

CS 277: Control and Reinforcement Learning

Winter 2026

Lecture 14: Bounded RL

Roy Fox

Department of Computer Science

School of Information and Computer Sciences

University of California, Irvine



Logistics

assignments

- Quiz 7 due **next Monday**
- Exercise 4 due **next Tuesday**

Today's lecture

Bounded RL

Control as inference

Bounded RL methods

Bounded optimality

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

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

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

Simplifying assumption

- **MaxEnt IRL** was approximate because it violated dynamical constraints
 - $p_{\pi}(\xi) \propto \pi_0(\xi)\exp(R(\xi))$, regardless of trajectory **feasibility**
- For simplicity, let's start with the same for RL
 - Suppose the environment is **fully controllable** $s_{t+1} = a_t$
 - **Bellman equation**:

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

Linearly-Solvable MDPs (LMDPs)

- Optimal policy for $V_\beta(s) = r(s) - \frac{1}{\beta} \min_{\pi} \mathbb{D} \left[\pi \parallel \frac{\pi_0(s'|s) \exp(\beta \gamma V_\beta(s'))}{Z'_\beta(s)} \right] + \frac{1}{\beta} \log Z'_\beta(s)$:

 - ▶ **Soft-greedy policy:** $\pi_\beta(s'|s) \propto \pi_0(s'|s) \exp(\beta \gamma V_\beta(s'))$

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

not Z'

$$Z_\beta(s) \stackrel{\text{def}}{=} \exp(\beta V_\beta(s)) = \exp(\beta r(s)) Z'_\beta(s) = \exp(\beta r(s)) \mathbb{E}_{(s'|s) \sim \pi_0} [Z'_\beta(s')]$$

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

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

[Todorov, Linearly-solvable Markov decision problems, NeurIPS 2006]

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** $Z(s) \rightarrow \exp(\beta r) Z^\gamma(s')$
- **Full-controllability** condition ($s_{t+1} = a_t$) can be relaxed with some $\pi_0(s' | s) = 0$
 - But we still allow **any transition** distribution $\pi(s' | s)$ over the remaining support
 - Later: the general case, $p(s' | s) = \sum_a \pi(a | s) p(s' | s, a)$

[Todorov, Linearly-solvable Markov decision problems, NeurIPS 2006]

Today's lecture

Bounded RL

Control as inference

Bounded RL methods

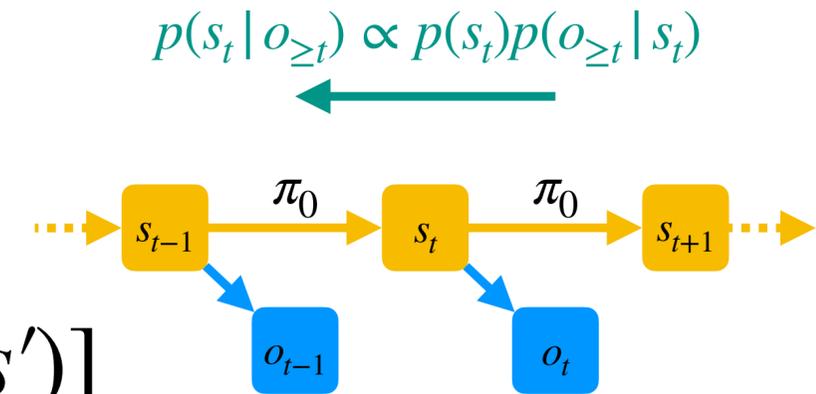
Duality between value and log prob

- 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**, $\log p(x | \hat{x}) = -\frac{1}{2}x^T \Sigma x + \text{const}$; costs / values are quadratic
 - In **value-based** algorithms, good **exploration** policy: $\pi(a | s) = \underset{a}{\text{softmax}} \beta Q(s, a)$
 - **Imitation Learning** can be viewed as RL with $\tilde{r}(s, a) = \log p^*(a | s)$
 - In **IRL**, a reward function can be viewed as a discriminator $D(s) = \exp(-\tilde{r}(s))$

