine



Today's lecture

Linearly Solvable MDPs

Bounded RL

SQL, SAC

Bounded optimality

- Bounded optimizer = trades off value and divergence from prior $\pi_0(a \mid s)$

$$\max_\pi \mathbb{E}_{s,a\sim p_\pi}[r(s,a)] - \tau \mathbb{D}[\pi \| \pi_0] = \max_\pi \mathbb{E}_{s,a\sim p_\pi}\left[\beta r(s,a) - \log \frac{\pi(a \mid s)}{\pi_0(a \mid s)}\right]$$

- $\beta = \frac{1}{\tau}$ is the tradeoff coefficient between value and relative entropy
  - Similar to the inverse-temperature in thermodynamics
  - As $\beta \to 0$, the agent will fall back to the prior $\pi \to \pi_0$
  - As $\beta \to \infty$, the agent will be a perfect value optimizer $\pi \to \pi^*$
- We'll see reasons to have finite β

Simplifying assumption

- MaxEnt IRL was approximate because it violated dynamical constraints
  - $p_\pi(\xi) \propto \exp(R(\xi))$ (regardless of trajectory feasibility)
- For simplicity, let's do the same for RL
  - Suppose the environment is fully controllable $s_{t+1} = a_t$
  - Bellman equation:

$$V^*_\beta(s) = \max_\pi \mathbb{E}_{s'|s\sim\pi}\left[r(s) - \frac{1}{\beta}\log\frac{\pi(s'|s)}{\pi_0(s'|s)} + \gamma V^*_\beta(s')\right]$$

$$= r(s) - \frac{1}{\beta}\min_\pi \mathbb{D}\left[\pi \,\middle\|\, \frac{\pi_0(s'|s)\exp(\beta\gamma V^*_\beta(s'))}{Z'_\beta(s)}\right] + \frac{1}{\beta}\log Z'_\beta(s)$$

Soft-greedy policy

- To solve the Bellman recursion

$$V_\beta^*(s) = \max_\pi \mathbb{E}_{s'|s\sim\pi}\left[r(s) - \frac{1}{\beta}\log\frac{\pi(s'|s)}{\pi_0(s'|s)} + \gamma V_\beta^*(s')\right]$$

$$= r(s) - \frac{1}{\beta}\min_\pi \mathbb{D}\left[\pi \,\middle\|\, \frac{\pi_0(s'|s)\exp(\beta\gamma V_\beta^*(s'))}{Z'_\beta(s)}\right] + \frac{1}{\beta}\log Z'_\beta(s)$$

  - Differentiate, with λ_s constraining $\sum_{s'} \pi(s'|s) = 1$

$$0 = \nabla_{\pi(s'|s)}\mathbb{E}_{s'|s\sim\pi}\left[-\frac{1}{\beta}\log\frac{\pi(s'|s)}{\pi_0(s'|s)} + \gamma V_\beta^*(s') - \lambda_s\right]$$

$$= -\frac{1}{\beta}\log\frac{\pi(s'|s)}{\pi_0(s'|s)} + \gamma V_\beta^*(s') - \lambda_s - \pi(s'|s)\,\nabla_{\pi(s'|s)}\log\pi(s'|s)$$

- Soft-greedy policy: $\pi_\beta^*(s'|s) \propto \pi_0(s'|s)\exp(\beta\gamma V_\beta^*(s'))$ (more general form later)

Linearly-Solvable MDPs (LMDPs)

- Plugging the soft-greedy policy back into the value recursion:

$$V_\beta^*(s) = r(s) - \frac{1}{\beta} \min_\pi \mathscr{D}\left[\pi \,\middle\|\, \frac{\pi_0(s'|s)\exp(\beta\gamma V_\beta^*(s'))}{Z'_\beta(s)} + \frac{1}{\beta}\log Z'_\beta(s)\right]$$

$$= r(s) + \frac{1}{\beta}\log Z'_\beta(s) = r(s) + \frac{1}{\beta}\log \mathbb{E}_{s'|s\sim\pi_0}[\exp(\beta\gamma V_\beta^*(s'))]$$

- Alternatively:

$$Z_\beta(s) = \exp(\beta V_\beta^*(s)) = \exp(\beta r(s)) Z'_\beta(s) = \exp(\beta r(s))\mathbb{E}_{s'|s\sim\pi_0}[Z_\beta^\gamma(s')]$$

