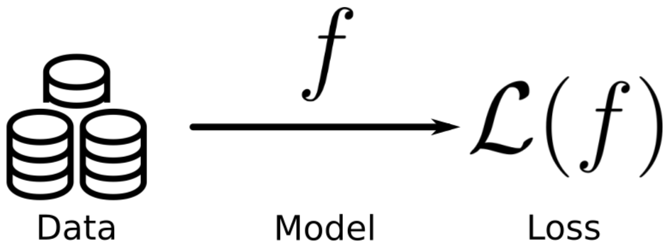
Entropy Regularization in RL

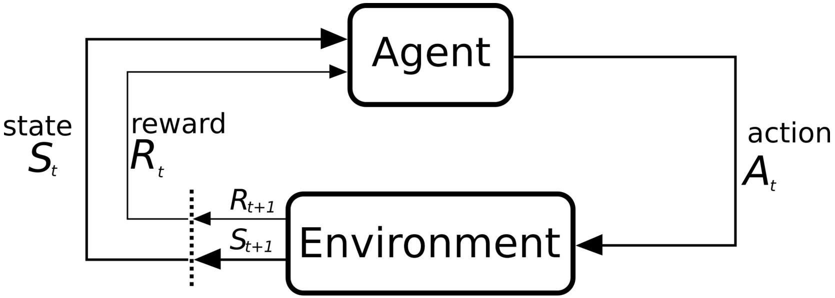
Riad Akrour

November 9, 2021

Data Sources in ML

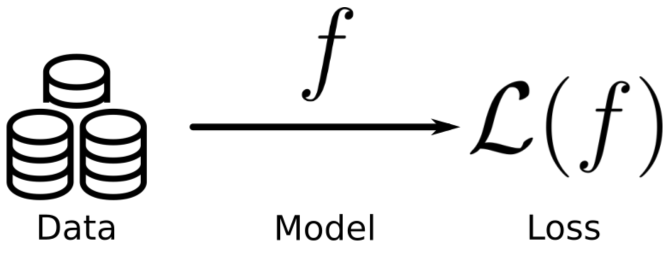


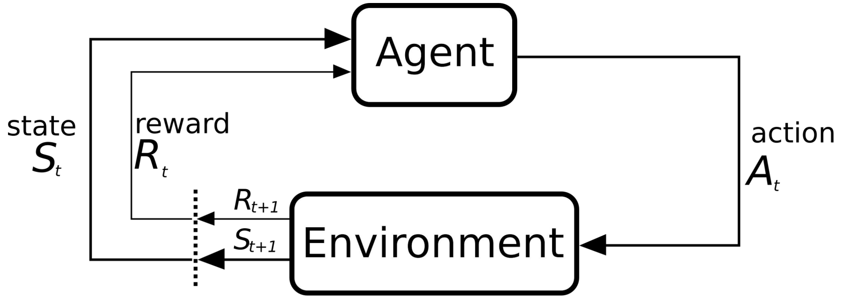
(Un)supervised Learning



Reinforcement Learning

Data Sources in ML





(Un)supervised Learning

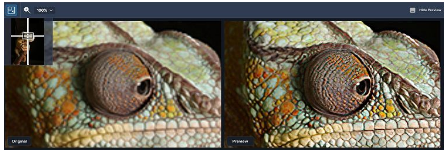
- i.i.d. dataset: agent learns from and is used on data with the same distribution

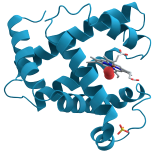
Reinforcement Learning

- Constant change to data generating process (at least in online RL)
- Large emphasis on out of distribution generalization

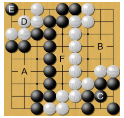
Successes of Machine Learning

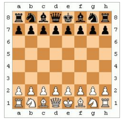
- **(Un)supervised Learning**
  - Machine translation, Speech recognition
  - AI image upscaling
  - Drug discovery