Full-controllability duality

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

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



- **Bounded control** in LMDP: $Z(s) = \exp(\beta r(s)) \mathbb{E}_{(s'|s) \sim \pi_0} [Z'(s')]$

- These two are **equivalent** if $Z(s) = p(o_{\geq t} | s_t)$ and $\exp(\beta r(s)) = p(o | s)$

- ▶ **Intuition**: find states that give good reward \Leftrightarrow high likelihood of observations

- Exact equivalence only in the **fully-controllable** case

- ▶ **Partially controllable** case takes more nuanced analysis

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

$$\mathcal{T}[V](s) = \max_{\pi} \mathbb{E}_{(a|s) \sim \pi} \left[r(s, a) - \frac{1}{\beta} \log \frac{\pi(a|s)}{\pi_0(a|s)} + \gamma \mathbb{E}_{(s'|s,a) \sim p} [V(s')] \right]$$

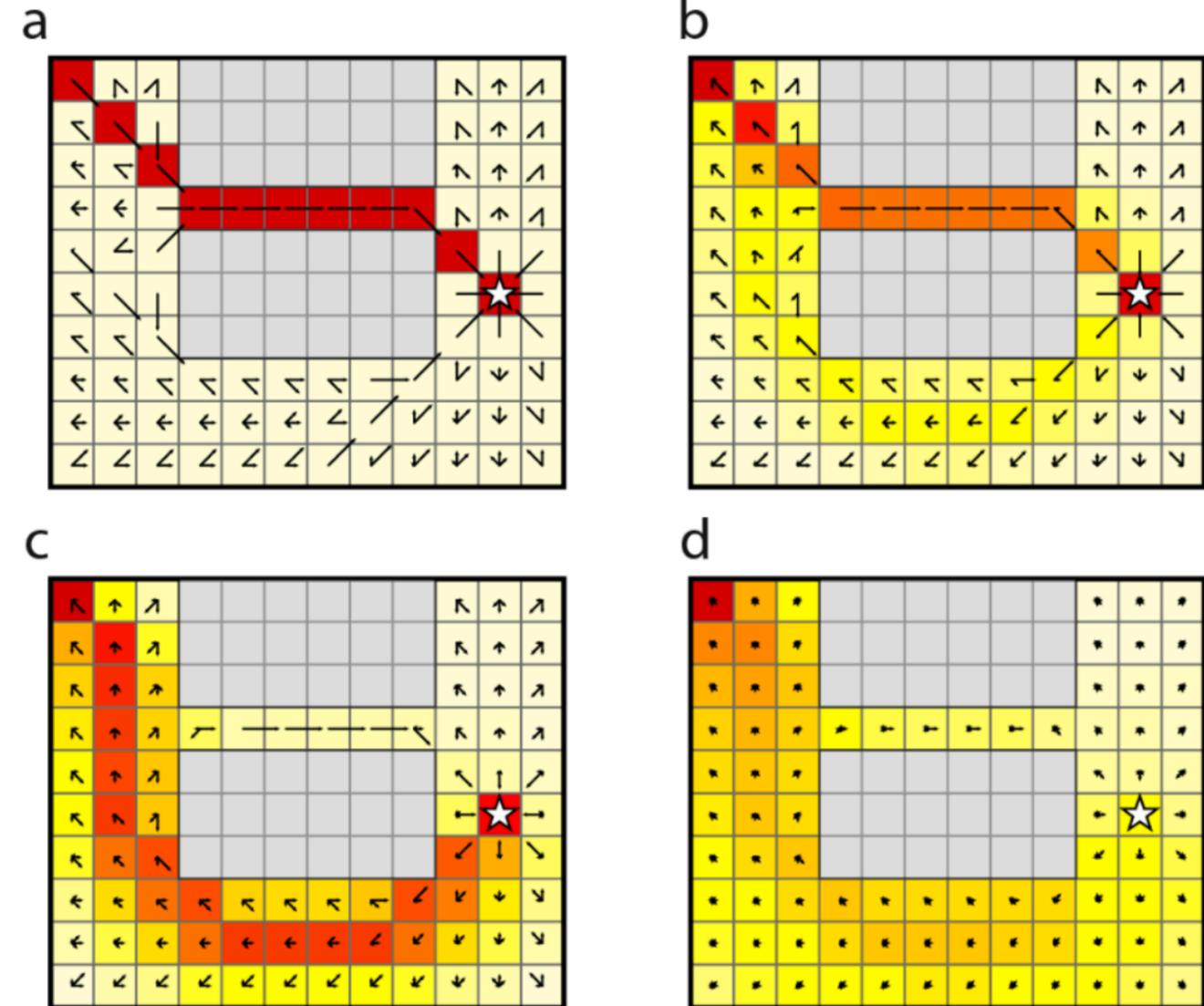
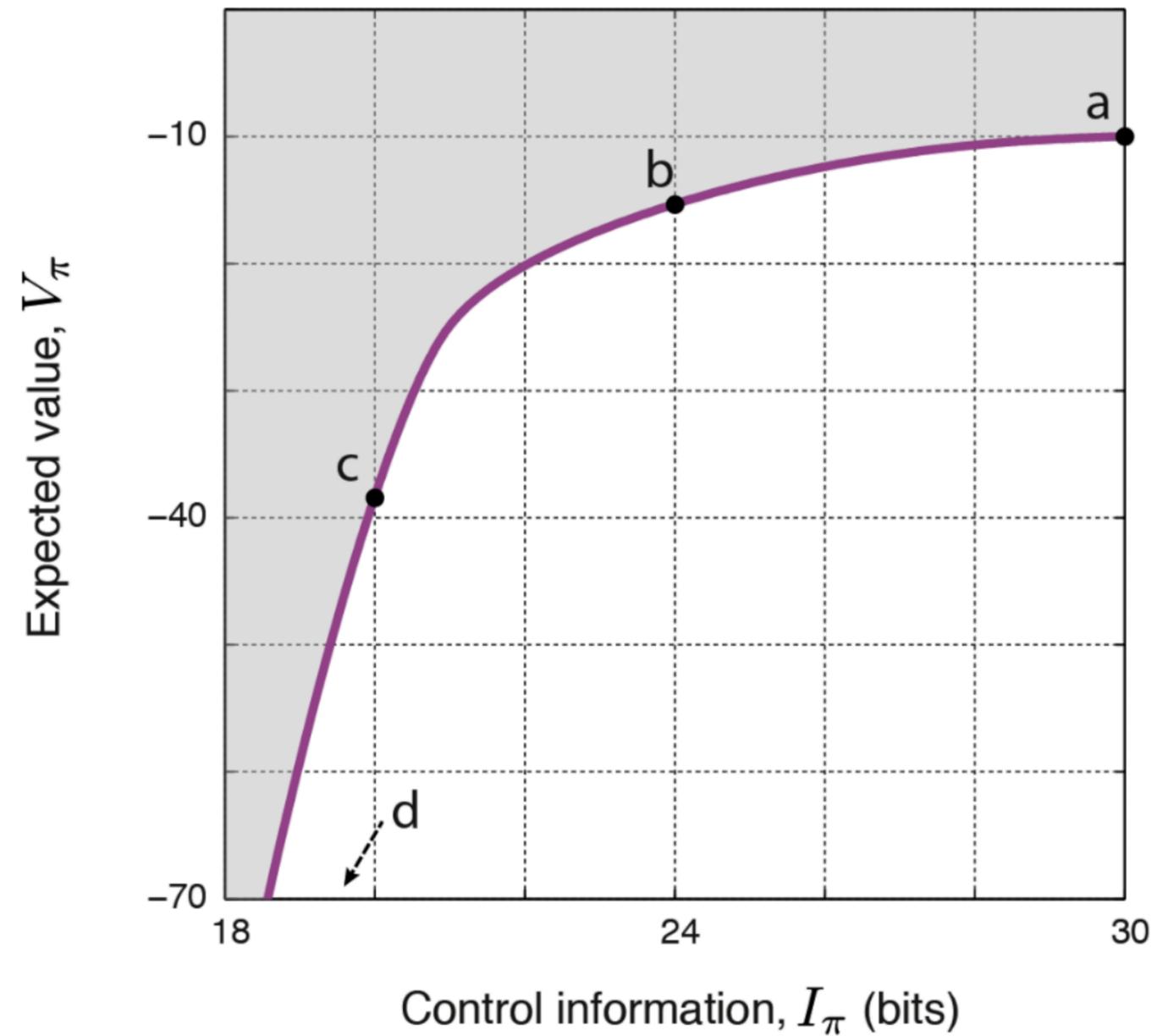
- As in the unbounded case ($\beta \rightarrow \infty$), this operator is **contracting**

