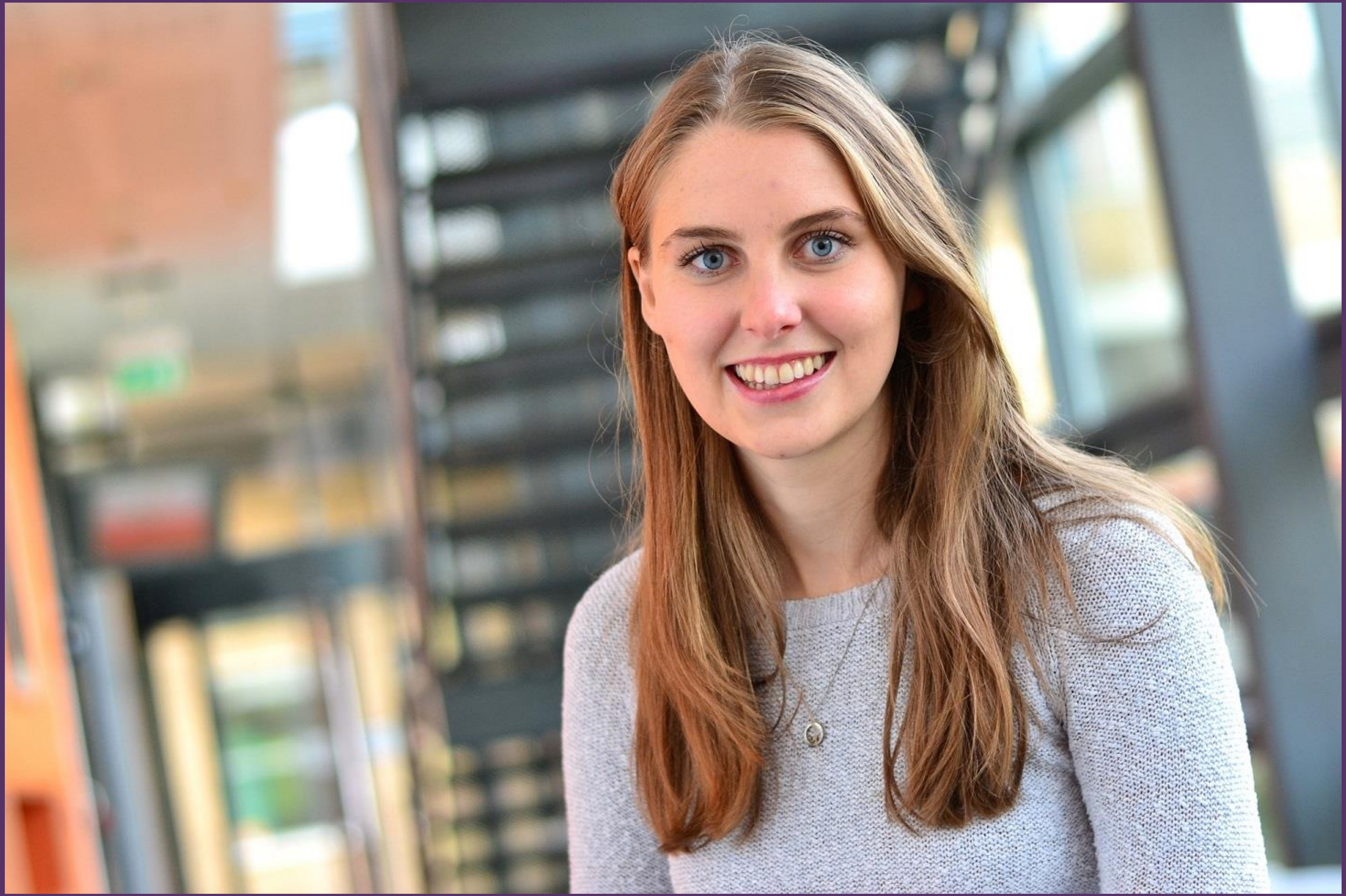
A UNIFYING VIEW OF OPTIMISM IN EPISODIC REINFORCEMENT LEARNING

Gergely Neu
(Universitat Pompeu Fabra, Barcelona)

based on joint work with
Ciara Pike-Burke



← your next invited speaker ;)

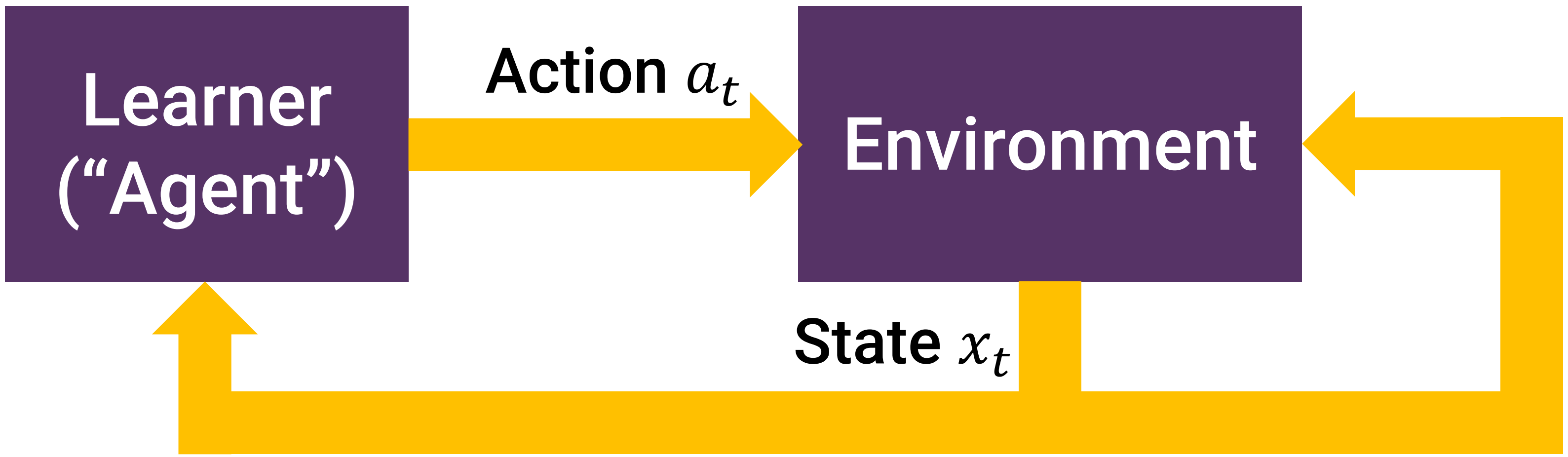
based on joint work with

Ciara Pike-Burke (UPF → Imperial College)