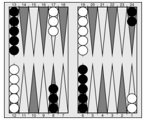




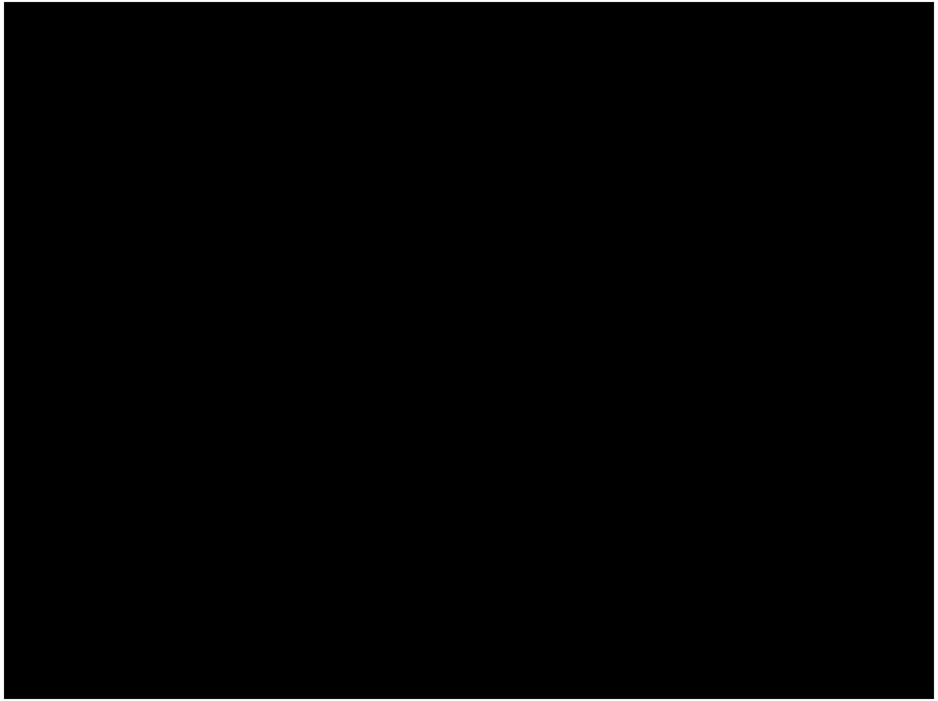
- **Reinforcement Learning**
  - Backgammon (Tesauro et al., 94), Chess (Hsu et al., 96), Atari (Mnih et al., 13), Go (Silver et al., 16)
  - Limited to game domains (closed world, unlimited data)







RL in the Physical World



High Acceleration Reinforcement Learning for Real-World Juggling with Binary Rewards
K. Ploeger, M. Lutter, J. Peters
CoRL20

- Small number of parameters (<20)
- Initialize from expert demonstrations
- Black-box optimization

Learning Goals

- **Entropy Regularization in RL**
  - Algorithms overview
    - From black-box optimization to deep reinforcement learning
  - Why is it important?

Relative Entropy Policy Search (REPS)

- For Gaussian search distribution $\pi_k(\theta) = \mathcal{N}(\theta|\mu_k, \Sigma_k)$, update following

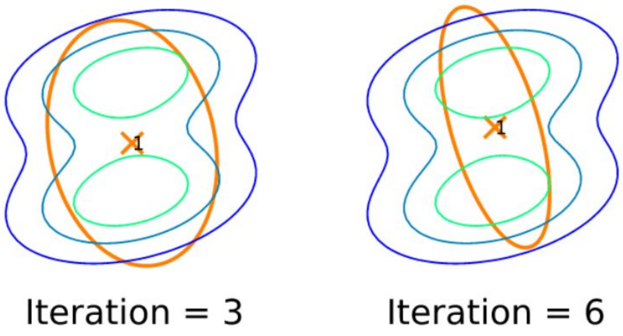
$$\max_{\pi_k} \quad \mathbb{E}_{\theta\sim\pi_k}\left[R(\theta)\right]$$

(Maximize rewards)

$$s.t. \quad \mathrm{KL}\left(\pi_k||\pi_{k-1}\right) \le \epsilon$$

(Do not change policy too much)

Deisenroth et al.

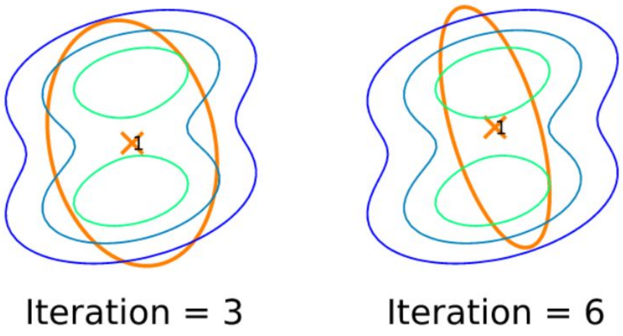


Relative Entropy Policy Search (REPS)

- For Gaussian search distribution $\pi_k(\theta) = \mathcal{N}(\theta|\mu_k, \Sigma$

$$\max_{\pi_k} \quad \mathbb{E}_{\theta\sim\pi_k}\left[R(\theta)\right]$$
$$s.t. \quad \mathrm{KL}\left(\pi_k || \pi_{k-1}\right) \leq \epsilon$$

Deisenroth et al.



Kullback-Leibler divergence

$$\mathrm{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$$

- KL-divergence is <u>always positive</u>
- is <u>equal to 0</u> if p is q
- For <u>Gaussians</u>, it grows if e.g. the
  - mean shifts
  - covariance matrix rotates
  - covariance matrix shrinks

Relative Entropy Policy Search (REPS)

- For Gaussian search distribution $\pi_k(\theta) = \mathcal{N}(\theta|\mu_k, \Sigma_k)$, update following

$$\max_{\pi_k} \quad \mathbb{E}_{\theta\sim\pi_k}[R(\theta)]$$ **(Maximize rewards)**

$$s.t. \quad \mathrm{KL}(\pi_k || \pi_{k-1}) \leq \epsilon$$ **(Do not change policy too much)**

- Closed-form solution in probability space

$$\pi_k \propto \pi_{k-1} \exp\left(\frac{R}{\eta}\right)$$ With η, the dual variable of the Lagrangian function