- **Soft-optimal policy:**

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

- **Soft-optimal value recursion:** $V(s) = \frac{1}{\beta} \log Z(s) = \frac{1}{\beta} \log \mathbb{E}_{(a|s) \sim \pi_0} [\exp \beta Q(s, a)]$

Value-RelEnt curve

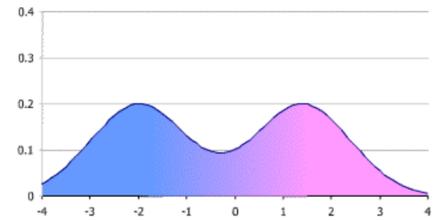


[Rubin et al., Trading Value and Information in MDPs, 2012]

Exact and approximate inference

- Suppose we want to **max log-likelihood** of a dataset $\max_{\theta} \mathbb{E}_{x \sim D}[\log p_{\theta}(x)]$

- ▶ With **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** $q_{\phi}(z | x)$

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

↙ $-\mathbb{D}[q_{\phi}(z|x) \| p_{\theta}(z, x)]$

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

- Two ways to decompose $p_\theta(z, x)$:

$$\log p_\theta(x) \geq -\mathbb{D}[q_\phi(z|x) \| p_\theta(z, x)]$$

objective

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

bounding gap between objective and proxy

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

what we actually compute

- Bounding gap:** $\mathbb{D}[q_\phi(z|x) \| p_\theta(z|x)] \geq 0$
 - The smaller the gap, the better the **guide** $q_\phi(z|x)$ **approximates** $p_\theta(z|x)$
- Bound (RHS) can be computed **efficiently** as a **proxy** for our objective

Control as inference

- Consider soft “success” indicators (assuming $r \leq 0$)