- In the undiscounted case $\gamma = 1$, with $D = \operatorname{diag}(\exp \beta r)$: $z = DP_0 z$

- We can solve for z, and therefore π, by finding a right-eigenvector of DP_0

Z-learning

$$Z(s) = \exp(\beta r(s))\mathbb{E}_{s'|s\sim\pi_0}[Z^\gamma(s')]$$

- We can do the same model-free:
  - Given experience (s, r, s') sampled by the prior policy π_0
  - Update $Z_\beta(s) \to \exp \beta r Z^\gamma(s')$
- Full-controllability condition ($s_{t+1} = a_t$) can be relaxed to allow $\pi_0(s' \mid s) = 0$
  - But we still allow any transition distribution $\pi(s' \mid s)$ over the remaining support
  - Later: the general case, $p(s' \mid s) = \sum_a \pi(a \mid s) p(s' \mid s, a)$

Today's lecture

Linearly Solvable MDPs

Bounded RL

SQL, SAC

Duality between value and log prob

- We've seen many cases where log-probs play the role of reward / value
  - Or values the role of logits (unnormalized log-probs)
- Examples:
  - In LQG, $\log p(x|\hat{x}) = -\frac{1}{2}x^\intercal \Sigma x + \text{const}$; costs / values are quadratic
  - In value-based algorithms, a good exploration policy is $\pi(a|s) = \underset{a}{\text{sm}}\, \beta Q(s,a)$
  - Imitation Learning can be viewed as RL with $r(s,a) = \log \pi_T(a|s)$
  - In IRL, a reward function can be viewed as a discriminator $D(s) = \exp r(s)$
  - etc.

Full-controllability duality

$$Z(s) = \exp(\beta r(s))\, \mathbb{E}_{s'|s \sim \pi_0}[Z^\gamma(s')]$$

- Backward filtering in a partially observable system with dynamics $\pi_0(s'|s)$

$$p(o_{\geqslant t}|s_t) = p(o_t|s_t)\, \mathbb{E}_{s_{t+1}|s_t \sim \pi_0}[p(o_{\geqslant t+1}|s_{t+1})]$$

- Equivalent if $r(s) = p(o|s)$ and $Z(s) = p(o_{\geqslant t}|s_t)$
  - With the actual observations that we see

- Can we say anything about the partially controllable case?

Bounded RL

- Back to the general case: $\max_\pi \mathbb{E}_{s,a\sim p_\pi}[\beta r(s,a)] - \mathbb{D}[\pi\|\pi_0]$

- Define an entropy-regularized Bellman optimality operator

$$\mathcal{B}[V](s) = \max_\pi \mathbb{E}_{a|s\sim\pi}\left[r(s,a) - \frac{1}{\beta}\log\frac{\pi(a|s)}{\pi_0(a|s)} + \gamma\,\mathbb{E}_{s'|s,a\sim p}[V(s')]\right]$$

