Sequential Decision-Making with Expert Demonstrations under Unobserved Heterogeneity

Vahid Balazadeh, Keertana Chidambaram, Viet Nguyen, Rahul G. Krishnan*, Vasilis Syrgkanis*

UNIVERSITY OF TORONTO
VECTOR INSTITUTE
CIFAR

Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Image credit to ChatGPT

Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Learn from online interaction with **no prior knowledge.**

Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Imitate a human (expert) teacher with **no online interaction.**

Student 1 → Teaching policy 1

Student 2 → Teaching policy 2

⋮

Student N → Teaching policy N

Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Imitate a human (expert) teacher with **no online interaction.**

Name: Alex
Previous Grades: B
Demographics: A
Family History: C
…

Use verbal presentations

Name: Sophia
Previous Grades: A
Demographics: B
Family History: C
…

Use visual aids

⋮

?

Name: Rosie
Previous Grades: A
Demographics: B
Family History: C
…

Provide self-study time

Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Imitate a human (expert) teacher with **no online interaction.**

Name: Alex
Previous Grades: B
Demographics: A
Family History: C
Auditory Learner

Use verbal presentations

Name: Sophia
Previous Grades: A
Demographics: B
Family History: C
Visual Learner

Use visual aids

⋮

Name: Rosie
Previous Grades: A
Demographics: B
Family History: C
Read/Write Learner

Provide self-study time

Experts can **observe and know more** than the agent.
Pure imitation leads to **sub-optimal** policies in such scenarios.

Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Combine the **prior Knowledge + online interaction**

Name: Loren
Previous Grades: A
Demographics: B
Family History: C

Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Combine the **prior Knowledge + online interaction**

Name: Loren
Previous Grades: A
Demographics: B
Family History: C

Use Prior Knowledge

Read/Write Learner: 50%
Visual Learner: 50%

Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Combine the **prior Knowledge + online interaction**