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

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

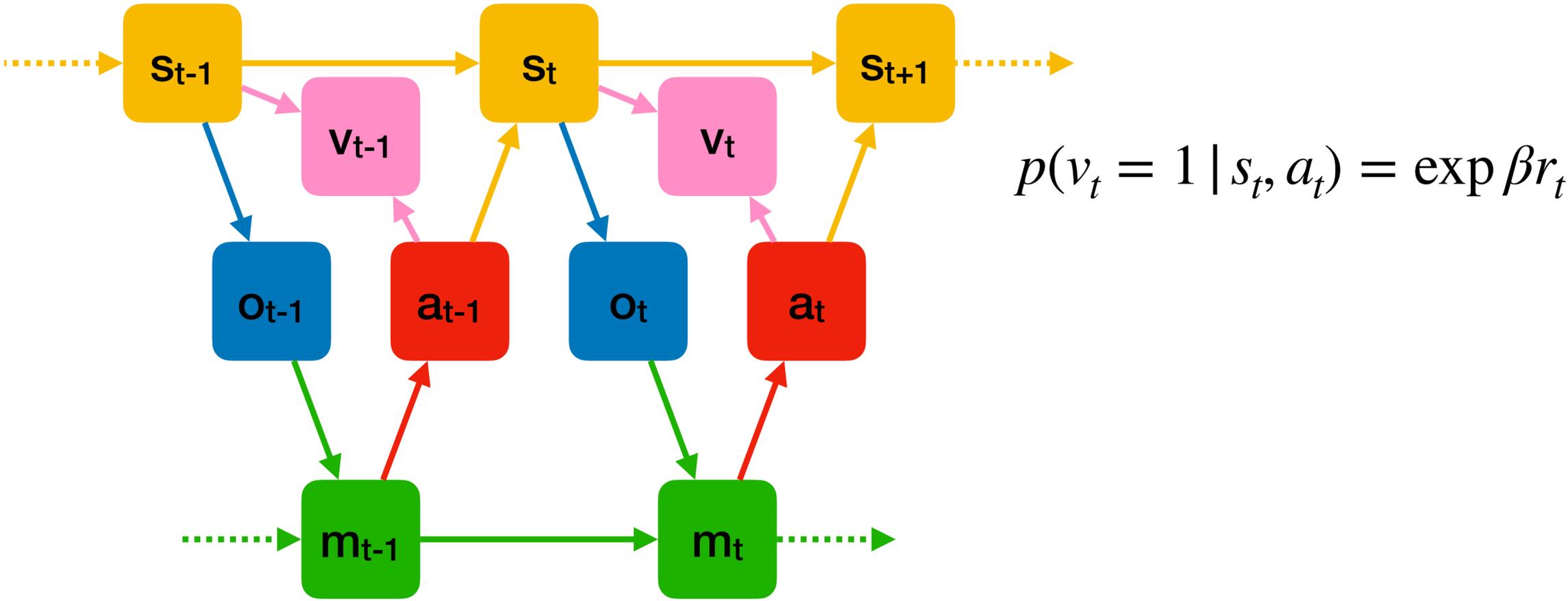
$$\log p(\mathcal{V} | \xi) = \sum_t \log p(v_t = 1 | s_t, a_t) = \beta \sum_t r(s_t, a_t) = \beta R(\xi)$$

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

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

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

Pseudo-observations



General duality between VI and bounded RL

- In VI, take $x = \mathcal{V}$, $z = \xi$, and $p_\theta(\xi) = p_0(\xi)$ (fix generator to prior)
- Optimize the ELBO with a realizable **guide distribution** $q_\phi(\xi | \mathcal{V}) = p_{\pi_\phi}(\xi)$

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

- ▶ Equivalent to the **bounded RL problem!** (a.k.a.: MaxEnt RL, energy-based RL)

Today's lecture

Bounded RL

Control as inference

Bounded RL methods

Soft Q-Learning (SQL)

- MaxEnt Bellman operator:

$$\mathcal{T}[Q](s, a) = r(s, a) + \gamma \mathbb{E}_{(s'|s, a) \sim p} \max_{\pi} \left[-\frac{1}{\beta} \log \frac{\pi(a'|s')}{\pi_0(a'|s')} + Q(s', a') \right]$$

- Maximum achieved for **soft-optimal** policy, soft-optimal value recursion

- With **tabular** parametrization: $Q(s, a) \rightarrow r + \frac{\gamma}{\beta} \log \mathbb{E}_{(a'|s') \sim \pi_0} [\exp \beta Q(s', a')]$

- With **differentiable** parametrization:

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

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

[Fox et al., Taming the noise in RL via soft updates, UAI 2016]

[Haarnoja et al., RL with deep energy-based policies, ICML 2017]

Soft Actor–Critic (SAC)

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

- In **continuous action spaces**, we can't explicitly softmax $Q(s, a)$ over a
- We can train a **critic** off-policy to **soft-evaluate** the actor