A Survey on Policy Search for Robotics (Chap. 2.4.3); M. Deisenroth, G. Neumann, J. Peters; Foundations and Trends in Robotics, 2013

REPS – Algorithm

- Sample $\{\theta_1, \dots, \theta_K\}$ from (Gaussian) π_{k-1}
- Evaluate parameters and get $\{R(\theta_1), \dots, R(\theta_K)\}$
- Optimize dual function and get η
- $\pi_k \propto \pi_{k-1} \exp\left(\frac{R}{\eta}\right)$:
  - Can evaluate distribution at sample points $\{\theta_1, \dots, \theta_K\}$ up to normalization factor
  - Want π_k to remain Gaussian for easy sampling

    -> maximum likelihood fit of Gaussian π_k to samples $\{\theta_1, \dots, \theta_K\}$ with weights

    $$w_i = \pi_{k-1}(\theta_i) \exp\left(\frac{R(\theta_i)}{\eta}\right)$$

A Survey on Policy Search for Robotics (Chap. 2.4.3); M. Deisenroth, G. Neumann, J. Peters; Foundations and Trends in Robotics, 2013

REPS – Limitations (1/2)

- Bias in density estimation
  - If environment is stochastic $R(\theta_1)$ is a random variable
    - Performance of θ_1 is given by expected return $E[R(\theta_1)]$
    - $E[R(\theta_1)]$ can be approximated by (unbiased) Monte Carlo estimate $\hat{R}(\theta_1) = \frac{1}{N} \sum R^{[i]}(\theta_1)$
  - Because of exp func., unbiased estimators of $E[R(\theta_1)]$ still yields biased estimate of density
    - $E\left[\exp\left(\frac{\hat{R}(\theta_1)}{\eta}\right)\right] \geq \exp\left(\frac{E\left[\hat{R}(\theta_1)\right]}{\eta}\right) \geq \exp\left(\frac{E[R(\theta_1)]}{\eta}\right)$ (Jensen's inequality + unbiasedness)
    - Equality if there is no variance otherwise overestimation

REPS – Limitations (2/2)

- <u>KL-divergence violation from maximum likelihood step</u>
  - Although $\pi_{k-1} \exp\left(\frac{R}{\eta}\right)$ satisfies KL-divergence cst., its Gaussian approximation π_k might not
  - Especially true if sample set $\{\theta_1, \dots, \theta_K\}$ is small
    - Gaussian distribution can quickly converge to a point mass with little to no variance

- <u>An alternative update</u>: π_{k-1} is Gaussian and if R is quadratic and concave then $\pi_{k-1} \exp\left(\frac{R}{\eta}\right)$ Gaussian!

Model-based REPS (MORE)

- For Gaussian policies $\pi_k(\theta) = \mathcal{N}(\theta|\mu_k, \Sigma_k)$

$$\begin{aligned} \max_{\pi_k} \quad & \mathbb{E}_{\theta\sim\pi_k}\left[\hat{R}(\theta)\right] && \textbf{(Maximize rewards)} \\ s.t. \quad & \mathrm{KL}\left(\pi_k || \pi_{k-1}\right) \leq \epsilon && \textbf{(Do not change policy too much)} \\ & \mathcal{H}(\pi_{k-1}) - \mathcal{H}(\pi_k) \leq \beta && \textbf{(Prevent premature convergence)} \end{aligned}$$

- Closed-form solution in probability space

$\pi_k \propto \pi_{k-1}^{\eta/(\eta+\omega)} \exp\left(\frac{\hat{R}}{\eta+\omega}\right)$ for dual variables η and ω of the Lagrangian function

- Closed-form solution in parameter space (no maximum likelihood step)

Model-Based Relative Entropy Stochastic Search; A. Abdolmaleki, R. Lioutikov, N. Lau, L. Reis, J. Peters, G. Neumann; NeurIPS15

Model-based REPS (MORE)

- For Gaussian policies $\pi_k(\theta) = \mathcal{N}(\theta | \mu_k, \Sigma_k)$

$$\begin{aligned} \max_{\pi_k} \quad & \mathbb{E}_{\theta \sim \pi_k}\left[\hat{R}(\theta)\right] \\ s.t. \quad & \mathrm{KL}\left(\pi_k || \pi_{k-1}\right) \leq \epsilon \\ & \mathcal{H}(\pi_{k-1}) - \mathcal{H}(\pi_k) \leq \beta \end{aligned}$$

(Max

(Do not c

(Preven

- Closed-form solution in probability space

$$\pi_k \propto \pi_{k-1}^{\eta/(\eta+\omega)} \exp\left(\frac{\hat{R}}{\eta+\omega}\right)$$ for dual var

- Closed-form solution in parameter space (no m

KL-divergence

- For Gaussians, it grows if e.g. the
  - mean shifts
  - covariance matrix rotates
  - covariance matrix shrinks

Entropy difference

- For Gaussians, it grows if the
  - covariance matrix shrinks

Having both **decouples** the control of matrix shrinkage from the rest.

