Relative Entropy Policy Search

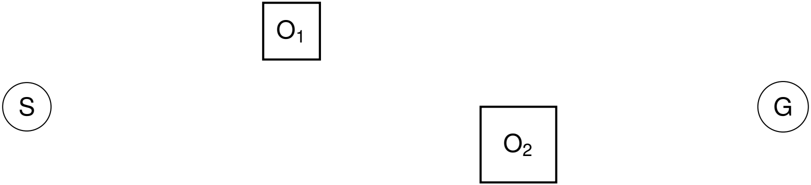
Jan Peters, Katharina Mülling, Yasemin Altün

Structure

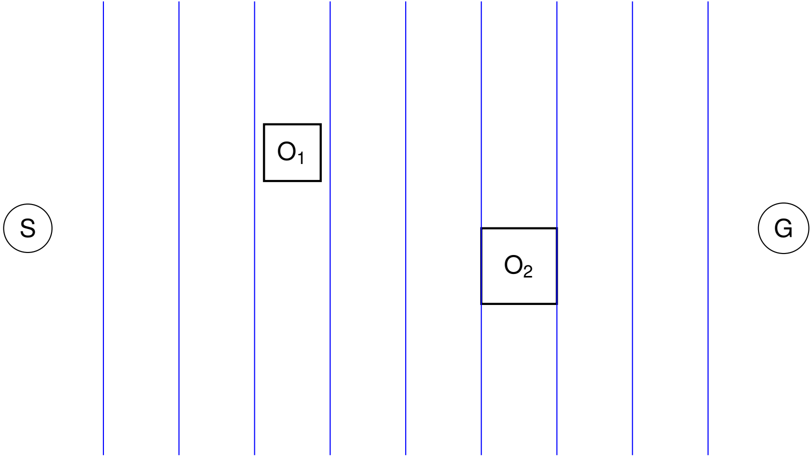
TECHNISCHE UNIVERSITÄT DARMSTADT

- Modeling Example
- Problem Statement
- REPS
- Policy Iteration with REPS

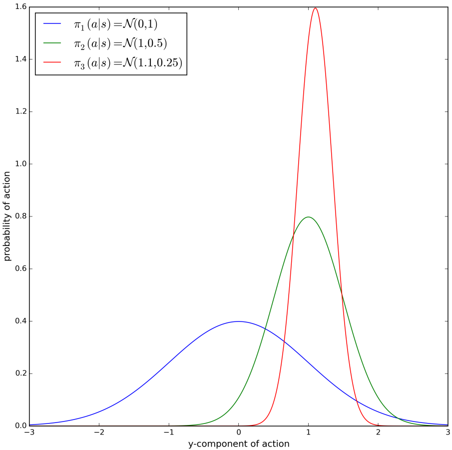
State-Action Space
Finding a trajectory



State-Action Space
Finding a trajectory



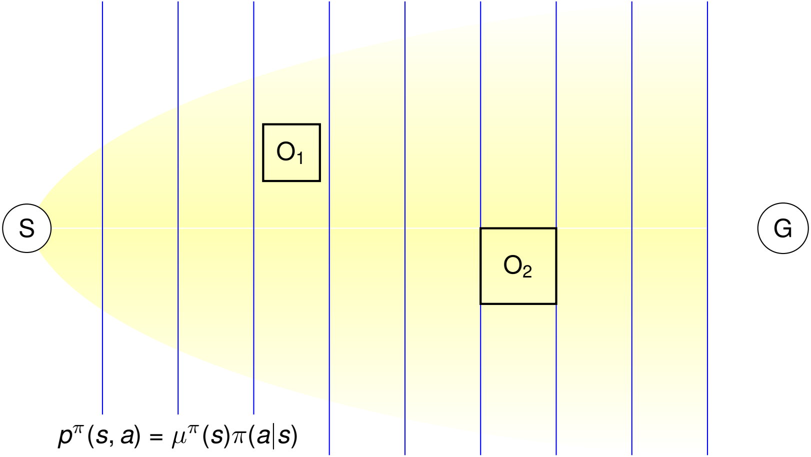
Policy



$\pi_1(a|s) = \mathcal{N}(0,1)$
$\pi_2(a|s) = \mathcal{N}(1,0.5)$
$\pi_3(a|s) = \mathcal{N}(1.1,0.25)$
probability of action
y-component of action
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
−3
−2
−1
0
1
2
3