THIS TALK

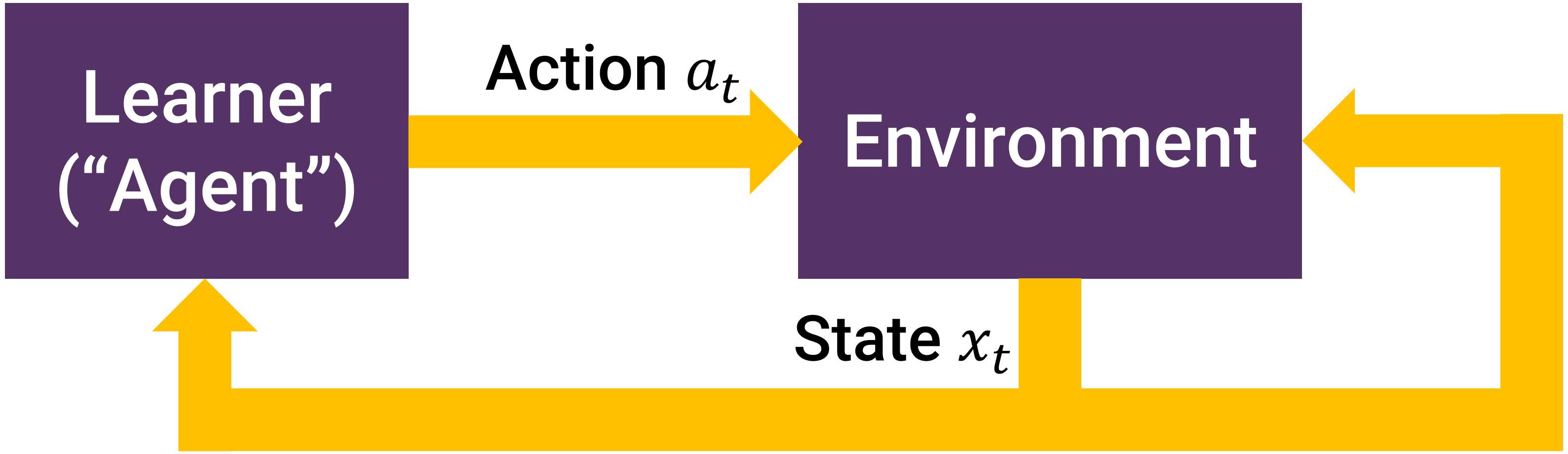
- The quickest intro to MDPs you've ever heard
- Optimistic exploration in RL
  - Model-optimism and value-optimism
  - A unifying view
- Linear function approximation
  - Local and global optimism

MARKOV DECISION PROCESSES



- Learner:
  - Observe state x_t, choose action a_t
  - Obtain reward $r(x_t, a_t)$
- Environment: Draw next state $x_{t+1} \sim P(\cdot\, | x_t, a_t)$
- Episode ends in round H

MARKOV DECISION PROCESSES



- Learner:
  - Observe state x_t, choose action a_t
  - Obtain reward $r(x_t, a_t)$
- Environment: Draw next state $x_{t+1} \sim P(\cdot\,|x_t, a_t)$
- Episode ends in round H

Goal:
get as much reward as possible!

OPTIMALITY IN MDPS

Primal: optimality in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $\sum_a q_{h+1,a} = \sum_a P_a^\top q_{h,a}$

$\sum_a q_1(x_0, a) = 1,\ q \geq 0$

Dual: optimality in value-function space

as characterized by the Bellman optimality equations

$$V_h^* = \max_a \{ r_a + P_a V_{h+1}^* \}$$

OPTIMALITY IN MDPS

Primal: optimality in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $\sum_a q_{h+1,a} = \sum_a P_a^\top q_{h,a}$

$\sum_a q_1(x_0, a) = 1, \; q \geq 0$

↕ Equivalent due to Linear Programming duality

Dual: optimality in value-function space

as characterized by the Bellman optimality equations

$$V_h^* = \max_a \{ r_a + P_a V_{h+1}^* \}$$

OPTIMALITY IN MDPS

Primal: optimality in trajectory space

$$\begin{aligned} \text{maximize} \quad & \sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle \\ \text{subject to} \quad & \sum_a q_{h+1,a} = \sum_a P_a^\top q_{h,a} \\ & \sum_a q_1(x_0, a) = 1,\ q \geq 0 \end{aligned}$$