Typically **in practice**: move mean/rotate covariance at faster rate than shrink covariance

Model-Based Relative Entropy Stochastic Search; A. Abdolmaleki, R. Lioutikov, N.

MORE – Algorithm

- Sample $\{\theta_1, \dots, \theta_K\}$ from (Gaussian) π_{k-1}
- Evaluate parameters and get $\{R(\theta_1), \dots, R(\theta_K)\}$
- Fit quadratic model $\hat{R}$ to data (regression problem)
- Optimize dual function and get η and ω
- Compute π_k from $\pi_{k-1}, \hat{R}$, η and ω
- <u>Limitation</u>: regression problem manageable for cleverly parameterized (and closed-loop) policies
  - Impractical if e.g. θ parameters of neural network

Model-Based Relative Entropy Stochastic Search; A. Abdolmaleki, R. Lioutikov, N. Lau, L. Reis, J. Peters, G. Neumann; NeurIPS15

Step-based MORE (MOTO)

- For linear-Gaussian policies $\pi_k^t(a_t|s_t) = \mathcal{N}\left(a_t|K_t s_t, \Sigma_t\right)$ **(Closed loop policy)**

$$\max_{\pi_k^t} \quad \mathbb{E}_{s\sim p_{k-1}^t, a\sim\pi_k^t(.|s)}\left[\hat{Q}_{k-1}^t(s,a)\right]$$ **(Maximize returns)**

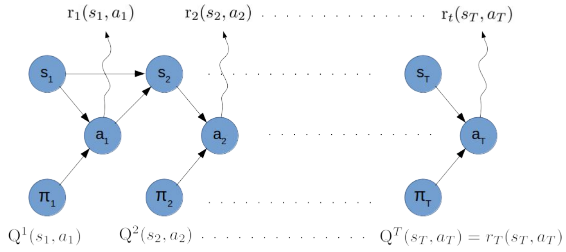
$$s.t. \quad \mathbb{E}_{s\sim p_{k-1}^t}\left[\mathrm{KL}\left(\pi_k^t(.|s)||\pi_{k-1}^t(.|s)\right)\right] \le \epsilon$$ **(Promote monotonic improvements)**

$$\mathbb{E}_{s\sim p_{k-1}^t}\left[\mathcal{H}(\pi_{k-1}^t(.|s)) - \mathcal{H}(\pi_k^t(.|s))\right] \le \beta$$ **(Prevent premature convergence)**

- Closed-form solution in probability space *and* parameter space

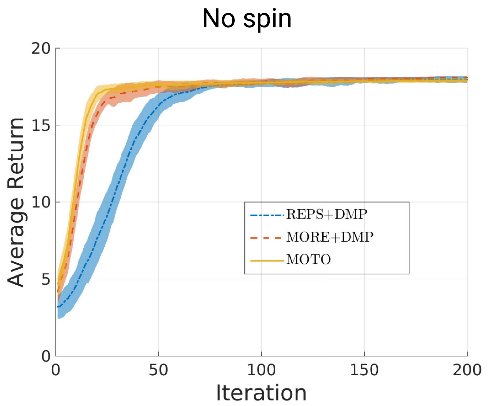
$$\pi_k^t(.|s) \propto \pi_{k-1}^t(.|s)^{\eta/(\eta+\omega)} \exp\left(\frac{\hat{Q}_{k-1}^t(s,.)}{\eta+\omega}\right)$$

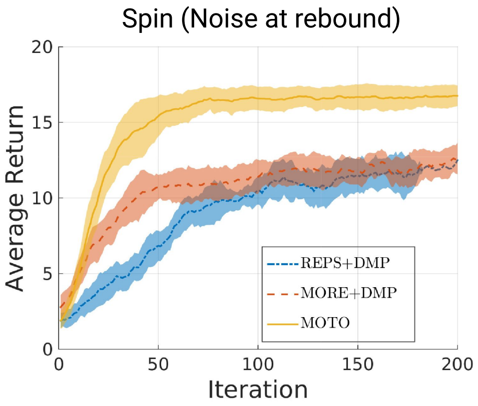
for dual variables η and ω of the Lagrangian function

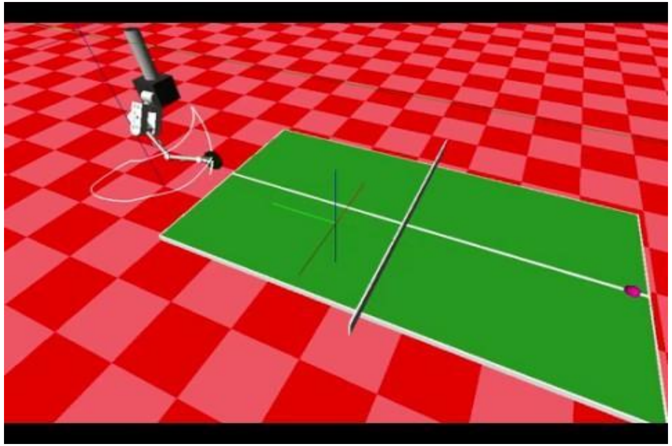


Model-free Trajectory-based Policy Optimization with Monotonic Improvement; R. Akrour, A. Abdolmaleki, H. Abdulsamad, J. Peters, G. Neumann; JMLR18