Name: Loren
Previous Grades: A
Demographics: B
Family History: C

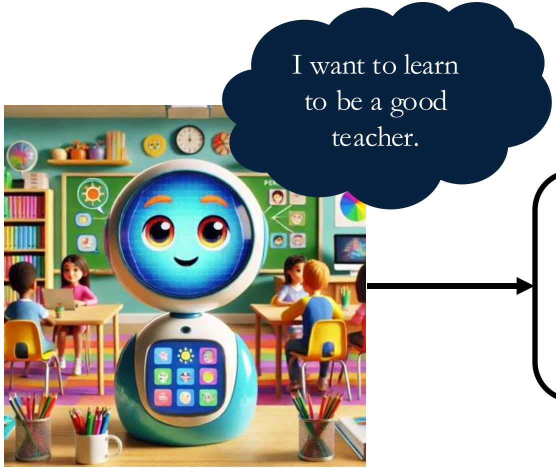
Use Prior Knowledge

Read/Write Learner: 50%
Visual Learner: 50%

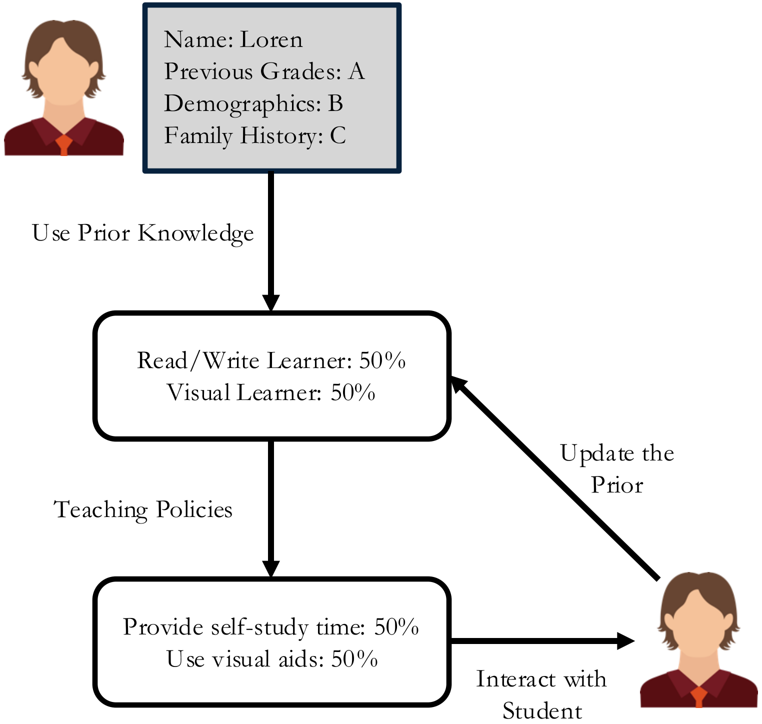
Teaching Policies

Provide self-study time: 50%
Use visual aids: 50%

Motivating Example: A Personalized AI Teacher



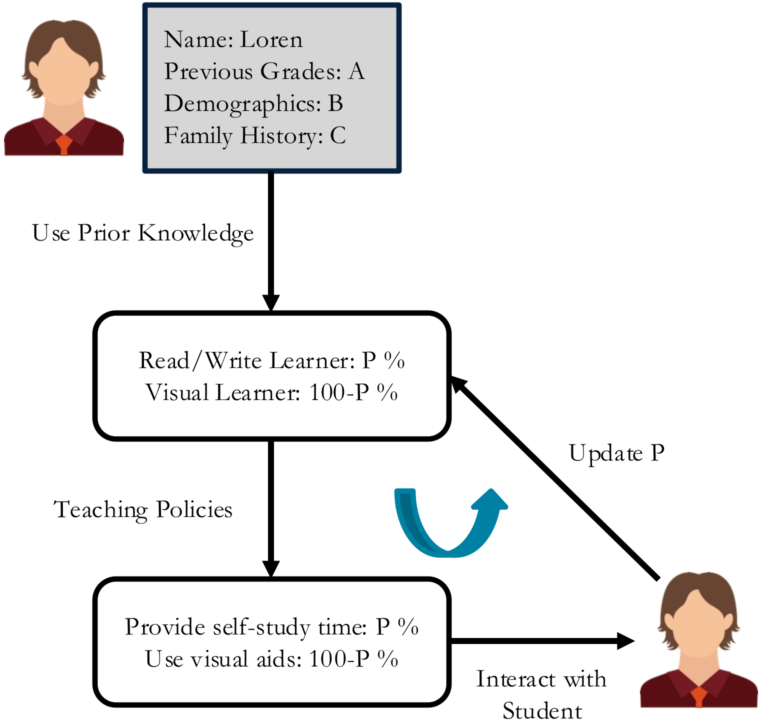
Combine the **prior Knowledge + online interaction**



Motivating Example: A Personalized AI Teacher

I want to learn to be a good teacher.

Combine the **prior Knowledge + online interaction**



Prior data **limits the space of exploration**.
Online interaction identifies **unobserved factors.**

Formalizing the Problem

Common Decision-Making Setting

Markov Decision Process (MDP):	$\mathcal{M} = (S, A, T, R, H, \rho)$
State	S
Actions	A
Transition Function	$T\colon S \times A \longrightarrow \Delta S$
Reward Function	$R\colon S \times A \longrightarrow \Delta\mathbb{R}$
Horizon	H
Episodes	L

Our Decision-Making Setting

Markov Decision Process (MDP):	$\mathcal{M} = (S, A, T, R, H, \rho, \boldsymbol{\mu}^*)$
State	S
Actions	A
Transition Function	$T\colon S \times A \times \boldsymbol{C} \longrightarrow \Delta S$
Reward Function	$R\colon S \times A \times \boldsymbol{C} \longrightarrow \Delta \mathbb{R}$
Horizon	H
Episodes	L
Initial State Distribution	$\rho \in \Delta S$
Distribution of Unobserved Factors (fixed distribution over learning styles)	$\boldsymbol{c} \sim \boldsymbol{\mu}^*$

Goal

Minimize the Bayesian Regret:

$$Reg := \mathbb{E}_{c\sim\mu^*}\left[\sum_{t=1}^{L} V_c(\pi_c) - \mathbb{E}_{\pi^t\sim p^t}[V_c(\pi^t)]\right]$$

Value Function: $$V_c(\pi) = \mathbb{E}\left[\sum_{h=1}^{H} r_h | \pi, c\right]$$

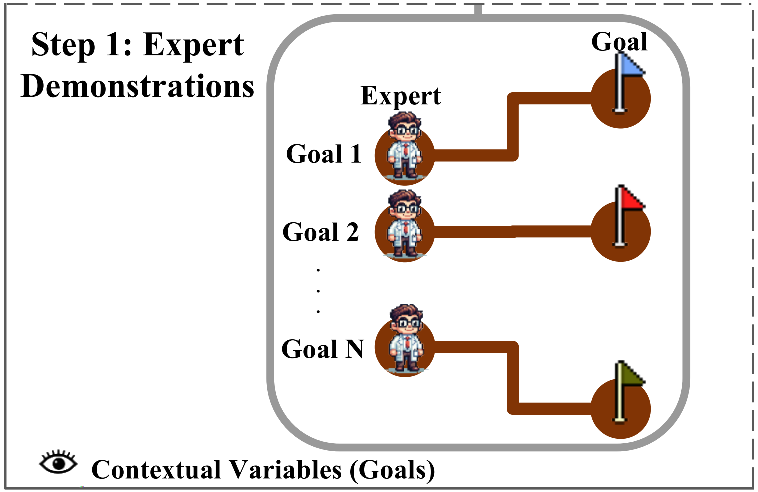
Optimal Policy: $$\pi_c = argmax_{\pi\in\Pi} V_c(\pi)$$

History-Dependent Policies: $$p^1, \dots., p^L \epsilon \Delta(\Pi)$$

Methodology: Experts-as-Priors (ExPerior)

ExPerior (Step 1): Experts Generate Demonstrations

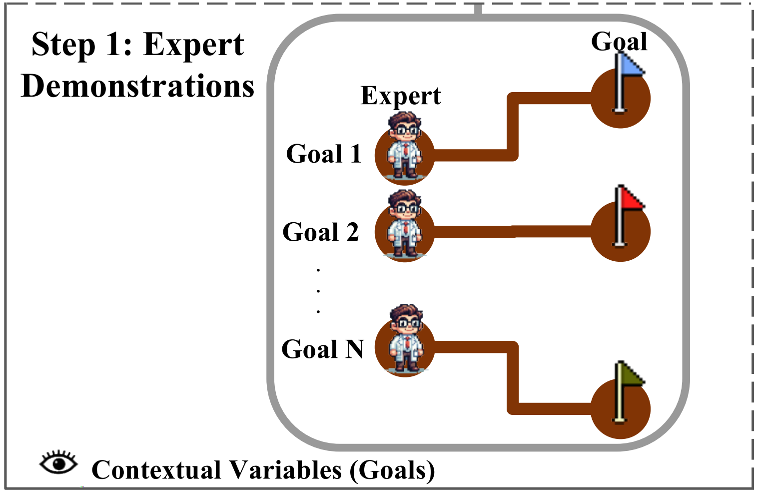
We assume experts can **observe** the “unobserved” factors