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

- And a soft-greedy **actor** to **imitate** the critic

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

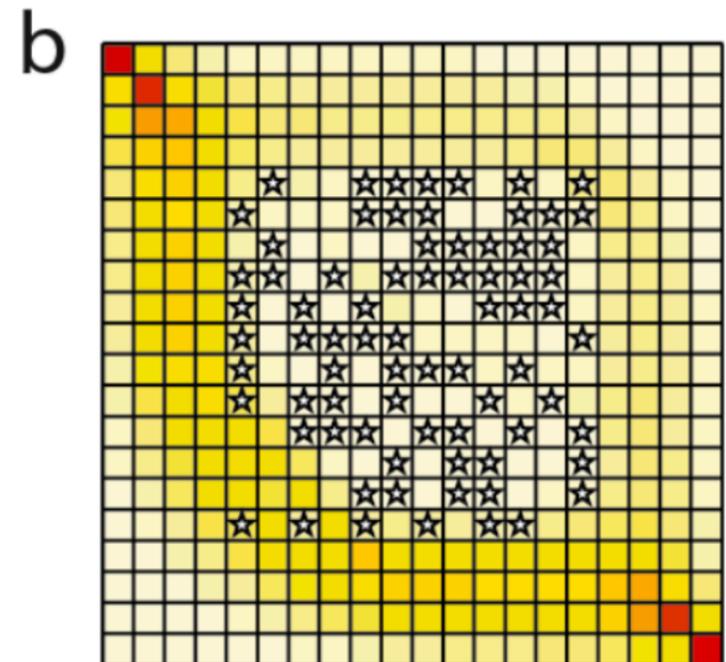
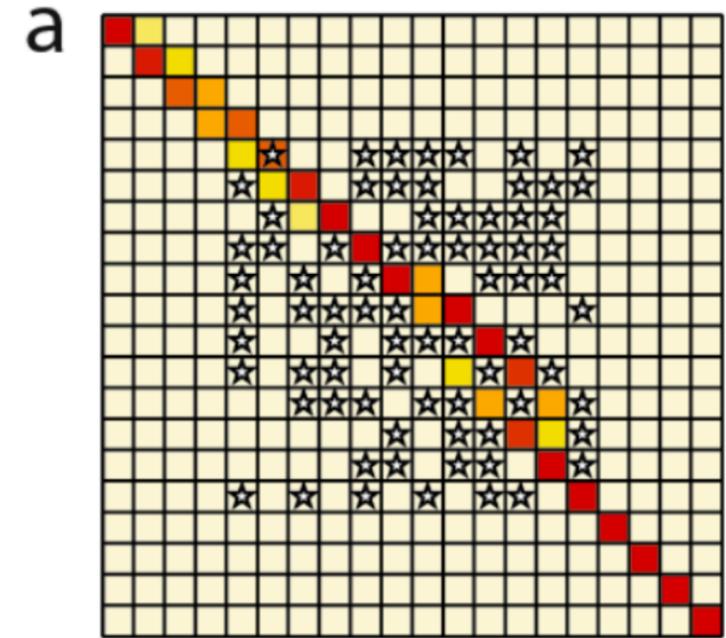
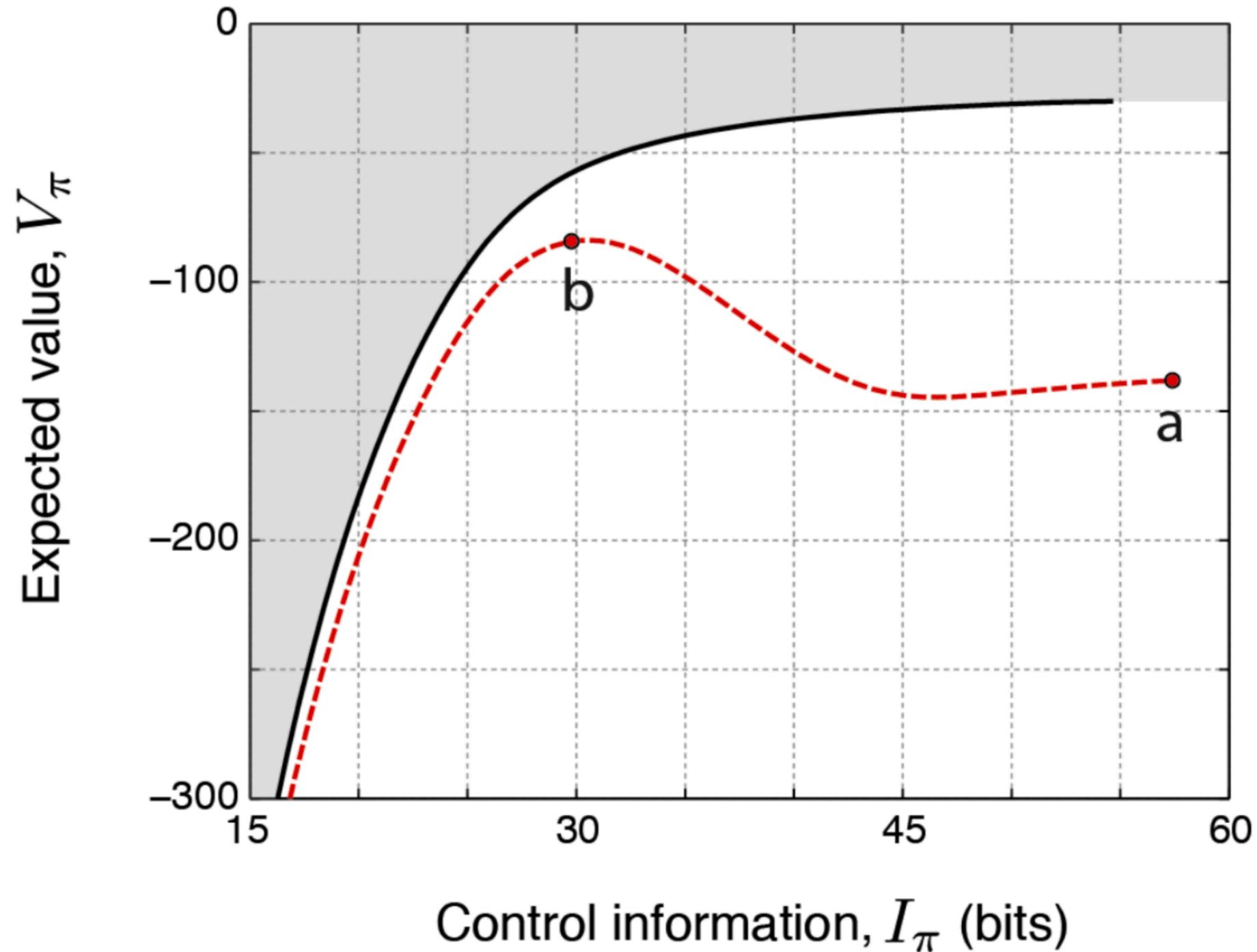
[Haarnoja et al., Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, ICML 20178]