Optimal policy:
$\pi_h^*(a|x) \propto q_h^*(x,a)$

Equivalent due to Linear Programming duality

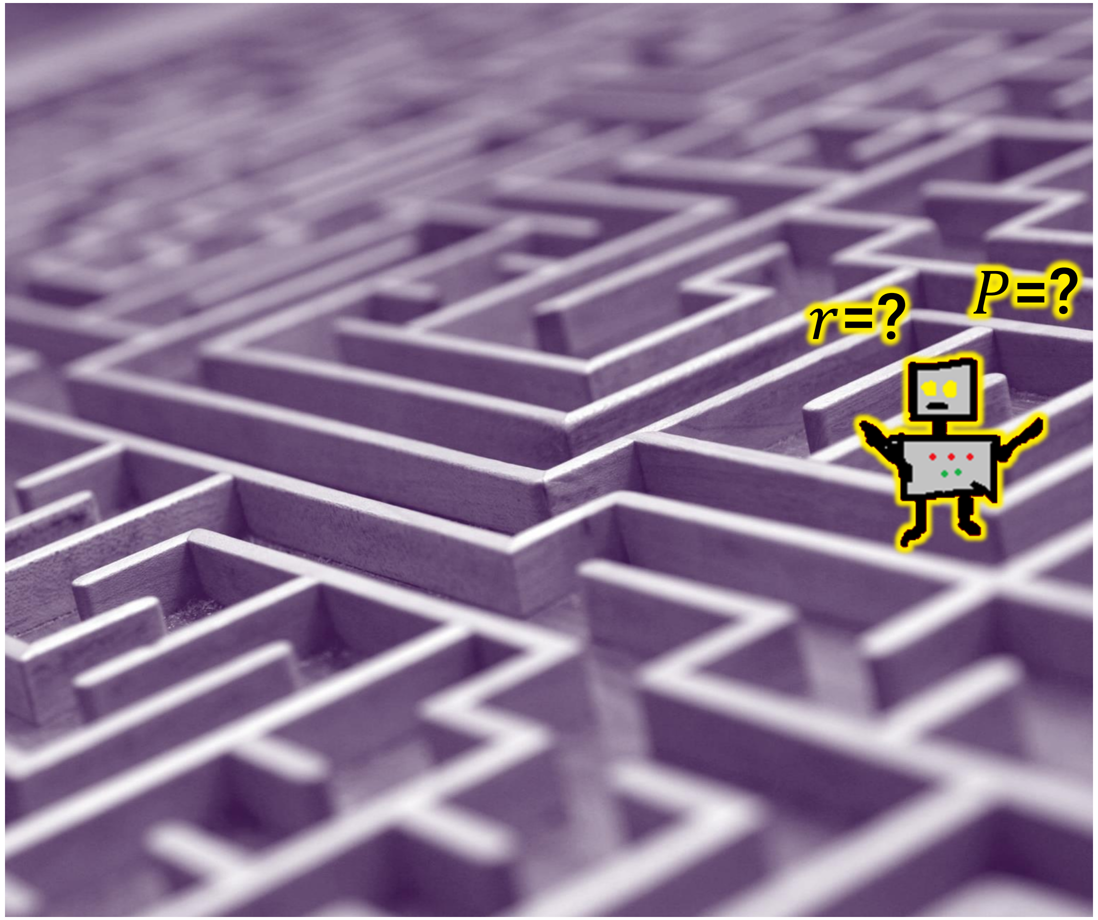
Dual: optimality in value-function space

as characterized by the Bellman optimality equations

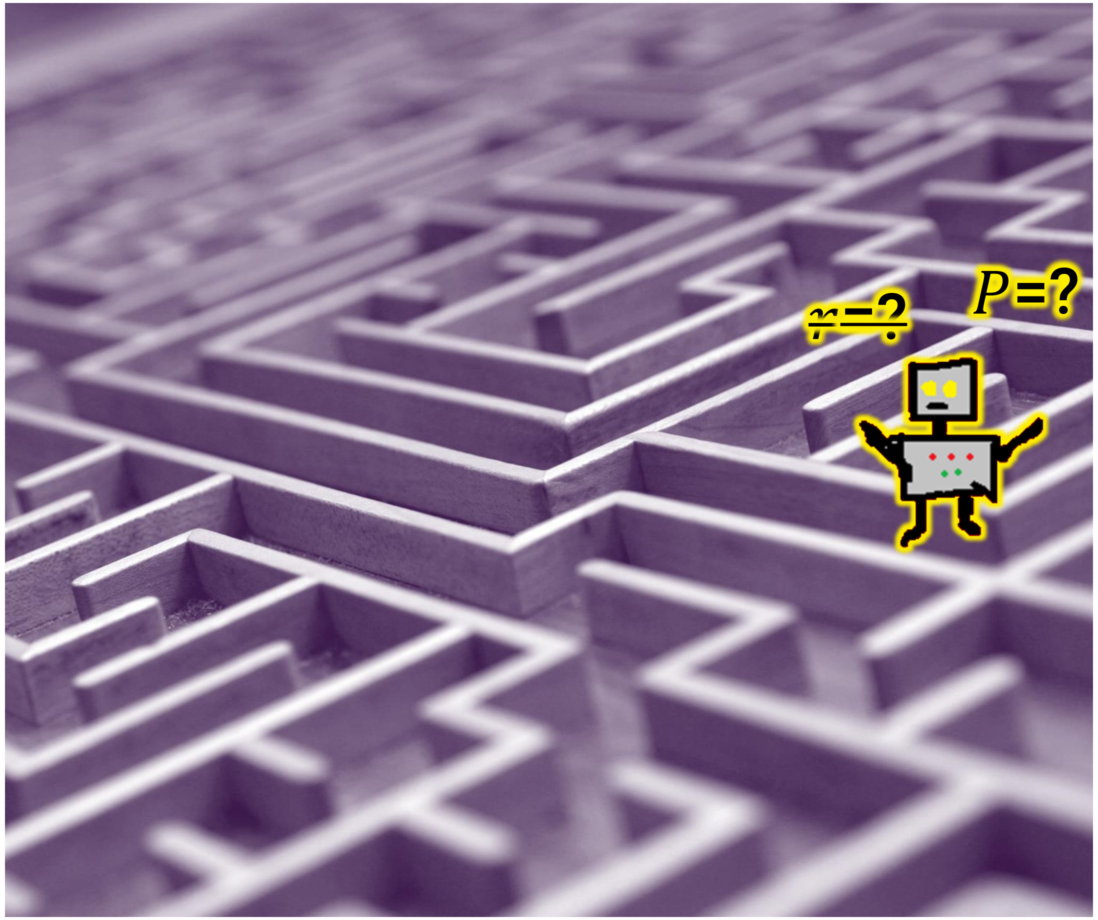
$$V_h^* = \max_a \{ r_a + P_a V_{h+1}^* \}$$

Optimal policy:
$\pi_h^*(a|x) \propto \mathbb{I}_{\{a = \operatorname{argmax}_{a'} Q^*(x,a')\}}$