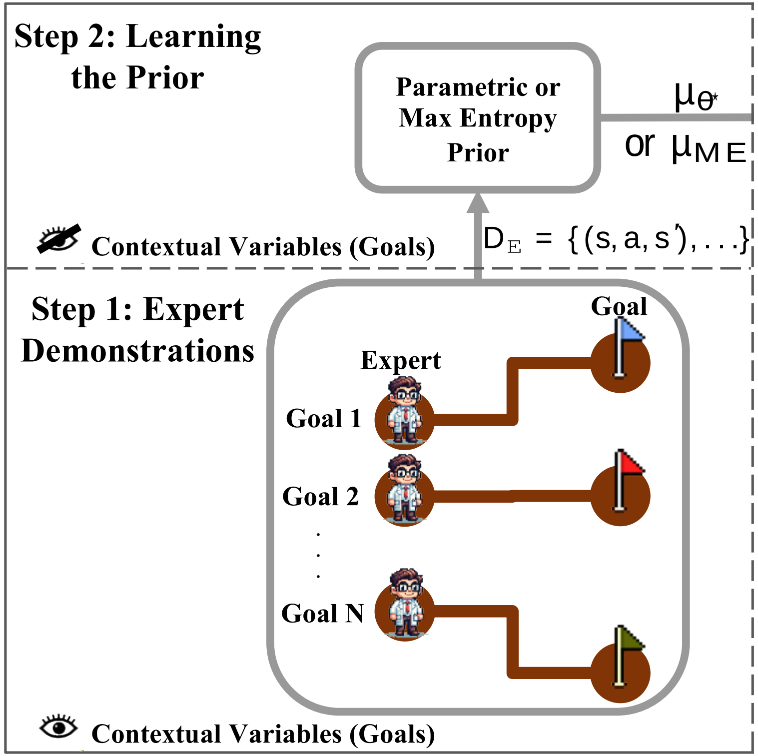
ExPerior (Step 1): Experts Generate Demonstrations

We assume experts can **observe** the "unobserved" factors and are **near-optimal (noisily-rational).**

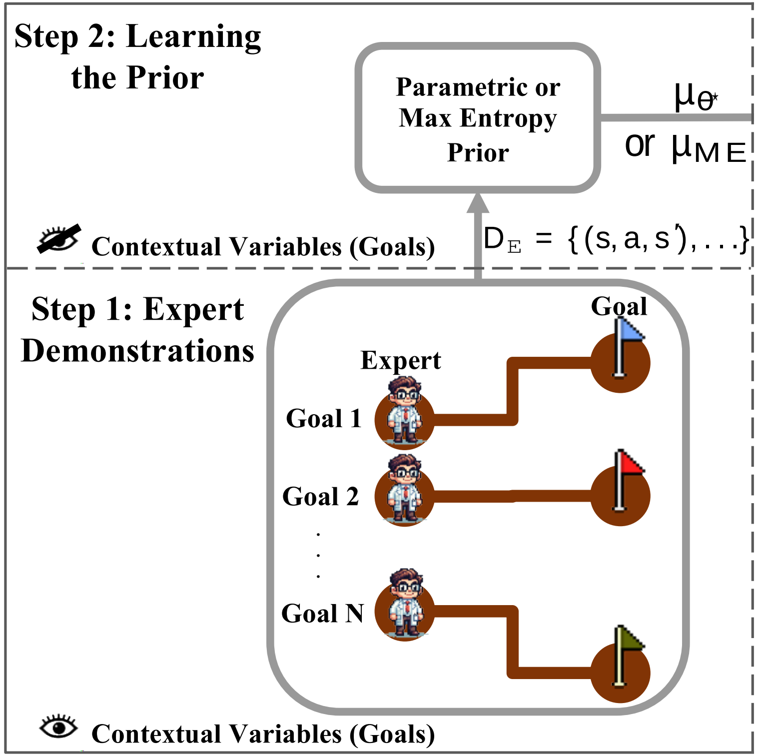
$$\forall s \in S, a \in A, c \in C : p_E(a \mid s;\ c) \propto \exp\{\beta . Q_c^{\pi_c}(s,a)\} \quad \text{where} \quad Q_c^{\pi}(s,a) \coloneqq \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'} \middle| s_h = s, a_h = a, \pi, c\right]$$