Open-Loop vs Closed Loop on Table Tennis







- Comparable to black-box methods with open-loop policies despite much larger search space (x800)
- Closed-loop nature of the policy copes with noisy environments
- Closed-form update outperforms control algorithms and TRPO (deep RL) on benchmark problems

Model-free Trajectory-based Policy Optimization with Monotonic Improvement; R. Akrour, A. Abdolmaleki, H. Abdulsamad, J. Peters, G. Neumann; JMLR18

MOTO – Limitation

- Linear Gaussian policies ill-suited for
  - Infinite horizon problems and in general problems with long (>150) planning horizons
  - Complex inputs (images, graphs...)

- <u>An alternative</u>: MPO algorithm [1]
  - Mean of policy given by neural network (or any other differentiable model)
  - Maximum likelihood step as in REPS but with additional KL-divergence constraint during fit

[1] Maximum A Posteriori Policy Optimisation; A. Abdolmaleki, J. Springenberg, Y. Tassa, R. Munos, N. Heess, M. Riedmiller; ICLR18

Natural Gradient and Deep RL - TRPO

- Around parameters θ of data generating policy π_{k-1}
  - Compute gradient *g* of objective $E_{s \sim p_{k-1}, a \sim \pi_k(. \mid s)} \left[\hat{A}_{k-1}(s,a) \right]$ for a first order approximation
  - Compute Hessian *F* of KL-divergence constraint for a second order approximation

- Approximated constrained problem admits closed form solution $\theta + \alpha F^{-1} g$ (so called Natural Gradient)
- Additional computational tricks:
  - Find NG by minimizing $\|Fp - g\|^2$ using conjugate gradient algorithm
  - Compute *Fp* by computing gradient(<gradient(KL), p>)
  - Add linesearch to ensure KL-divergence constraint is satisfied
- Works well on medium scale problems and more stable than other deep RL algorithms

Trust Region Policy Optimization; S. Schulman, S. Levine, P. Moritz, M. Jordan, P. Abbeel; ICML15

Gradient Clipping and Soft Constraints – PPO

- Solves the same optimization problem as TRPO (maximize advantage under KL-divergence constraint)
- Optimization routine of TRPO can be a bit slow when dealing with very large networks
  - PPO introduces tricks to (heuristically) tackle the optim. problem with vanilla gradient descent
- Main trick is clipped loss
  - Zero contribution (the gradient) of state-action pairs if $\frac{\pi_k(a \mid s)}{\pi_{k-1}(a \mid s)}$ deviates too much from 1
  - Implies a total variation 'constraint' between π_k and π_{k-1}

Proximal Policy Optimization Algorithms; S. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov; 2017

Gradient Clipping and Soft Constraints – PPO

- Solves the same optimization problem as TRPO (maximize adv
- Optimization routine of TRPO can be a bit slow when dealin
  - PPO introduces tricks to (heuristically) tackle the op
- Main trick is clipped loss
  - Zero contribution (the gradient) of state-ac
  - Implies a total variation ‘constraint’ between

$$E_{a\sim\pi_{k-1}(a\,|\,s)}\left[\left|\frac{\pi_k(a\,|\,s)}{\pi_{k-1}(a\,|\,s)} - 1\right|\right] \le \epsilon$$

$$\int_{\mathcal{A}} |\pi_k(a\,|\,s) - \pi_{k-1}(a\,|\,s)|\,\mathrm{d}a \le \epsilon$$

$$TV(\pi_k\,|\,\pi_{k-1}) \le \epsilon$$

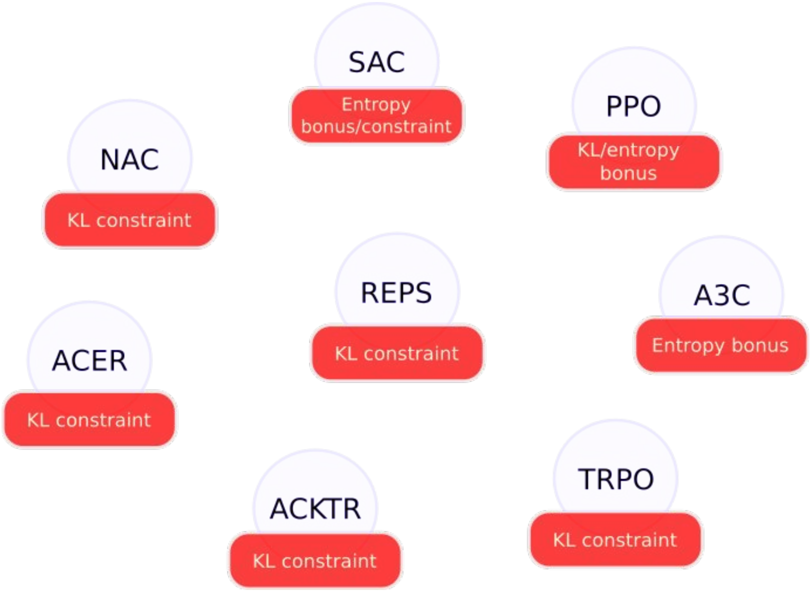
Proximal Policy Optimization Algorithms; S. Schulman, F. Wolski, P. Dhariwal, A. R