  - As in the unbounded case $\beta \to \infty$, this operator is contracting

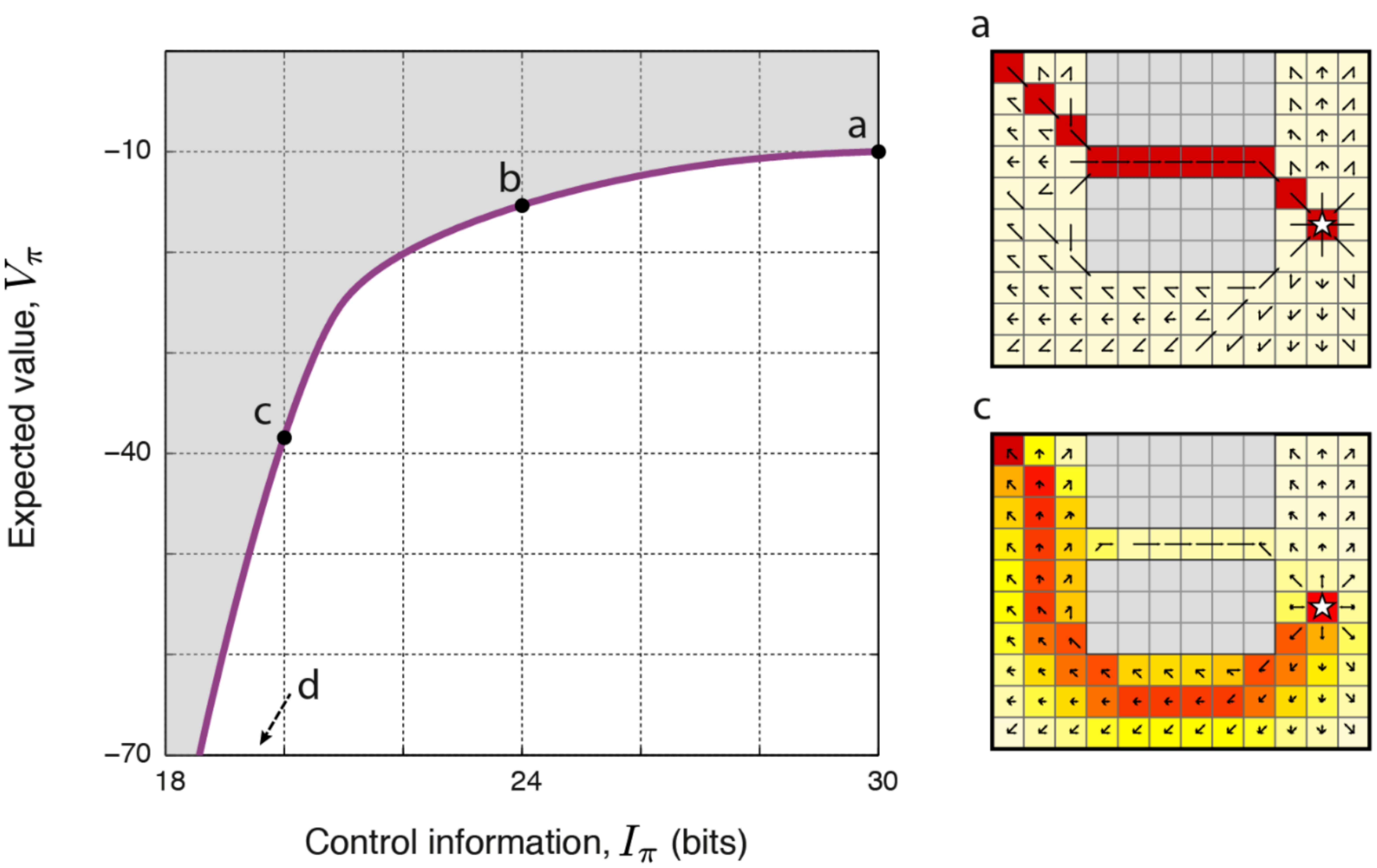
- Optimal policy:

$$\pi(a|s) \propto \pi_0(a|s)\exp\beta(r(s,a) + \gamma\,\mathbb{E}_{s'|s,a\sim p}[V(s')]) = \pi_0(a|s)\exp\beta Q(s,a)$$

- Optimal value recursion:

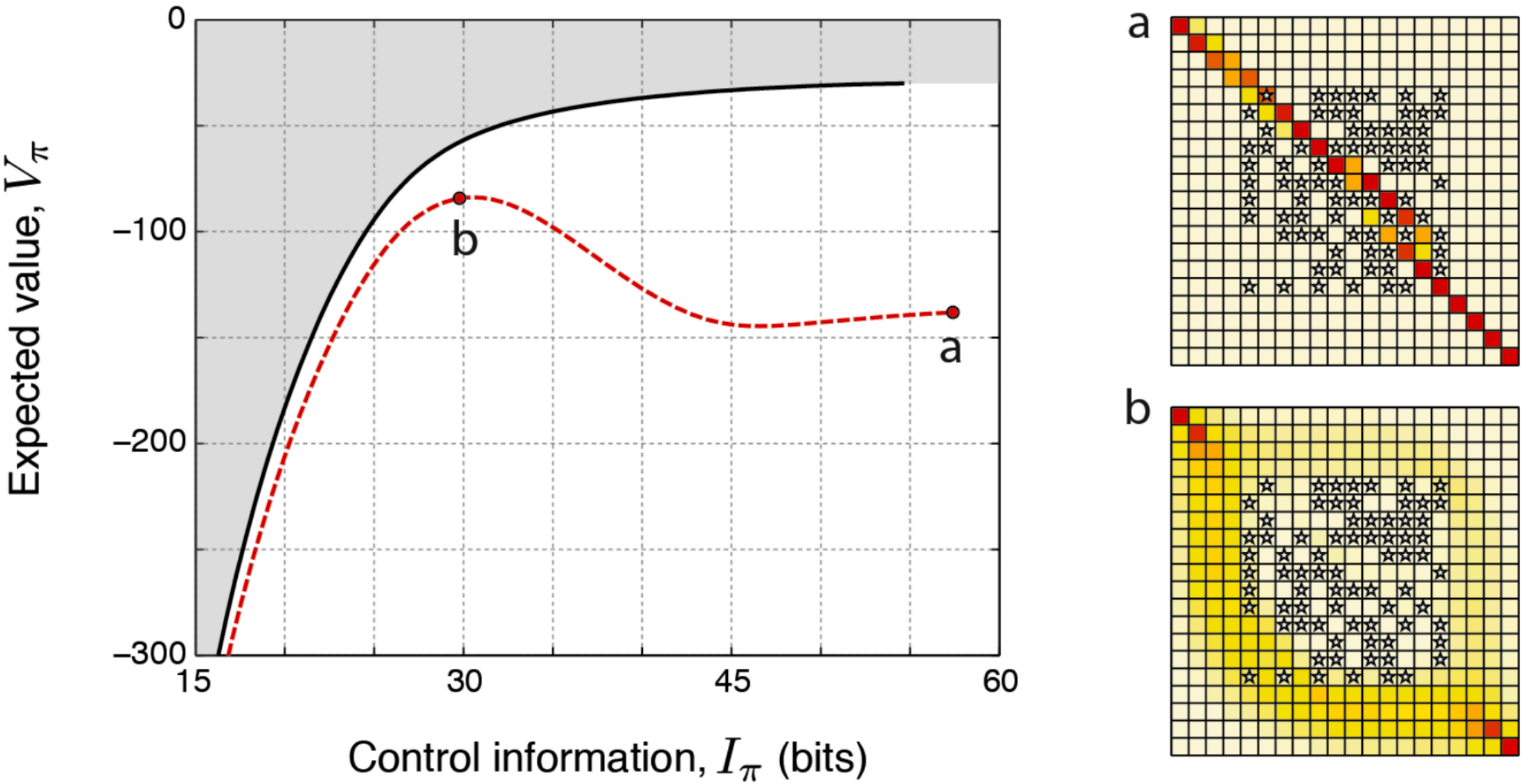
$$V(s) = \frac{1}{\beta}\log Z(s) = \frac{1}{\beta}\log\mathbb{E}_{a|s\sim\pi_0}[\exp\beta(r(s,a) + \gamma\,\mathbb{E}_{s'|s,a\sim p}[V(s')])]$$

Value–RelEnt curve



[Rubin et al., 2012]

Robustness to model uncertainty



Today's lecture

Linearly Solvable MDPs

Bounded RL

SQL, SAC

Exact and approximate inference

- Suppose we want to max log-likelihood of a dataset $\max_\theta \mathbb{E}_{x\sim\mathcal{D}}[\log p_\theta(x)]$
  - And computing it is easier with a latent intermediate variable $p_\theta(z)p_\theta(x|z)$
- Expectation–Gradient (EG):

$$\nabla_\theta \log p_\theta(x) = \mathbb{E}_{z|x\sim p_\theta}[\nabla_\theta \log p_\theta(z,x)]$$

- But what if sampling from the exact posterior $p_\theta(z|x)$ is also hard?
- Let's do importance sampling from any approximate posterior $q_\phi(z|x)$

$$\log p_\theta(x) = \log \mathbb{E}_{z|x\sim q_\phi}\left[\frac{p_\theta(z)}{q_\phi(z|x)}p_\theta(x|z)\right] \geqslant \mathbb{E}_{z|x\sim q_\phi}\left[\log \frac{p_\theta(z,x)}{q_\phi(z|x)}\right]$$

Variational Inference (VI): Evidence Lower Bound (ELBO)

- Two ways of decomposing $p_\theta(z, x)$:

$$\log p_\theta(x) \geqslant -\mathbb{D}[q_\phi(z|x) \| p_\theta(z, x)]$$
$$= \log p_\theta(x) + \mathbb{E}_{z|x \sim q_\phi}\left[\log \frac{p_\theta(z|x)}{q_\phi(z|x)}\right] \qquad (1)$$
$$= \mathbb{E}_{z|x \sim q_\phi}\left[\log \frac{p_\theta(z)}{q_\phi(z|x)} + \log p_\theta(x|z)\right] \qquad (2)$$

- (1) shows that the bounding gap is $\mathbb{D}[q_\phi(z|x) \| p_\theta(z|x)] \geqslant 0$
    - It is smaller the better we can approximate $p_\theta(z|x)$ using $q_\phi(z|x)$
- (2) shows how the bound can be computed efficiently
    - We can use it as a proxy for our objective

Control as inference

- Consider soft "success" indicators

$$p(v_t = 1 | s_t, a_t) = \exp \beta r(s_t, a_t)$$

- What is the log-probability that an entire trajectory ξ "succeeds"?