ExPerior (Step 2): Infer a Prior Distribution over Unobserved Contexts



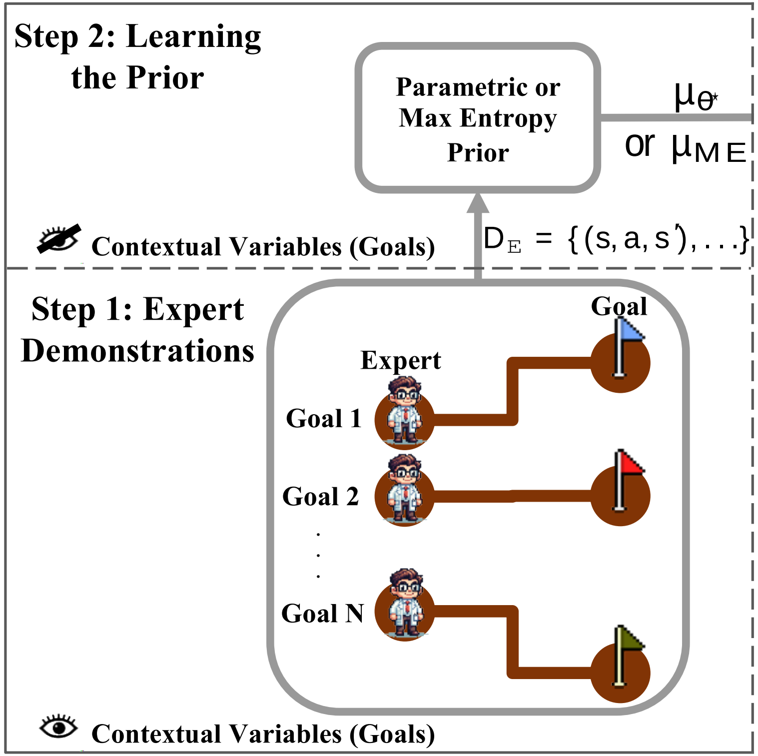
ExPerior (Step 2): Infer a Prior Distribution for Unobserved Factors



Trajectory

$$\tau_E = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H, s_{H+1})$$

ExPerior (Step 2): Infer a Prior Distribution for Unobserved Factors



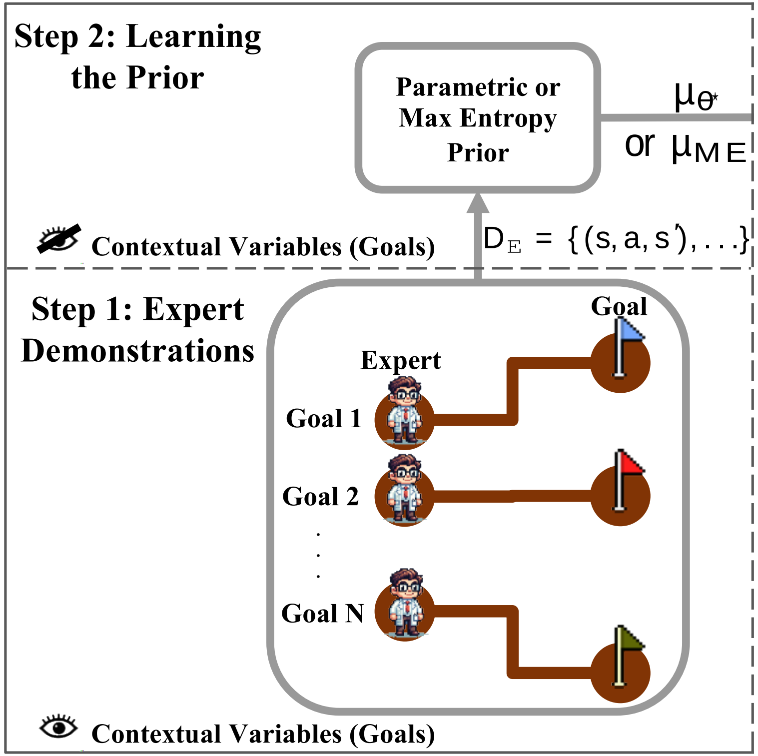
Trajectory

$$\tau_E = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H, s_{H+1})$$

Marginal likelihood of from expert dataset

$$P_E(\tau_E; \mu) = \mathbb{E}_{c\sim\mu^*}\left[\rho(s_1)\ \prod_{h=1}^{H} p_E(a_h | s_h\ ; c)\ T(s_{h+1} | s_h, a_h; c)\right]$$

ExPerior (Step 2): Infer a Prior Distribution for Unobserved Factors



Trajectory

$$\tau_E = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H, s_{H+1})$$

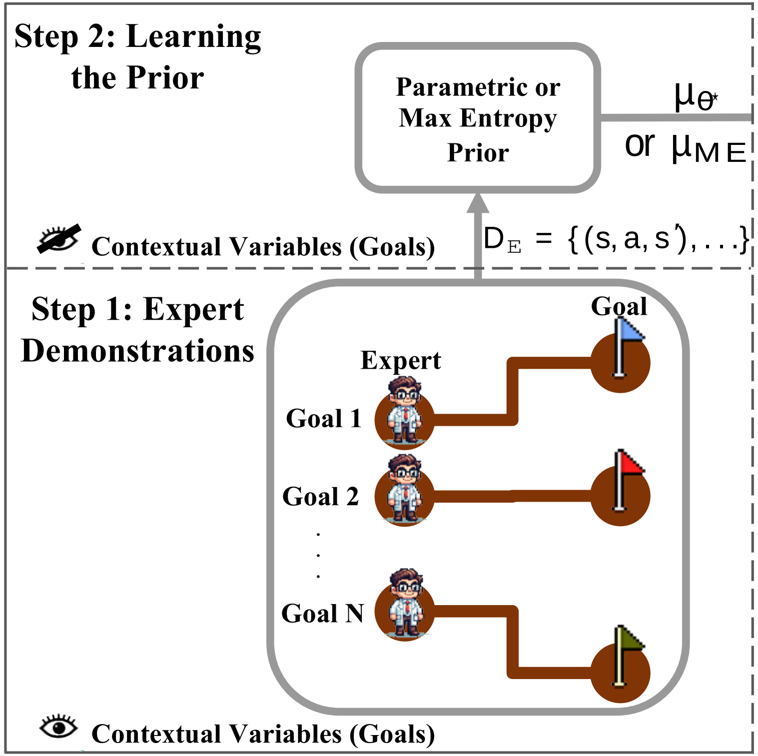
Marginal likelihood of from expert dataset

$$P_E(\tau_E; \mu) = \mathbb{E}_{c\sim\mu^*}\left[\rho(s_1)\ \prod_{h=1}^{H} p_E(a_h|s_h\,; c)\ T(s_{h+1}|s_h, a_h; c)\right]$$