State-Action Space

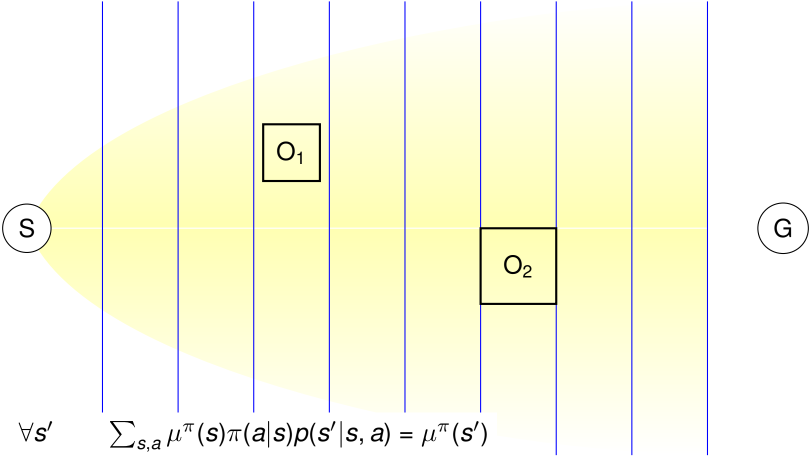
Finding a trajectory



$p^{\pi}(s, a) = \mu^{\pi}(s)\pi(a|s)$

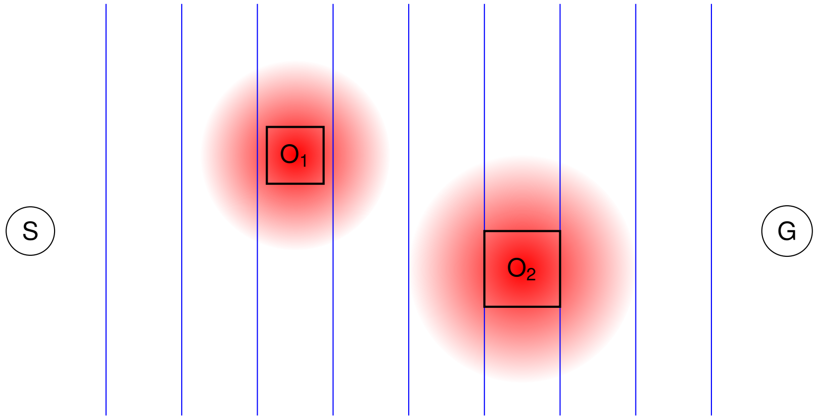
State-Action Space

Finding a trajectory



$\forall s' \qquad \sum_{s,a} \mu^{\pi}(s)\pi(a|s)p(s'|s,a) = \mu^{\pi}(s')$

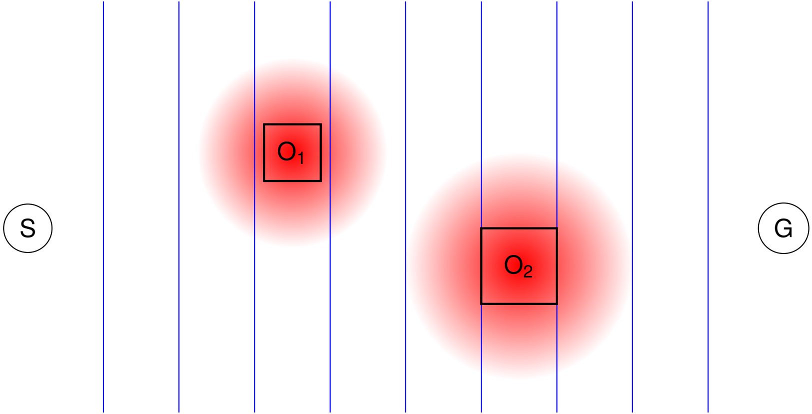
State-Action Space
Finding a trajectory



$r(s, a) =$

State-Action Space
Finding a trajectory

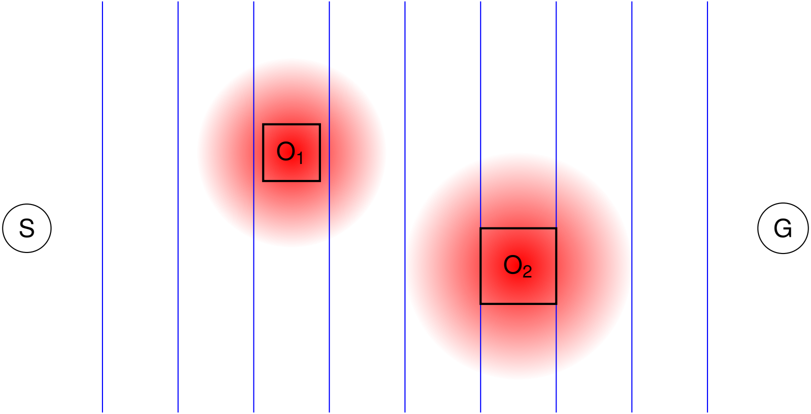
TECHNISCHE UNIVERSITÄT DARMSTADT



$$r(s, a) = -\sum_i \max(0, \mathrm{dist}(O_i, s, a) - \mathrm{radius}(O_i))^2$$

State-Action Space