Gradient Clipping and Soft Constraints – PPO

- Solves the same optimization problem as TRPO (maximize advantage under KL-divergence constraint)
- Optimization routine of TRPO can be a bit slow when dealing with very large networks
    - PPO introduces tricks to (heuristically) tackle the optim. problem with vanilla gradient descent
- Main trick is clipped loss
    - Zero contribution (the gradient) of state-action pairs if $\frac{\pi_k(a \mid s)}{\pi_{k-1}(a \mid s)}$ deviates too much from 1
    - Implies a total variation 'constraint' between π_k and π_{k-1}
    - TV and KL are related through inequality and serve similar purposes (see second part)
    - KL-divergence term sometimes added to loss too

Proximal Policy Optimization Algorithms; S. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov; 2017

Entropy regularization in (deep) RL



Hard vs Soft Constraints

- Formulation with soft-constraints/bonuses is more common in the theory of (convex) optimization
- In practical algorithms it is common
  - That the KL-divergence is hard constrained
  - That entropy is added as a bonus
    - Notable exception is latter version of SAC
- Still an open question how to best formulate and solve policy update in approximate policy iteration

Optimization Issues in KL-Constrained Approximate Policy Iteration; N. Lazic, B. Hao, Y. Abbasi-Yadkori, D. Schuurmans, C. Szepesvari; 2021

Summary of Algorithms Overview

- KL-divergence (a.k.a. relative entropy) and entropy are widespread regularizers in RL
  - From black-box formulations typical in robotics to deep RL
- Constrained entropy regularized policy update can be solved in closed form in some cases
  - But a lot of tricks are involved to tackle the problems in deep RL
  - Finding the best formulation is still an active research area
- Soft formulation with entropy terms added to policy update loss trivial to implement
  - But does not seem to be popular in practice, especially for KL-divergence constraint
  - ...why is it important to bound the KL-divergence anyway?

Learning Goals

- **Entropy Regularization in RL**
  - Algorithms and applications in robotics
  - Why is it important?
    - Monotonic improvements in Approximate Policy Iteration (API)
    - Exploration in RL

Why does Entropy Regularization Helps?

- Typical entropy regularized policy update in Approximate Policy Iteration (API)

$$
\begin{aligned}
\max_{\pi_k^t} \quad & \mathbb{E}_{s\sim p_{k-1}^t, a\sim\pi_k^t(.|s)} \left[\hat{Q}_{k-1}^t(s,a)\right] \\
s.t. \quad & \mathbb{E}_{s\sim p_{k-1}^t} \left[\mathrm{KL}\left(\pi_k^t(.|s) || \pi_{k-1}^t(.|s)\right)\right] \leq \epsilon \\
& \mathbb{E}_{s\sim p_{k-1}^t} \left[\mathcal{H}(\pi_{k-1}^t(.|s)) - \mathcal{H}(\pi_k^t(.|s))\right] \leq \beta
\end{aligned}
$$

- Strict compliance with KL-divergence constraint important in practice… why?
- Objective and constraints expressed in terms of p_{k-1}^t… is it reasonable?

Policy Improvement

- In tabular RL, $\pi_k(s) = \arg\max Q_{k-1}(s, .)$
  - Take better action in **all states**
  - Immediately implies that $Q_k \geq Q_{k-1}$, i.e. for any *(s,a)* $Q_k(s,a) \geq Q_{k-1}(s,a)$
- Relaxation when using function approximators
  - $\pi_k = \arg\max_{\pi} \quad \mathbb{E}_{s\sim p_{k-1}, a\sim\pi(.|s)}\left[Q_{k-1}(s,a)\right]$
  - Take better actions in average of **previous state** distribution
  - What about average under the **current state** distribution $\mathbb{E}_{s\sim p_k, a\sim\pi(.|s)}\left[Q_{k-1}(s,a)\right]$?

Notation

- Π matrix representation of policy π

- Π matrix of size $|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|$

$$\Pi_{(s,(s',a))} = \pi(a|s) \text{ if } s = s', 0 \text{ else}$$

Notation

- Π matrix representation of policy π
- P transition matrix

- P matrix of size $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$

  $P_{((s,a),s')} = p(s'|s,a)$

Notation

- Π matrix representation of policy π
- P transition matrix
- Value function $V^\pi = \Pi R + \gamma \Pi P V^\pi$

- R matrix of size $|\mathcal{S}||\mathcal{A}| \times 1$

  $R_{((s,a),1)} = r(s,a)$

Notation

- Π matrix representation of policy π
- P transition matrix
- Value function $V^\pi = (I - \gamma \Pi P)^{-1} \Pi R$

- Policy induced state distribution

$$(I - \gamma \Pi P)^{-1}_{(s,s')} = \sum_{t=0}^{\infty} \gamma^t \left(\Pi P\right)^t_{(s,s')},$$
$$= \sum_{t=0}^{\infty} \gamma^t Pr(s_t = s' | s_0 = s; \pi),$$