Parametric Prior: **Maximum marginal likelihood**

$$\mu_{\theta^*} \in argmin \sum_{\tau\,\in\,expert\ data} -\log P_E(\tau; \mu_\theta)$$

ExPerior (Step 2): Infer a Prior Distribution for Unobserved Factors



Trajectory

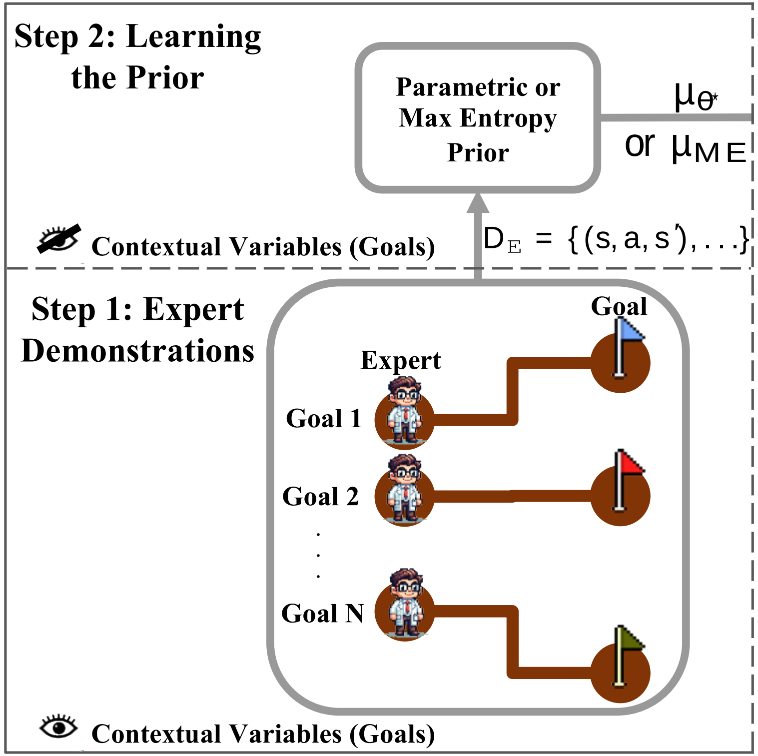
$$\tau_E = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H, s_{H+1})$$

Marginal likelihood of from expert dataset

$$P_E(\tau_E; \mu) = \mathbb{E}_{c\sim\mu^*}\left[\rho(s_1)\ \prod_{h=1}^{H} p_E(a_h|s_h\,; c)\ T(s_{h+1}|s_h, a_h; c)\right]$$

What if there is no existing knowledge of the parametric form of the prior?

ExPerior (Step 2): Infer a Prior Distribution for Unobserved Factors



Trajectory

$$\tau_E = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H, s_{H+1})$$

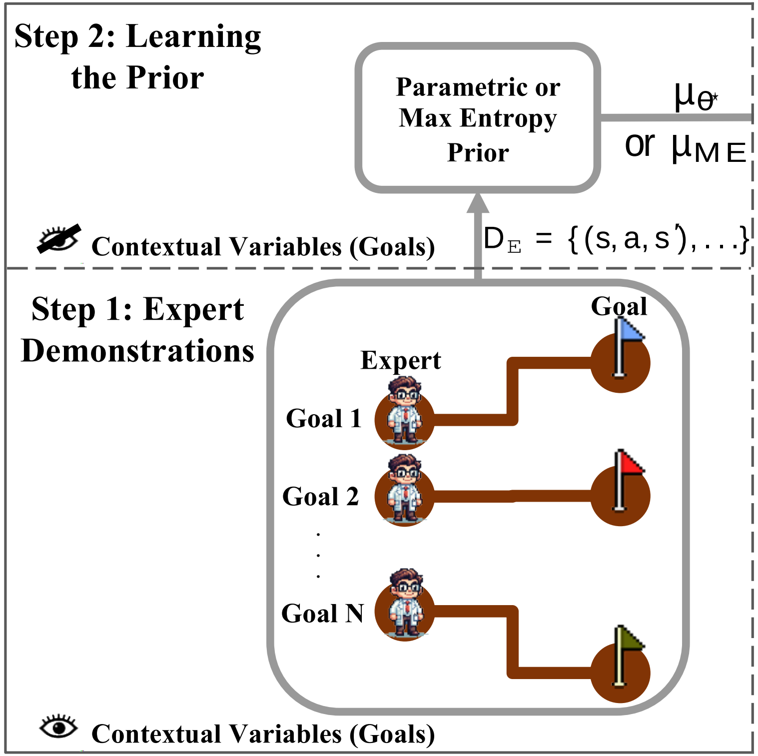
Marginal likelihood of from expert dataset

$$P_E(\tau_E; \mu) = \mathbb{E}_{c\sim\mu^*}\left[\rho(s_1)\ \prod_{h=1}^{H} p_E(a_h|s_h\,; c)\ T(s_{h+1}|s_h, a_h; c)\right]$$

What if there is no existing knowledge of the parametric form of the prior?

Choose the Maximum-Entropy Prior.