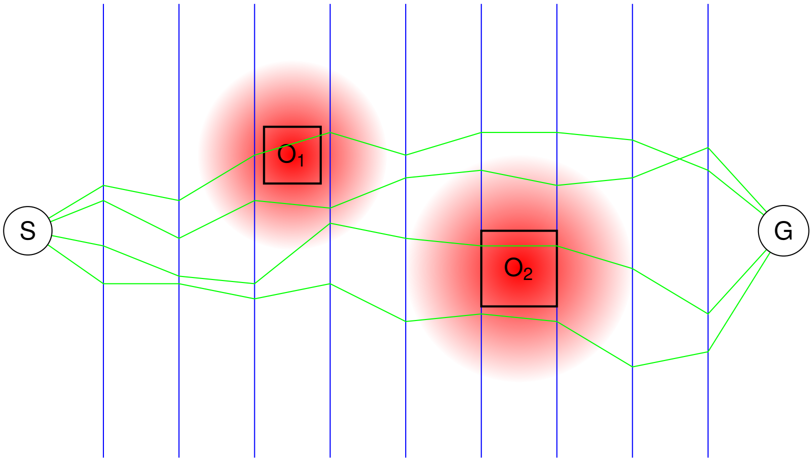
Finding a trajectory



$$r(s, a) = -\sum_i \max(0, \mathrm{dist}(\mathrm{O}_i, s, a) - \mathrm{radius}(\mathrm{O}_i))^2 - \mathrm{length}(a)$$

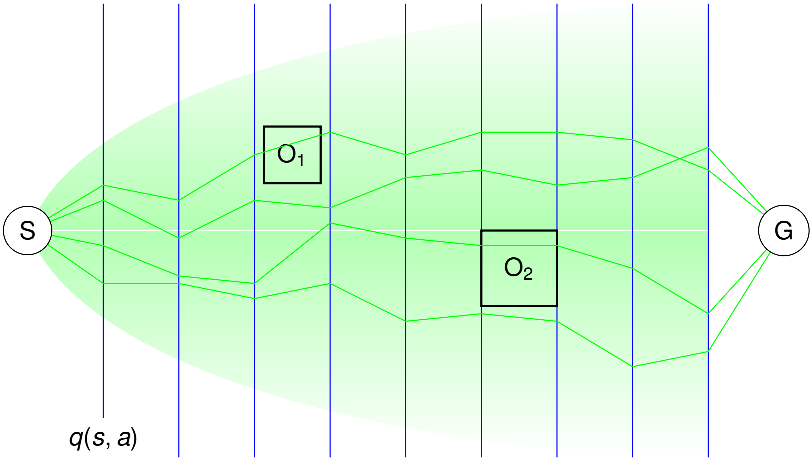
State-Action Space

Finding a trajectory



State-Action Space

Finding a trajectory



Kullback-Leibler Divergence – *Relative Entropy*

TECHNISCHE UNIVERSITÄT DARMSTADT

$$h_d(x) = -\int p(x) \log p(x) \, \mathrm{d}x$$

$$h_c(x) = -\int p(x) \log q(x) \, \mathrm{d}x$$

$$h_d(x) - h_c(x) = -\int p(x) \log p(x) \, \mathrm{d}x - \left(-\int p(x) \log q(x) \, \mathrm{d}x \right)$$