- Can optimize $\beta = \frac{1}{\tau}$ to match a **target entropy** $L_\tau(s, a) = -\tau \log \pi_\theta(a | s) - \tau H$

Why use a finite β

- To model **suboptimal** agents / teachers
- To be robust to model **misspecification** / avoid **overfitting**
- With uncertainty in Q , to eliminate **bias** due to winner's curse
 - For $\beta \rightarrow \infty$: **positive bias** $\mathbb{E}[\max_a Q(a)] \geq \max_a \mathbb{E}[Q(a)]$
 - For $\beta \rightarrow 0$: **negative bias** $\mathbb{E}[\mathbb{E}_{a \sim \pi_0}[Q(a)]] = \mathbb{E}_{a \sim \pi_0}[\mathbb{E}[Q(a)]] \leq \max_a \mathbb{E}[Q(a)]$
 - Somewhere in between there must be an **unbiased** β
- To be robust to **non-stationary** environment, **multi-agent**, etc.

Robustness to model uncertainty



Recap

- We can model bounded rationality with **KL cost** to diverge from prior π_0
- Equivalent to a form of **variational inference**
- Can be optimized with **Soft Q-Learning (SQL)**
 - In continuous action spaces, **Soft Actor–Critic (SAC)**
- Value–entropy **trade-off coefficient β** shouldn't be annealed too fast
 - **Schedule** with a target entropy or by other principles