ExPerior (Step 2): Infer a Prior Distribution for Unobserved Factors

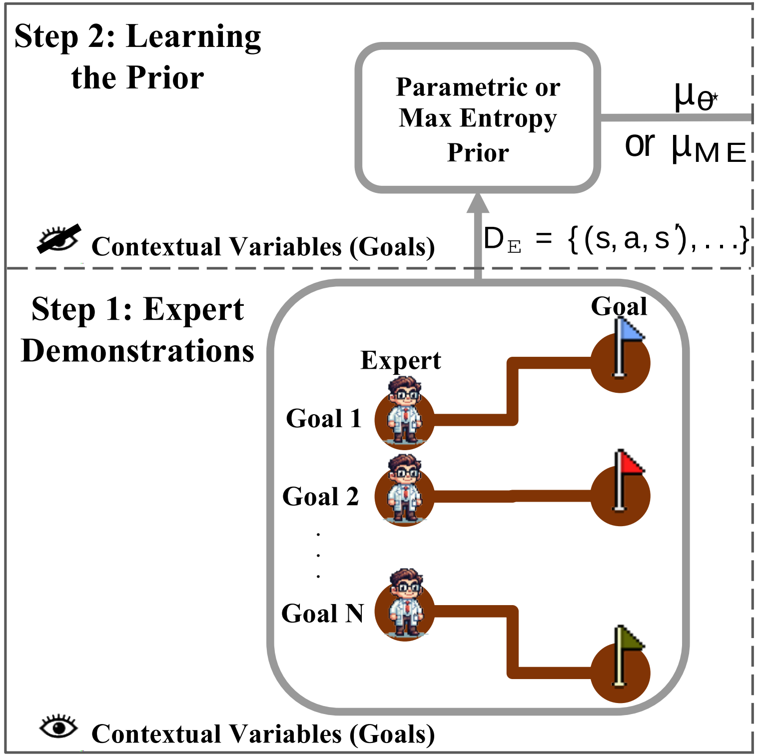


The high-probability set of all plausible prior distribution

$$\mathcal{P}(\varepsilon) := \left\{\mu;\ D_{KL}\left(\hat{P}_E \middle\| P_E(\tau_E;\mu)\right) \leq \varepsilon\right\}$$

Choose the Maximum-Entropy Prior.

ExPerior (Step 2): Infer a Prior Distribution for Unobserved Factors



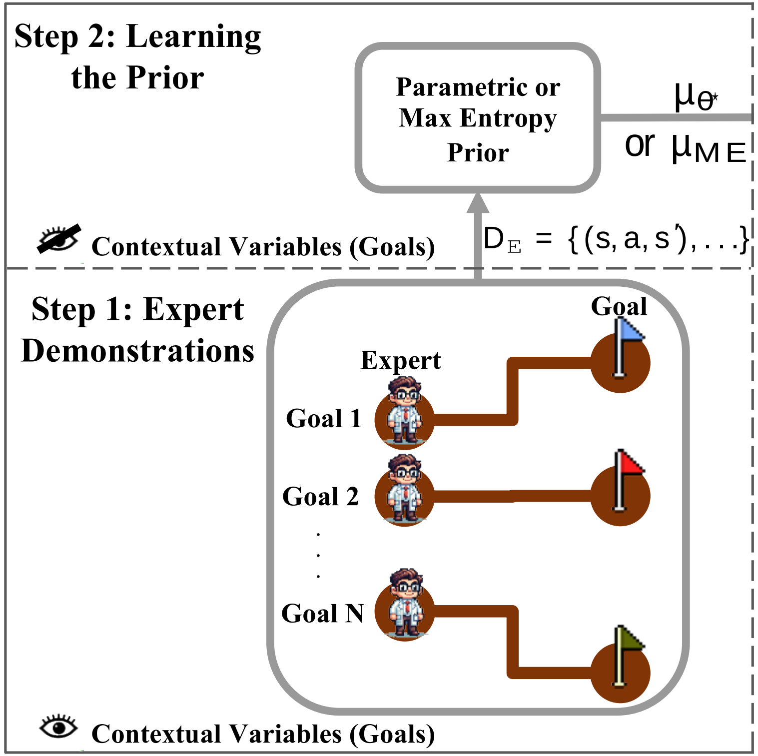
The high-probability set of all plausible prior distribution

$$\mathcal{P}(\varepsilon) := \left\{\mu;\ D_{KL}\left(\hat{P}_E \middle\| P_E(\tau_E;\mu)\right) \leq \varepsilon\right\}$$

$$\boxed{\mu_{ME} = argmax_{\mu\in\mathcal{P}(\varepsilon)} H(\mu)}$$

Choose the Maximum-Entropy Prior.