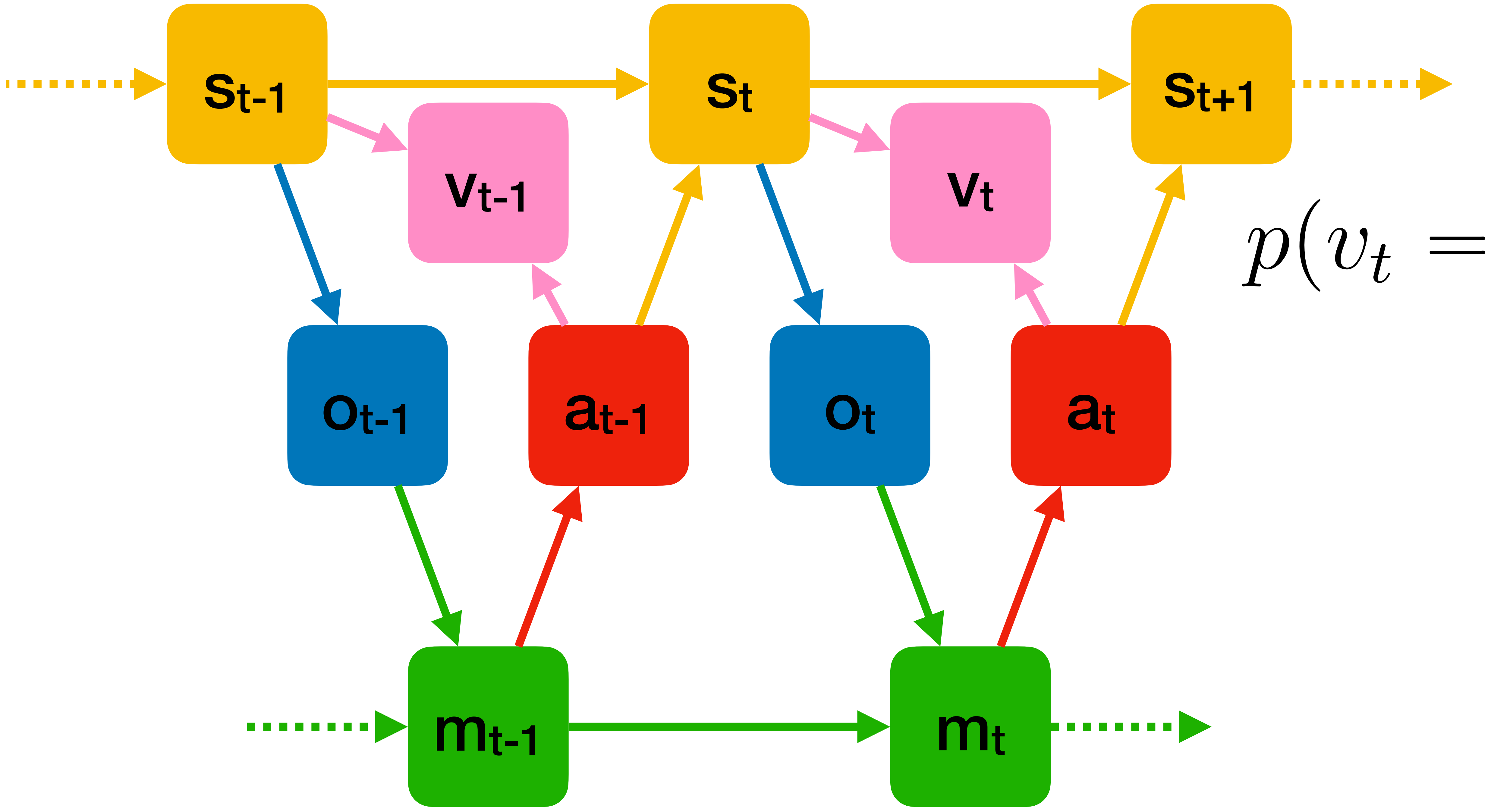
$$\log p(\mathcal{V} | \xi) = \sum_t \log p(v_t = 1 | s_t, a_t) = \beta \sum_t r(s_t, a_t) = \beta R$$