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

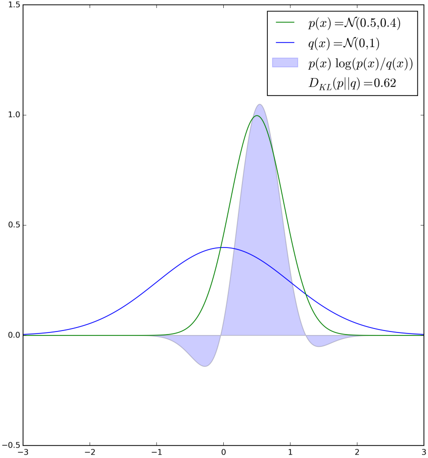
Kullback-Leibler Divergence – *Relative Entropy*

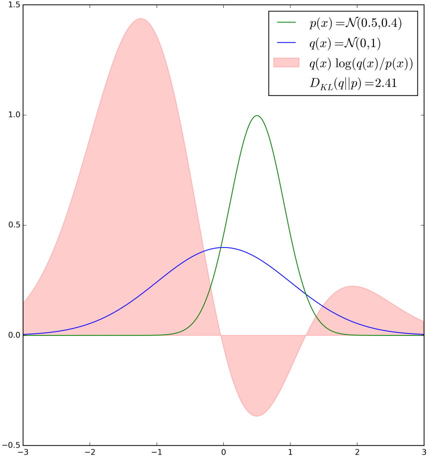
TECHNISCHE UNIVERSITÄT DARMSTADT

$$h_d(x) = -\int p(x) \log p(x) \, \mathrm{d}x$$
$$h_c(x) = -\int p(x) \log q(x) \, \mathrm{d}x$$

$$h_d(x) - h_c(x) = -\int p(x) \log p(x) \, \mathrm{d}x - \left(- \int p(x) \log q(x) \, \mathrm{d}x \right)$$
$$D_{\mathrm{KL}}(p \| q) = -\int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

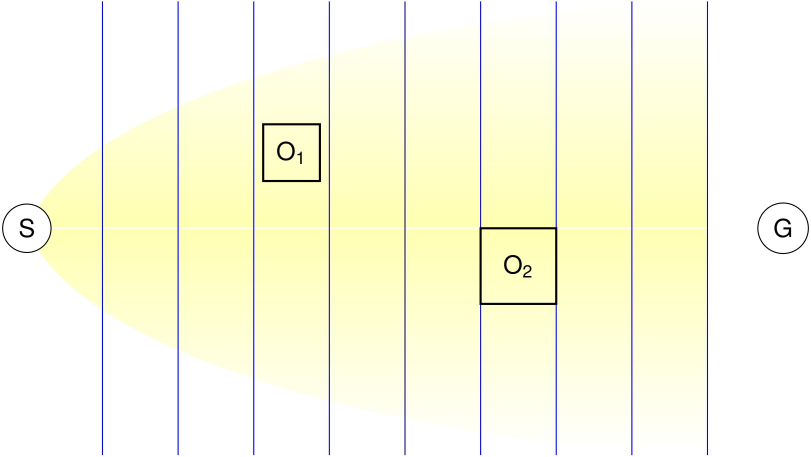
Kullback-Leibler Divergence – *Relative Entropy*





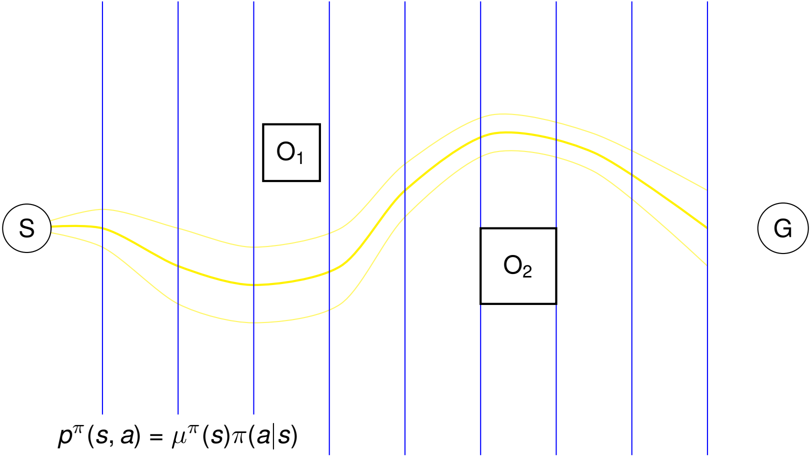
State-Action Space

Finding a trajectory



State-Action Space

Finding a trajectory



$$p^{\pi}(s,a) = \mu^{\pi}(s)\pi(a|s)$$

Problem Statement - *informal*

Maximize expected reward $J(\pi)$ while:

- Limiting information loss through Kullback-Leibler divergence
- Ensuring stationary features under π
- Enforcing that $p^\pi(s, a) = \mu^\pi(s)\pi(a|s)$ is a probability distribution

Problem Statement - *formal*

$$\max_{\pi,\mu^\pi} J(\pi) = \sum_{s,a} \mu^\pi(s)\pi(a|s)r(s,a)$$

$$s.t. \quad \epsilon \geq D(p^\pi || q)$$

$$\sum_{s'} \mu^\pi(s')\phi_{s'} = \sum_{s,a,s'} \mu^\pi(s)\pi(a|s)p(s'|s,a)\phi_{s'}$$

$$1 = \sum_{s,a} \mu^\pi(s)\pi(a|s) = \sum_{s,a} p^\pi(s,a)$$

Problem Statement - *Lagrangian*

$$
\begin{aligned}
L = &\sum_{s,a} p^{\pi}(a,s) r(s,a) \\
&\cdot \eta \left(\epsilon - \sum_{s,a} p^{\pi}(s,a) \log \frac{p^{\pi}(s,a)}{q(s,a)} \right) \\
&\cdot \theta \left(- \sum_{s',a'} p^{\pi}(s',a') \phi_{s'} + \sum_{s,a,s'} p^{\pi}(s,a) p(s'|s,a) \phi_{s'} \right) \\
&\cdot \lambda \left(1 - \sum_{s,a} p^{\pi}(s,a) \right)
\end{aligned}
$$

Derivation of REPS – Part I

TECHNISCHE UNIVERSITÄT DARMSTADT

$$L = \sum_{s,a} p^{\pi}(s,a)\left(r(s,a) - \eta \log \frac{p^{\pi}(s,a)}{q(s,a)} - \theta\phi_s + \theta \sum_{s'} p(s'|s,a)\phi_{s'} - \lambda\right) + \eta\epsilon + \lambda$$

Derivative w.r.t $p^{\pi}(s,a)$

$$\begin{aligned}\frac{\partial}{\partial p^{\pi}(s,a)} - p^{\pi}(s,a)\eta \log \frac{p^{\pi}(s,a)}{q(s,a)} &= -\eta \log \frac{p^{\pi}(s,a)}{q(s,a)} - p^{\pi}(s,a)\eta \frac{q(s,a)}{p^{\pi}(s,a)} \frac{1}{q(s,a)} \\ &= -\eta \log \frac{p^{\pi}(s,a)}{q(s,a)} - \eta\end{aligned}$$

$$\frac{\partial}{\partial p^{\pi}(s,a)} L = r(s,a) - \eta \log \frac{p^{\pi}(s,a)}{q(s,a)} - \eta - \theta\phi_s + \sum_{s'} p(s'|s,a)\theta\phi_{s'} - \lambda$$

Derivation of REPS – Part II

TECHNISCHE UNIVERSITÄT DARMSTADT

$\theta\phi_s \Rightarrow V_s$

$$\frac{\partial}{\partial p^\pi(s,a)} L = -\eta \log \frac{p^\pi(s,a)}{q(s,a)} - \eta + r(s,a) - \lambda + \sum_{s'} p(s'|s,a) V_{s'} - V_s$$

Derivation of REPS – Part III

TECHNISCHE UNIVERSITÄT DARMSTADT

$r(s,a) - \lambda + \sum_{s'} p(s'|s,a)V_{s'} - V_s \Rightarrow \delta_\theta(s,a)$ (Bellman error)

$$
\begin{aligned}
\frac{\partial}{\partial p^\pi(s,a)} L = -\eta \log \frac{p^\pi(s,a)}{q(s,a)} - \eta - \lambda + \delta_\theta(s,a) &\stackrel{!}{=} 0 \\
-\eta - \lambda + \delta_\theta(s,a) &= \eta \log \frac{p^\pi(s,a)}{q(s,a)} \\
\exp \frac{-\eta - \lambda}{\eta} \exp \frac{\delta_\theta(s,a)}{\eta} &= \frac{p^\pi(s,a)}{q(s,a)} \qquad (1) \\
q(s,a) \exp \frac{-\eta - \lambda}{\eta} \exp \frac{\delta_\theta(s,a)}{\eta} &= p^\pi(s,a)
\end{aligned}
$$