OPTIMISTIC EXPLORATION IN RL



OPTIMISTIC EXPLORATION IN RL

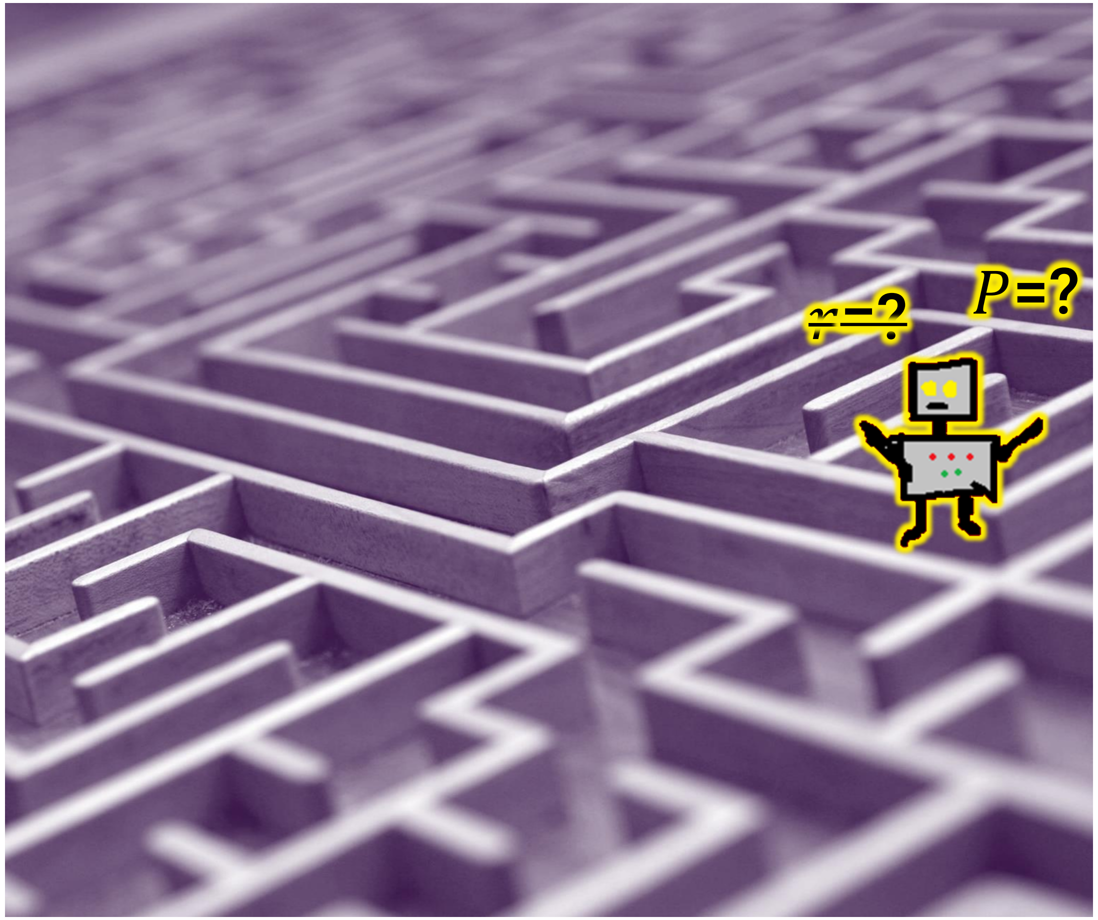


OPTIMISTIC EXPLORATION IN RL

"Optimism in the face of uncertainty"