- What is the posterior distribution over trajectories, given success?

$$p(\xi | \mathcal{V}) = \frac{p_0(\xi) p(\mathcal{V} | \xi)}{p_0(\mathcal{V})} = \frac{p_0(\xi) \exp \beta R}{Z}$$

- But this distribution is not realizable, due to dynamical constraints

Pseudo-observations



$$p(v_t = 1 | s_t, a_t) = \exp \beta r_t$$

General duality between VI and bounded RL

- Take $x = \mathcal{V}$, $z = \xi$, and $p_\theta(\xi) = p_0(\xi)$
- Optimize the ELBO with a realizable proposal distribution $q_\phi(\xi|\mathcal{V}) = p_{\pi_\phi}(\xi)$
- The ELBO becomes

$$\mathbb{E}_{\xi|\mathcal{V}\sim q_\phi}\left[\log p_0(\mathcal{V}|\xi) + \log \frac{p_0(\xi)}{q_\phi(\xi|\mathcal{V})}\right] = \mathbb{E}_{\xi\sim p_{\pi_\phi}}\left[\beta R - \log \frac{p_{\pi_\phi}(\xi)}{p_0(\xi)}\right]$$
$$= \mathbb{E}_{s,a\sim p_{\pi_\phi}}\left[\beta r(s,a) - \log \frac{\pi_\phi(a|s)}{\pi_0(a|s)}\right]$$

  - Equivalent to the bounded RL problem!

Soft Q-Learning (SQL)

- TD off-policy algorithm for model-free bounded RL

- With tabular parametrization:

$$\Delta Q(s,a) = r + \frac{\gamma}{\beta} \log \mathbb{E}_{a'|s' \sim \pi_0}[\exp \beta Q(s',a')] - Q(s,a)$$

- With differentiable parametrization:

$$\mathcal{L}_\theta(s,a,r,s') = \left(r + \frac{\gamma}{\beta} \log \mathbb{E}_{a'|s' \sim \pi_0}[\exp \beta Q_{\bar{\theta}}(s',a')] - Q_\theta(s,a) \right)^2$$

- As $\beta \to \infty$, this becomes (Deep) Q-Learning

Soft Actor–Critic (SAC)

- AC off-policy algorithm for model-free bounded RL

- Optimally:

$$\pi(a|s) = \frac{\pi_0(a|s)\exp \beta Q(s,a)}{\exp \beta V(s)} \qquad \forall a: \; V(s) = Q(s,a) - \frac{1}{\beta}\log\frac{\pi(a|s)}{\pi_0(a|s)}$$

- We can train the critic off-policy

$$\mathcal{L}_\phi(s,a,r,s',a') = \left(r + \gamma\left(Q_{\bar\phi}(s',a') - \frac{1}{\beta}\log\frac{\pi_\theta(a'|s')}{\pi_0(a'|s')}\right) - Q_\phi(s,a)\right)^2$$

- And the actor to be soft-greedy = distill / imitate the critic

$$\mathcal{L}_\theta(s) = \mathbb{E}_{a|s\sim\pi_\theta}[\log \pi_\theta(a|s) - \log \pi_0(a|s) - \beta Q_\phi(s,a)]$$

- Allows continuous action spaces

Why use a finite β

- Model suboptimal agents / teachers
- Robustness to model misspecification / avoid overfitting
- Eliminate bias due to winner's curse
  - For $\beta \to \infty$ $\qquad \mathbb{E}[\max_a Q(a)] \geqslant \max_a \mathbb{E}[Q(a)]$
  - For $\beta \to 0$ $\qquad \mathbb{E}[\mathbb{E}_{a\sim\pi_0}[Q(a)]] = \mathbb{E}_{a\sim\pi_0}[\mathbb{E}[Q(a)]] \leqslant \max_a \mathbb{E}[Q(a)]$
  - Somewhere in between there must be an unbiased β
- More reasons...

Recap

- Duality: rewards and values are like log-probs
- Can use inference methods to plan and learn
- Fall back to “optimal” methods in the 0-temperature case
- But many reasons to keep finite temperature, during training and often after