ExPerior (Step 2): Infer a Prior Distribution for Unobserved Factors



The high-probability set of all plausible prior distribution

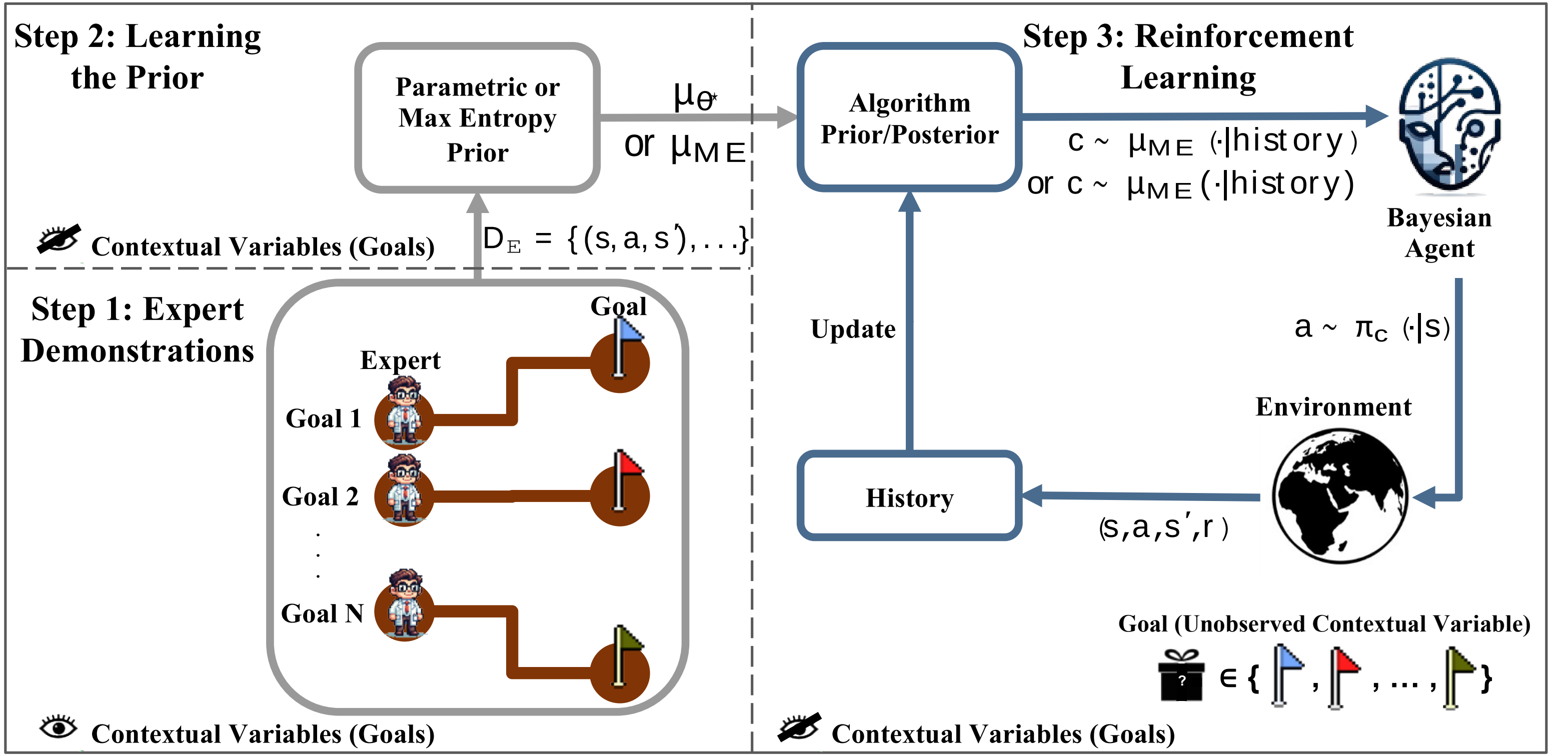
$$\mathcal{P}(\varepsilon) := \left\{\mu;\ D_{KL}\left(\hat{P}_E \middle\| P_E(\tau_E;\mu)\right) \leq \varepsilon\right\}$$

$$\boxed{\mu_{ME} = argmax_{\mu\in\mathcal{P}(\varepsilon)} H(\mu)}$$

$\boldsymbol{\mathcal{P}(\varepsilon)}$ **is convex! Solve using Fenchel's duality theorem.**

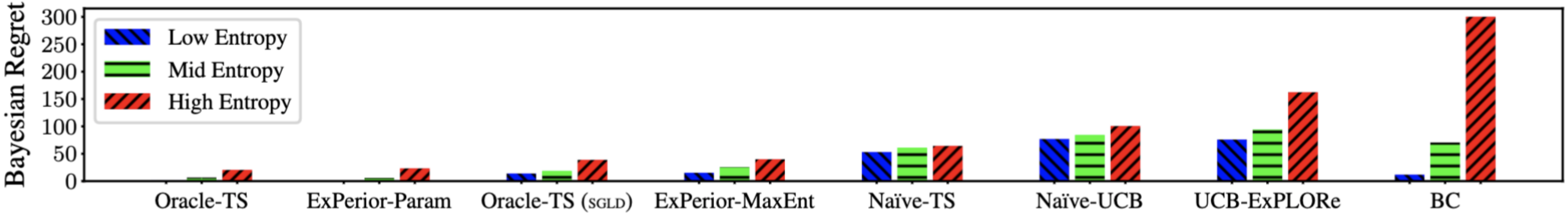
Choose the Maximum-Entropy Prior.