≈

imagine you're in the
best statistically plausible world
and plan accordingly

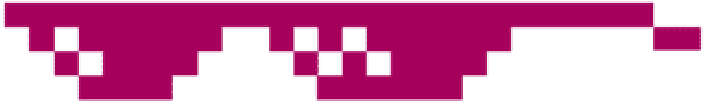


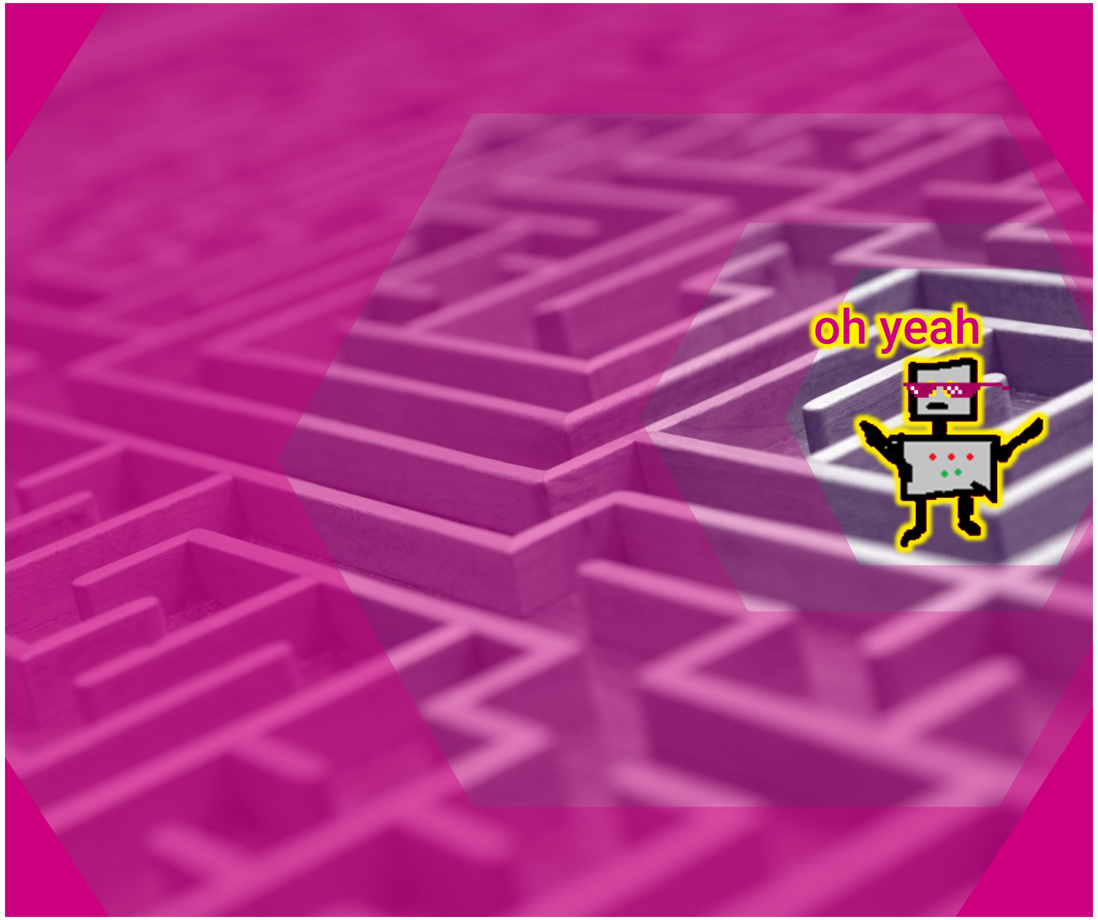
OPTIMISTIC EXPLORATION IN RL

"Optimism in the face of uncertainty"

$\approx$

imagine you're in the best statistically plausible world and plan accordingly





THE TWO KINDS OF OPTIMISM

Optimism in model space:
construct a confidence set around P and jointly optimize over models & policies

Optimism in value space:
construct upper confidence bounds directly on the optimal value function V^*

THE TWO KINDS OF OPTIMISM

Optimism in model space: construct a confidence set around P and jointly optimize over models & policies

- $\mathcal{P}$ = confidence set of transition functions $\tilde{P}$ centered around empirical transition function $\hat{P}$ such that
  $$D\left(\tilde{P}(\cdot\,|x,a), \hat{P}(\cdot\,|x,a)\right) \leq \epsilon(x,a),$$
  holds for all (x,a)
- Calculate optimistic policy-model pair
  $$(\pi^+, P^+) = \arg\max_{\pi, \tilde{P}\in\mathcal{P}} V_{\tilde{P}}^{\pi}(x_0)$$
- E.g., UCRL2 (Jaksch et al., 2010) uses $\left\|\tilde{P}(\cdot\,|x,a) - \hat{P}(\cdot\,|x,a)\right\|_1 \leq C\sqrt{S/N(x,a)}$ and "extended value iteration"

Optimism in value space: construct upper confidence bounds directly on the optimal value function V^*

$N(x,a)$ =#visits to (x,a) so far

$\hat{P}(x'|x,a) = \frac{N(x,a,x')}{N(x,a)}$

THE TWO KINDS OF OPTIMISM

Optimism in model space:

construct a confidence set around P and jointly optimize over models & policies

- $\mathcal{P}$ = confidence set of transition functions $\tilde{P}$ centered around empirical transition function $\hat{P}$ such that
  $$D\left(\tilde{P}(\cdot\,|x,a), \hat{P}(\cdot\,|x,a)\right) \le \epsilon(x,a),$$
  holds for all (x,a)
- Calculate optimistic policy-model pair
  $$(\pi^+, P^+) = \arg\max_{\pi, \tilde{P} \in \mathcal{P}} V^{\pi}_{\tilde{P}}(x_0)$$
- E.g., UCRL2 (Jaksch et al., 2010) uses
  $$\left\|\tilde{P}(\cdot\,|x,a) - \hat{P}(\cdot\,|x,a)\right\|_1 \le C\sqrt{S/N(x,a)}$$
  and "extended value iteration"

Optimism in value space:

construct upper confidence bounds directly on the optimal value function V^*

- Compute exploration bonus $\mathrm{CB}(x,a)$ for each (x,a) and solve the optimistic Bellman optimality equations with the empirical transition function $\hat{P}$:
  $$V^+_{h+1} = \max_a\left\{r_a + \mathrm{CB}_a + \hat{P}_a V^+_h\right\}$$
- E.g., UCB-VI (Azar et al., 2017) uses
  $$\mathrm{CB}(x,a) = CH\sqrt{1/N(x,a)}$$

$N(x,a)$ =#visits to (x,a) so far

$\hat{P}(x'|x,a) = \frac{N(x,a,x')}{N(x,a)}$

PROS AND CONS

Optimism in model space:
construct a confidence set around P and jointly optimize over models & policies

Optimism in value space:
construct upper confidence bounds directly on the optimal value function V^*

☺ simple probabilistic analysis

just show that $P \in \mathcal{P}$!

☹ complicated to implement

need to search jointly over models and policies

☹ loose bounds

best known regret guarantees are suboptimal $O(HS\sqrt{AT})$

☹ complicated to analyze

need recursive arguments to show optimistic property of V^+

☺ easy to implement

dynamic programming with $\hat{P}$ and $r + CB$

☺ tight bounds

optimal regret bounds $O(H\sqrt{SAT})$

UNIFYING THE TWO VIEWS

Main result

"Every model-optimistic algorithm can be written as a value-optimistic algorithm"

UNIFYING THE TWO VIEWS

Main result

"Every model-optimistic algorithm can be written as a value-optimistic algorithm"

Consider any divergence D that is a) convex in its arguments and b) positive homogeneous, and define its conjugate D as

$$D_*(v|\hat{p},\epsilon) = \max_{p\in\Delta}\{\langle v, p - \hat{p}\rangle | D(p,\hat{p}) \leq \epsilon\}$$

UNIFYING THE TWO VIEWS

Main result

"Every model-optimistic algorithm can be written as a value-optimistic algorithm"

Consider any divergence D that is a) convex in its arguments and b) positive homogeneous, and define its conjugate D as

$$D_*(v|\hat{p},\epsilon) = \max_{p\in\Delta}\{\langle v, p - \hat{p}\rangle | D(p,\hat{p}) \leq \epsilon\}$$