- We define $\Pi_s = (I - \gamma \Pi P)^{-1}$

Notation

- Π matrix representation of policy π
- P transition matrix
- Value function $V^\pi = \Pi_s \Pi R$

Notation

- Π matrix representation of policy π
- P transition matrix
- Value function $V^\pi = \Pi_s \Pi R$
- Policy return $J(\pi) = \mu^T V^\pi$ for initial state distribution matrix μ

Performance Difference Lemma

- $V^{\pi} - V^{\pi'} = \Pi_s \Pi A^{\pi'}$

Performance Difference Lemma

- $V^{\pi} - V^{\pi'} = \Pi_s \Pi A^{\pi'}$

- Value difference

$$
\begin{aligned}
V^{\pi} - V^{\pi'} &= \Pi\left(\mathcal{R} + \gamma P V^{\pi}\right) - V^{\pi'}, \\
&= \Pi\left(\mathcal{R} + \gamma P\left(V^{\pi} + V^{\pi'} - V^{\pi'}\right)\right) - V^{\pi'}, \\
&= \gamma \Pi P\left(V^{\pi} - V^{\pi'}\right) + \Pi\left(\mathcal{R} + \gamma P V^{\pi'}\right) - V^{\pi'} \\
&= \gamma \Pi P\left(V^{\pi} - V^{\pi'}\right) + \Pi Q^{\pi'} - V^{\pi'}, \\
&= \gamma \Pi P\left(V^{\pi} - V^{\pi'}\right) + \Pi A^{\pi'}, \\
&= \left(I - \gamma \Pi P\right)^{-1} \Pi A^{\pi'}.
\end{aligned}
$$

Performance Difference Lemma

- $$\begin{aligned} V^{\pi} - V^{\pi'} &= \Pi_s \Pi A^{\pi'} \\ &= \Pi'_s \Pi A^{\pi'} + \left(\Pi_s - \Pi'_s\right) \Pi A^{\pi'} \end{aligned}$$

Performance Difference Lemma

- $V^{\pi} - V^{\pi'} = \Pi_s \Pi A^{\pi'}$

$$= \Pi'_s \Pi A^{\pi'} + (\Pi_s - \Pi'_s)$$

- State distribution difference

$$\begin{aligned}
\Pi_s - \Pi'_s &= \gamma \Pi P \Pi_s - \gamma \Pi' P \Pi'_s \\
&= \gamma \left(\Pi - \Pi' + \Pi'\right) P \Pi_s - \gamma \Pi' P \Pi'_s \\
&= \gamma \Pi' P (\Pi_s - \Pi'_s) + \gamma (\Pi - \Pi') P \Pi_s \\
&= \gamma \Pi'_s (\Pi - \Pi') P \Pi_s
\end{aligned}$$

Performance Difference Lemma

- $$\begin{aligned} V^{\pi} - V^{\pi'} &= \Pi_s \Pi A^{\pi'} \\ &= \Pi'_s \Pi A^{\pi'} + (\Pi_s - \Pi'_s) \Pi A^{\pi'} \\ &= \Pi'_s \Pi A^{\pi'} + \gamma \Pi'_s (\Pi - \Pi') P \Pi_s \Pi A^{\pi'} \end{aligned}$$

Performance Difference Lemma

- $$\begin{aligned} V^{\pi} - V^{\pi'} &= \Pi_s \Pi A^{\pi'} \\ &= \Pi'_s \Pi A^{\pi'} + \left(\Pi_s - \Pi'_s\right) \Pi A^{\pi'} \\ &= \Pi'_s \Pi A^{\pi'} + \gamma \Pi'_s \left(\Pi - \Pi'\right) P \Pi_s \Pi A^{\pi'} \end{aligned}$$
- $$J^{\pi} - J^{\pi'} = \mu^T \Pi'_s \Pi A^{\pi'} + \gamma \mu^T \Pi'_s \left(\Pi - \Pi'\right) P \Pi_s \Pi A^{\pi'}$$
- Expressed policy return as a function of old advantage under old state distribution
    - + term small when new policy is close to old one

Lower Bounding the Policy Return

- Previous expression contains Π_s which is hard to quantify
  - Prior work will mainly differ in bounding $||\mu^T\left(\Pi_s - \Pi'_s\right)||_\infty$