ExPerior (Step 3): Exploration with Posterior Sampling



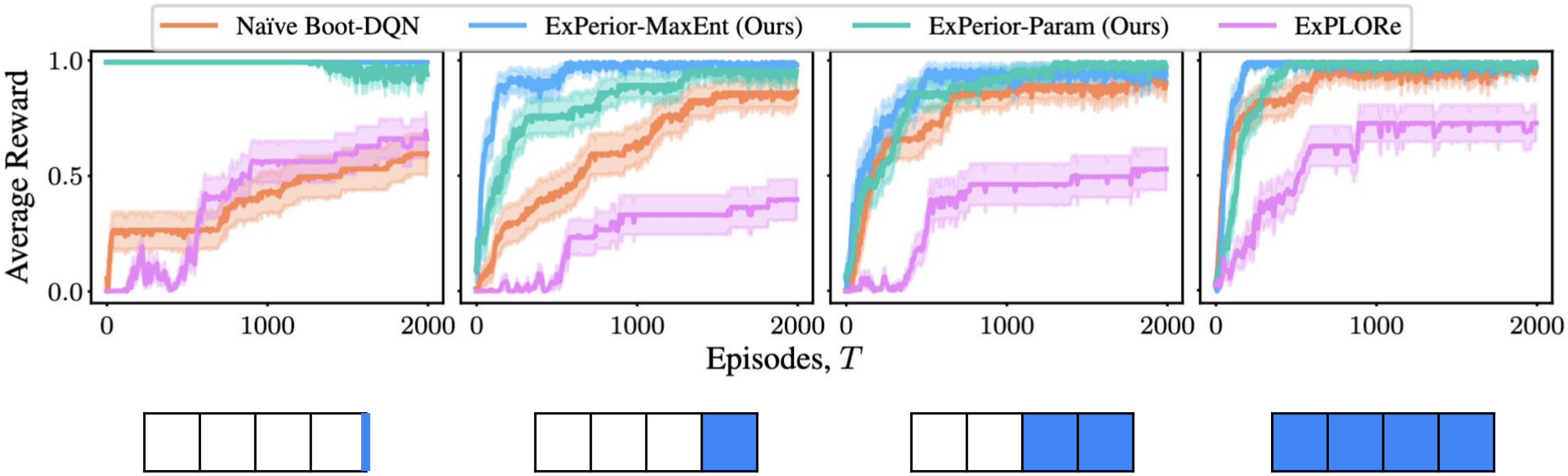
Experiments and Implications

Bandit Experiments – Bernoulli Multi-Armed Bandit



Additional Experiments – MDPs and Partially Observable MDPs

- Deep Sea Environment (MDP)



Additional Experiments – MDPs and Partially Observable MDPs

- Frozen Lake Environment (POMDP)

	Fixed # Hazard = 9				**Fixed $\beta = 1$**			
	$\beta = 0.1$	$\beta = 1$	$\beta = 2.5$	$\beta = 10$	# Hazard = 2	# Hazard = 5	# Hazard = 7	# Hazard = 9
(POMDP)								
ExPerior-MaxEnt	-22.58 ± 1.17	**6.00 ± 0.00**	3.58 ± 0.89	1.62 ± 1.85	11.47 ± 0.52	**5.71 ± 0.67**	**6.00 ± 0.00**	**6.00 ± 0.00**
ExPerior-Param	-23.32 ± 0.69	-4.31 ± 1.80	5.27 ± 0.51	**6.00 ± 0.00**	**12.00 ± 0.37**	2.11 ± 1.41	5.42 ± 0.40	-4.31 ± 1.80
Naïve Boot-DQN	-23.32 ± 0.69	-23.32 ± 0.69	-23.32 ± 0.69	-23.32 ± 0.69	-14.36 ± 5.88	-20.57 ± 2.91	-20.39 ± 1.75	-23.32 ± 0.69
ExPLORe	**5.99 ± 0.00**	**6.00 ± 0.00**	**6.00 ± 0.00**	**6.00 ± 0.00**	-30.68 ± 12.40	-10.64 ± 16.64	-13.00 ± 19.00	**6.00 ± 0.00**
Optimal	6.00 ± 0.00	6.00 ± 0.00	6.00 ± 0.00	6.00 ± 0.00	12.00 ± 0.37	6.53 ± 0.31	6.00 ± 0.00	6.00 ± 0.00
(MDP)								
ExPerior-MaxEnt	-23.36 ± 1.26	12.26 ± 0.29	12.68 ± 0.03	**12.71 ± 0.03**	**13.02 ± 0.18**	**12.78 ± 0.11**	**12.78 ± 0.06**	12.26 ± 0.29
ExPerior-Param	-25.53 ± 2.35	**12.64 ± 0.08**	**12.70 ± 0.03**	12.68 ± 0.03	13.00 ± 0.18	**12.78 ± 0.12**	12.73 ± 0.07	**12.64 ± 0.08**
Naïve Boot-DQN	-23.32 ± 0.69	-23.32 ± 0.69	-23.32 ± 0.69	-23.32 ± 0.69	-14.39 ± 5.22	-20.99 ± 2.86	-20.39 ± 1.75	-23.32 ± 0.69
ExPLORe	**11.74 ± 0.41**	11.75 ± 0.63	11.96 ± 0.28	12.3 ± 0.22	-113.84 ± 17.50	-54.89 ± 13.75	-10.00 ± 7.60	11.75 ± 0.63
Optimal	12.71 ± 0.03	12.71 ± 0.03	12.71 ± 0.03	12.71 ± 0.03	13.02 ± 0.18	12.78 ± 0.11	12.76 ± 0.06	12.64 ± 0.03

Conclusion and Implication

- ExPerior provides a principled approach to combining offline prior data with online learning under unobserved heterogeneity in general decision-making settings.

Conclusion and Implication

- ExPerior provides a principled approach to combining offline prior data with online learning under unobserved heterogeneity in general decision-making settings.

data $\longrightarrow$ (max-entropy) prior distribution $\longrightarrow$ posterior sampling

Conclusion and Implication

- ExPerior provides a principled approach to combining offline prior data with online learning under unobserved heterogeneity in general decision-making settings.

data $\longrightarrow$ (max-entropy) prior distribution $\longrightarrow$ posterior sampling

- Our work opens new directions for more complex open-ended decision-making tasks, such as personalized adaptation of large language models.

Thank You!