Solution of
$$(\pi^+, P^+) = \arg\max_{\pi,\tilde{P}\in\mathcal{P}} V^{\pi}_{\tilde{P}}(x_0)$$

⟷

Solution of
$$V^+_{h+1} = \max_a\{r_a + \mathrm{CB}_{h,a} + \hat{P}_a V^+_h\}$$

$$\mathrm{CB}_h(x,a) = D_*\left(V^+_{h+1} \middle| \hat{P}_h(\cdot\,|x,a), \epsilon(x,a)\right)$$

EXAMPLES

Algorithm	Divergence	ϵ	Conjugate bound	Regret
UCRL2	$\lVert p - \hat{p} \rVert_1$	$\sqrt{S/N}$	$\epsilon \cdot \mathrm{span}(V)$	$SH^{3/2}\sqrt{AT}$
UCRL2B	$\max_x \frac{\left(p(x) - \hat{p}(x)\right)^2}{\hat{p}(x)}$	$1/N$	$\sum_x \sqrt{\epsilon \hat{p}(x)} \lvert V - \hat{p}V \rvert$	$H\sqrt{S\Gamma AT}$
KL-UCRL	$KL(p \vert \hat{p})$	S/N	$\sqrt{\epsilon \, \mathrm{Var}_{\hat{p}}(V)}$	$HS\sqrt{AT}$
χ^2-UCRL	$\sum_x \frac{\left(p(x) - \hat{p}(x)\right)^2}{\hat{p}(x)}$	S/N	$\sqrt{\epsilon \, \mathrm{Var}_{\hat{p}}(V)}$	$HS\sqrt{AT}$

Jaksch et al. (2010), Fruit et al. (2019), Filippi et al. (2010), Maillard et al. (2014)

EXAMPLES

Algorithm	Divergence	ϵ	Conjugate bound	Regret
UCRL2	$\|p - \hat{p}\|_1$	$\sqrt{S/N}$	$\epsilon \cdot \mathrm{span}(V)$	$SH^{3/2}\sqrt{AT}$
UCRL2B	$\max_x \frac{(p(x) - \hat{p}(x))^2}{\hat{p}(x)}$	$1/N$	$\sum_x \sqrt{\epsilon \hat{p}(x)}\|V - \hat{p}V\|$	$H\sqrt{S\Gamma AT}$
KL-UCRL	$KL(p\|\hat{p})$	S/N	$\sqrt{\epsilon\, \mathrm{Var}_{\hat{p}}(V)}$	$HS\sqrt{AT}$
χ^2-UCRL	$\sum_x \frac{(p(x) - \hat{p}(x))^2}{\hat{p}(x)}$	$S/$ [illegible]	$\sqrt{\epsilon\, \mathrm{Var}_{\hat{p}}(V)}$	$HS\sqrt{AT}$

“Data-dependent” exploration bonuses!

Jaksch et al. (2010), Fruit e [illegible] l et al. (2014)

PROOF IDEA: DUALITY

Primal: optimality in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $\sum_a q_{h+1,a} = \sum_a P_a^\top q_{h,a}$

$\sum_a q_1(x_0, a) = 1,\ q \geq 0$

PROOF IDEA: DUALITY

Primal: optimality in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $\sum_a q_{h+1,a} = \sum_a P_a^\top q_{h,a}$

PROOF IDEA: DUALITY

Primal: optimism in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $q_{h+1,a} = \sum_a \tilde{P}_a^\top q_{h,a}$ $\Big|$ $D\left(\tilde{P}(\cdot\,|x,a), \hat{P}(\cdot\,|x,a)\right) \leq \epsilon(x,a)$

PROOF IDEA: DUALITY

Primal: optimism in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$
subject to $q_{h+1,a} = \sum_a \tilde{P}_a^\top q_{h,a}$ | $D\left(\tilde{P}(\cdot\,|x,a), \hat{P}(\cdot\,|x,a)\right) \leq \epsilon(x,a)$

- Nonconvex due to bilinear constraint $\tilde{P}q$!
- Convex reparametrization: $J(x,a,x') = q(x,a)\tilde{P}(x'|x,a)$.
- Use assumptions on D to rewrite confidence constraint as

$$D\left(J(x,a,\cdot), q(x,a)\hat{P}(\cdot\,|x,a)\right) \leq q(x,a)\epsilon(x,a).$$

PROOF IDEA: DUALITY

Primal: optimism in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$ | $D\left(\tilde{P}(\cdot\,|x,a), \hat{P}(\cdot\,|x,a)\right) \leq \epsilon(x,a)$

subject to $q_{h+1,a} = \sum_a \tilde{P}_a^\top q_{h,a}$

- Nonconvex due to bilinear constraint $\tilde{P}q$!
- Convex reparametrization: $J(x,a,x') = q(x,a)\tilde{P}(x'|x,a)$.
- Use assumptions on D to rewrite confidence constraint as

$$D\left(J(x,a,\cdot), q(x,a)\hat{P}(\cdot\,|x,a)\right) \leq q(x,a)\epsilon(x,a).$$

- Establish strong duality: $\max_{q,\tilde{P}} \min_V \mathcal{L}(q,\tilde{P};V) = \min_V \max_{q,\tilde{P}} \mathcal{L}(q,\tilde{P};V)$.
- Exploit the local nature of confidence constraints.

PROOF IDEA: DUALITY

Primal: optimism in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $q_{h+1,a} = \sum_a \tilde{P}_a^\top q_{h,a}$ | $D\left(\tilde{P}(\cdot\,|x,a), \hat{P}(\cdot\,|x,a)\right) \leq \epsilon(x,a)$

Equivalent due to Lagrangian duality

Dual: optimism in value-function space

as characterized by the Bellman optimality equations

$$V_h^+ = \max_a \left\{ r_a + \mathrm{CB}_{h,a} + \hat{P}_a V_{h+1}^+ \right\}$$

IMPLICATIONS

Optimism in model space:
construct a confidence set around P and jointly optimize over models & policies

Optimism in value space:
construct upper confidence bounds directly on the optimal value function V^*

☺ simple probabilistic analysis
just show that $P \in \mathcal{P}$!