Lower Bounding the Policy Return

- Previous expression contains Π_s which is hard to quantify
  - Prior work will mainly differ in bounding $||\mu^T\left(\Pi_s - \Pi'_s\right)||_\infty$
- CPI (Kakade et al. ICML02): $||\mu^T\left(\Pi_s - \Pi'_s\right)||_\infty \leq \frac{2\alpha\gamma}{(1-\gamma)^2}$
  - Where $\Pi = \alpha\Pi^g - (1-\alpha)\Pi'$ mixes previous policy with policy maximizing old advantage
  - Improvement of policy return can be guaranteed for small enough α

$$J^\pi - J^{\pi'} \geq \frac{\left(\mu^T \Pi'_s \Pi^g A^{\pi'}\right)^2}{8}$$

Lower Bounding the Policy Return

- Previous expression contains Π_s which is hard to quantify
  - Prior work will mainly differ in bounding $||\mu^T \left(\Pi_s - \Pi'_s\right)||_\infty$
- CPI (Kakade et al. ICML02): $||\mu^T \left(\Pi_s - \Pi'_s\right)||_\infty \leq \frac{2\alpha\gamma}{(1-\gamma)^2}$
- USPI (Pirotta et al. ICML13): $||\mu^T \left(\Pi_s - \Pi'_s\right)||_\infty \leq \frac{\gamma}{(1-\gamma)^2}||\Pi - \Pi'||_\infty$
  - $$\begin{aligned}\left\|\Pi - \Pi'\right\|_\infty &= \max_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi'(a|s)| \\ &= 2\max_{s\in\mathcal{S}} \mathrm{TV}(\pi(.|s) \parallel \pi'(.|s))\end{aligned}$$

Lower Bounding the Policy Return

- Previous expression contains Π_s which is hard to quantify
  - Prior work will mainly differ in bounding $||\mu^T \left(\Pi_s - \Pi_s'\right)||_\infty$
- CPI (Kakade et al. ICML02): $||\mu^T \left(\Pi_s - \Pi_s'\right)||_\infty \leq \frac{2\alpha\gamma}{(1-\gamma)^2}$
- USPI (Pirotta et al. ICML13): $||\mu^T \left(\Pi_s - \Pi_s'\right)||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{s \in \mathcal{S}} \mathrm{TV}(\pi(.|s) \parallel \pi'(.|s))$
- TRPO (Schulman et al. ICML15): $||\mu^T \left(\Pi_s - \Pi_s'\right)||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{s \in \mathcal{S}} \sqrt{\frac{1}{2}\mathrm{KL}(\pi(.|s) \parallel \pi'(.|s))}$
  - Pinsker's inequality: $\mathrm{TV} \leq \sqrt{\frac{1}{2}\mathrm{KL}}$

Lower Bounding the Policy Return

- Previous expression contains Π_s which is hard to quantify
  - Prior work will mainly differ in bounding $||\mu^T \left(\Pi_s - \Pi'_s\right)||_\infty$
- CPI (Kakade et al. ICML02): $||\mu^T \left(\Pi_s - \Pi'_s\right)||_\infty \leq \frac{2\alpha\gamma}{(1-\gamma)^2}$
- USPI (Pirotta et al. ICML13): $||\mu^T \left(\Pi_s - \Pi'_s\right)||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{s\in\mathcal{S}} \mathrm{TV}(\pi(.|s) \parallel \pi'(.|s))$
- TRPO (Schulman et al. ICML15): $||\mu^T \left(\Pi_s - \Pi'_s\right)||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{s\in\mathcal{S}} \sqrt{\frac{1}{2}\mathrm{KL}(\pi(.|s) \parallel \pi'(.|s))}$
- CPO (Achiam et al. ICML17): $||\mu^T \left(\Pi_s - \Pi'_s\right)||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \mathbb{E}_{s\sim\pi'} \left[\mathrm{TV}(\pi(.|s) \parallel \pi'(.|s))\right]$

Pessimism of (TV) Constrained Policy Updates

- For never seen (s,a), fair to assume that it has the worst reward and leads to the worst states
  - Necessary if we want to provide guarantees that hold in the worst case
- Bounds previously discussed assume the worst when bounding $\Pi_s \Pi A^{\pi'}$ even for pairs explored by π'
  - Only saving grace is to minimize divergence to π'
  - In that sense similar to imitation learning [1]
  - Bounds too pessimistic to be useful in practice
- Open question: how to incorporate pessimism that takes into account visited (s,a) pairs in deep RL

[1] The Importance of Pessimism in Fixed Dataset Policy Optimization; J. Buckman, C. Gelada, M. Bellemare; ICLR2021

Optimism in RL

- For never seen (s,a), fair to assume that it has the best reward and leads to the best states
  - Necessary if we want to provide guarantees that we find the optimal policy
- For explored (s,a) pairs, high probability bounds of reward and future state distribution depend on
  - Number of times (s,a) has been visited and optionally its variance (concentration inequalities)
- Estimating these quantities is a similar open problem as in previous slide
  - Entropy bonus/constraint is the optimistic pendant of TV/KL-divergence constraint
  - Ensures that (s,a) pairs are explored sufficiently many times
  - Can be overly optimistic towards very bad pairs and is mostly used as an heuristic

Summary

- For principled RL algorithm development, learner should memorize visitation counts for (s,a) pairs
  - Doable in tabular settings or with trees (MCTS)
  - Remains an open problem for large MDPs requiring function approximators
- Entropy regularization provides useful tools to cope with lack of visitation counts and incorporates
  - Optimistic view: to ensure sufficient exploration through entropy lower bound
  - Pessimistic view: to ensure policy improvement through TV/KL-divergence upper bound
- Future research needed to correct for excessive optimism/pessimism