☹ complicated to implement
need to search jointly over models and policies

☹ loose bounds
best known regret guarantees are suboptimal $O\left(HS\sqrt{AT}\right)$

☹ complicated to analyze
need recursive arguments to show optimistic property of V^+

☺ easy to implement
dynamic programming with $\hat{P}$ and $r + CB$

☺ tight bounds
optimal regret bounds $O\left(H\sqrt{SAT}\right)$

IMPLICATIONS

Optimism in model space:
construct a confidence set around P and jointly optimize over models & policies

Optimism in value space:
construct upper confidence bounds directly on the optimal value function V^*

Best of both worlds!

- Simple probabilistic analysis and easy implementation!
- Simple regret bound: $\mathrm{Regret}_T \leq \sum_{t=1}^{T} \sum_{h=1}^{H} \mathrm{CB}_{h,t}(x_{h,t}, a_{h,t}) + O(H\sqrt{SAT})$
- If the exact form of CB is difficult to calculate, you can use a tractable upper bound CB^+ and retain the guarantees

IMPLICATIONS

Optimism in model space:
construct a confidence set around P and jointly optimize over models & policies

Optimism in value space:
construct upper confidence bounds directly on the optimal value function V^*

Best of both worlds!

- Simple probabilistic analysis and easy implementation!
- Simple regret bound: $\mathrm{Regret}_T \leq \sum_{t=1}^{T} \sum_{h=1}^{H} \mathrm{CB}_{h,t}(x_{h,t}, a_{h,t}) + O(H\sqrt{SAT})$
- If the exact form of CB is difficult to calculate, you can use a tractable upper bound CB^+ and retain the guarantees

Downside: bounds still loose by a factor $\sqrt{S}$ ☹

LINEAR FUNCTION APPROXIMATION

Assumption: factored linear MDP

The transition matrix factorizes as

$$P_a = \Phi M_a,$$

where the rows of Φ correspond to some known feature vectors $\varphi(x) \in \mathbb{R}^d$

Implies realizability of Q-function approximation:
every Q function can be written as $Q(x,a) = \langle \theta_a, \varphi(x) \rangle$

LINEAR FUNCTION APPROXIMATION

Assumption: factored linear MDP

The transition matrix factorizes as

$$P_a = \Phi M_a,$$

where the rows of Φ correspond to some known feature vectors $\varphi(x) \in \mathbb{R}^d$

Implies realizability of Q-function approximation:
every Q function can be written as $Q(x,a) = \langle \theta_a, \varphi(x) \rangle$

Dual: optimality in value-function space

as characterized by the projected Bellman optimality equations

$$Q^*_{h,a} = \Pi_\Phi \left[r_a + P_a \max_{a'} Q^*_{h+1,a'} \right]$$

PRIMAL-DUAL FORMULATION

NEW

Primal: optimality in trajectory space

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle \\ \text{subject to} \quad & \textstyle\sum_a q_{h+1,a} = \sum_a P_a W_{h,a} \Phi \omega_{h,a} \\ & \Phi^\top q_{h,a} = \Phi^\top W_{h,a} \Phi \omega_{h,a} \end{aligned}$$

Equivalent due to Linear Programming duality

Dual: optimality in value-function space

as characterized by the projected Bellman optimality equations

$$Q^*_{h,a} = \Pi_\Phi \left[r_a + P_a \max_{a'} Q^*_{h+1,a'} \right]$$

BUILDING A REFERENCE MODEL

Idea:

Construct confidence sets around LSTD reference model $\widehat{P}_{t,a} = \Phi \widehat{M}_{t,a}$ with

$$\widehat{M}_{t,a} = \Sigma_{t,a}^{-1} \sum_{k=1}^{t} \mathbb{I}_{\{a_k = a\}} \varphi(x_k) e_{x'_k}$$

and observe that $\left(\widehat{M}_{t,a} - M_a\right) v$ is a vector-valued martingale for any v!

$$\Sigma_{t,a} = I + \sum_{k=1}^{t} \varphi(x_k) \varphi(x_k)^\top$$

Bradtke and Barto (1996), Boyan (1998), Parr et al. (2008)

BUILDING A REFERENCE MODEL

Idea:

Construct confidence sets around LSTD reference model $\widehat{P}_{t,a} = \Phi \widehat{M}_{t,a}$ with

$$\widehat{M}_{t,a} = \Sigma_{t,a}^{-1} \sum_{k=1}^{t} \mathbb{I}_{\{a_k = a\}} \varphi(x_k) e_{x'_k}$$

and observe that $\left(\widehat{M}_{t,a} - M_a\right)v$ is a vector-valued martingale for any v!

Lemma

$$\left\|\left(\widehat{M}_{t,a} - M_a\right)v\right\|_{\Sigma_{t,a}} \leq C\sqrt{d}\|v\|_\infty$$

Abbasi-Yadkori, Pál and Szepesvári (2011)

$$\Sigma_{t,a} = I + \sum_{k=1}^{t} \varphi(x_k)\varphi(x_k)^\top$$

Bradtke and Barto (1996), Boyan (1998), Parr et al. (2008)

LOCAL AND GLOBAL OPTIMISM

Local confidence sets

$$D\left(\tilde{P}(\cdot\,|x,a), \hat{P}(\cdot\,|x,a)\right) \le \epsilon(x,a)$$

Least-squares VI with local exploration bonuses

$$\mathrm{CB}(x,a) = C\|\varphi(x)\|_{\Sigma_{t,a}^{-1}}$$

- Equivalent to LSVI-UCB by Jin et al. (COLT 2020)!
- Regret= $O\left(\sqrt{H^3 d^3 T}\right)$
- Efficient implementation

LOCAL AND GLOBAL OPTIMISM

Local confidence sets

$$D\left(\tilde{P}(\cdot\,|x,a), \hat{P}(\cdot\,|x,a)\right) \le \epsilon(x,a)$$

Least-squares VI with local exploration bonuses

$$\mathrm{CB}(x,a) = C\|\varphi(x)\|_{\Sigma_{t,a}^{-1}}$$

- Equivalent to LSVI-UCB by Jin et al. (COLT 2020)!
- Regret= $O\left(\sqrt{H^3 d^3 T}\right)$
- Efficient implementation

Global confidence sets

$$\left\|\left(\widehat{M}_{t,a} - \widetilde{M}_a\right)v\right\|_{\Sigma_{t,a}} \le \epsilon$$

Least-squares VI with global exploration bonuses

$$\mathrm{CB}(x,a) = \langle B_a, \varphi(x)\rangle$$

with $\|B_a\|_{\Sigma_{t,a}} \le \epsilon$

- Equivalent to ELEANOR by Zanette et al. (ICML 2020)!
- Regret= $O\left(d\sqrt{H^3 T}\right)$
- No efficient implementation

LOCAL AND GLOBAL OPTIMISM

Model-based perspective
(=simple probabilistic analysis)

Least-squares VI with local exploration bonuses

$\mathrm{CB}(x,a) = C\|\varphi(x)\|_{\Sigma_{t,a}^{-1}}$

- Equivalent to LSVI-UCB by Jin et al. (COLT 2020)!
- Regret= $O\left(\sqrt{H^3 d^3 T}\right)$
- Efficient implementation

Least-squares VI with global exploration bonuses

$\mathrm{CB}(x,a) = \langle B_a, \varphi(x)\rangle$

with $\|B_a\|_{\Sigma_{t,a}} \leq \epsilon$

- Equivalent to ELEANOR by Zanette et al. (ICML 2020)!
- Regret= $O\left(d\sqrt{H^3 T}\right)$
- No efficient implementation

LOCAL AND GLOBAL OPTIMISM

Model-based perspective

(=simple probabilistic analysis)

Least-squares VI with local exploration bonuses

$$\mathrm{CB}(x,a) = C\|\varphi(x)\|_{\Sigma_{t,a}^{-1}}$$

- Equivalent to LSVI-UCB by Jin et al. (COLT 2020)!
- Regret= $O\left(\sqrt{H^3 d^3 T}\right)$
- Efficient implementation

Least-squares VI with global exploration bonuses

$$\mathrm{CB}(x,a) = \langle B_a, \varphi(x)\rangle$$

with $\|B_a\|_{\Sigma_{t,a}} \leq \epsilon$

- Equivalent to ELEANOR by Zanette et al. (ICML 2020)!
- Regret= $O\left(d\sqrt{H^3 T}\right)$
- No efficient implementation

IMPLEMENTING ELEANOR

Eleanor in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $q_{h+1,a} = \sum_a \widetilde{M}_a^\top \Phi^\top W_{h,a} \Phi \omega_{h,a}$

$\Phi^\top q_{h,a} = \Phi^\top W_{h,a} \Phi \omega_{h,a}$

$$\sup_{v \in \mathcal{V}} \left\| \left(\widetilde{M}_a - \widehat{M}_a \right) v \right\|_\Sigma \leq \epsilon$$

IMPLEMENTING ELEANOR

ELEANOR in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $q_{h+1,a} = \sum_a \widetilde{M}_a^\top \Phi^\top W_{h,a} \Phi \omega_{h,a}$

$\Phi^\top q_{h,a} = \Phi^\top W_{h,a} \Phi \omega_{h,a}$

$\sup_{v \in \mathcal{V}} \left\| (\widetilde{M}_a - \widehat{M}_a) v \right\|_\Sigma \leq \epsilon$

- Nonconvex due to bilinear constraint $\Phi \widetilde{M}_a W_{h,a} \Phi \omega_{h,a}$!
- Previous tricks (convex reparametrization, etc.) don't work!!

IMPLEMENTING ELEANOR

ELEANOR in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $q_{h+1,a} = \sum_a \widetilde{M}_a^\top \Phi^\top W_{h,a} \Phi \omega_{h,a}$

$\Phi^\top q_{h,a} = \Phi^\top W_{h,a} \Phi \omega_{h,a}$

$$\sup_{v \in \mathcal{V}} \left\| (\widetilde{M}_a - \widehat{M}_a) v \right\|_\Sigma \leq \epsilon$$

- Nonconvex due to bilinear constraint $\widetilde{M}_a^\top \Phi^\top W_{h,a} \Phi \omega_{h,a}$!
- Previous tricks (convex reparametrization, etc.) don't work!!
- Can be written as convex maximization problem essentially identical to LinUCB / OFUL

☹

CONCLUSION

- Current optimistic exploration methods may be closer to each other than we thought!
- Model-based view allows simpler algorithm design & analysis
- Open challenges:
  - Closing the gaps between the bounds?
  - Model-based theory for misspecified models?
    (some concurrent results by Lykouris et al., 2020)
  - More general function approximation?
  - …

CONCLUSION

- Current optimistic exploration methods may be closer to each other than we thought!
- Model-based view allows simpler algorithm design & analysis
- Open challenges:
  - Closing the gaps between the bounds?
  - Model-based theory for misspecified models?
    (some concurrent results by Lykouris et al., 2020)
  - More general function approximation?
  - …

Model-based optimism is alive! 

Thanks!!!

PRIMAL REALIZABILITY

Primal: optimality in trajectory space

maximize $\sum_{h=1}^{H} \langle q_{h,a}, r_{h,a} \rangle$

subject to $q_{h+1,a} = \sum_a P_a^\top W_{h,a} \Phi \omega_{h,a}$

$\Phi^\top q_{h,a} = \Phi^\top W_{h,a} \Phi \omega_{h,a}$

If transition model is factored as $P_a = \Phi M_a$, all feasible q's are feasible in the original LP:

$$q_{h+1,a} = \sum_a P_a^\top W_{h,a} \Phi \omega_{h,a} = \sum_a M_a^\top \Phi^\top W_{h,a} \Phi \omega_{h,a} = \sum_a M_a^\top \Phi^\top q_{h,a} = \sum_a P^\top q_{h,a}$$