Derivation of REPS – Part IV

Since $\sum_{s,a} p^{\pi}(s,a) = 1$

$$\sum_{s,a} q(s,a) \exp \frac{\delta_\theta(s,a)}{\eta} \exp \frac{-\eta - \lambda}{\eta} = 1$$

$$\exp \frac{-\eta - \lambda}{\eta} = \frac{1}{\sum_{s,a} q(s,a) \exp \frac{1}{\eta} \delta_\theta(s,a)} \tag{2}$$

$$p^{\pi}(s,a) = \frac{q(s,a) \exp \frac{1}{\eta} \delta_\theta(s,a)}{\sum_{s,a} q(s,a) \exp \frac{1}{\eta} \delta_\theta(s,a)}$$

Derivation of REPS – Part V

TECHNISCHE UNIVERSITÄT DARMSTADT

Using $p^\pi(s,a) = \mu^\pi(s)\pi(a|s)$ and $\mu^\pi(s) = \sum_a p^\pi(s,a)$

$$\mu^\pi(s)\pi(a|s) = \frac{q(s,a)\exp\frac{1}{\eta}\delta_\theta(s,a)}{\sum_{s,b} q(s,b)\exp\frac{1}{\eta}\delta_\theta(s,b)}$$

$$\pi(a|s) = \frac{q(s,a)\exp\frac{1}{\eta}\delta_\theta(s,a)}{\sum_a p^\pi(s,a)\cdot\sum_{s,b} q(s,b)\exp\frac{1}{\eta}\delta_\theta(s,b)}$$

Optimal Policy

$$\pi(a|s) = \frac{q(s, a) \exp\left(\frac{1}{\eta}\delta_\theta(s, a)\right)}{\sum_b q(s, b) \exp\left(\frac{1}{\eta}\delta_\theta(s, b)\right)}$$

Bellman error: $\delta_\theta(s, a) = r(s, a) + \sum_{s'} p(s'|s, a) V_\theta(s') - V_\theta(s)$
Dual: $g(\eta, \theta) = \eta\epsilon + \eta \log \sum_{s,a} q(s, a) \exp \frac{\delta_\theta(s,a)}{\eta}$

Derivation of the Dual – Part I

Start from Lagrangian using the Bellman error

$$g(\eta, \lambda) = \sum_{s,a} p^{\pi}(s, a) \left(-\eta \log \frac{p^{\pi}(s, a)}{q(s, a)} + \delta_{\theta}(s, a) - \lambda \right) + \eta\epsilon + \lambda$$

Substituting $\frac{p^{\pi}(s,a)}{q(s,a)}$ with Equation (1)

$$\begin{aligned}
&-\eta \log \left(\exp \frac{-\eta - \lambda}{\eta} \exp \frac{\delta_{\theta}(s, a)}{\eta} \right) \\
&-\eta \frac{-\eta - \lambda + \delta_{\theta}(s, a)}{\eta}
\end{aligned}$$

$$g(\eta, \lambda) = \sum_{s,a} p^{\pi}(s, a) \left(\eta + \lambda - \delta_{\theta}(s, a) + \delta_{\theta}(s, a) - \lambda \right) + \eta\epsilon + \lambda$$

Derivation of the Dual – Part II

$$g(\eta, \lambda) = \eta \sum_{s,a} p^{\pi}(s, a) + \eta\epsilon + \lambda = \eta + \eta\epsilon + \lambda = \eta\epsilon \cdot \eta \log \exp \frac{\eta + \lambda}{\eta}$$

Substituting $\exp(\frac{\eta+\lambda}{\eta})$ with Equation (2)

$$g(\eta, \theta) = \eta\epsilon + \eta \log \sum_{s,a} q(s, a) \exp \frac{\delta_\theta(s, a)}{\eta}$$

Solving the Dual

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\min_{\eta,\theta} \eta\epsilon + \eta \log \sum_{s_i, a_i} \frac{1}{N} \exp \frac{\delta_\theta(s_i, a_i)}{\eta}$$

- Draw samples from current policy
- Evaluate policy for η and θ by solving the dual
  - Using the samples from this or more iterations
- Compute new policy
- Repeat until convergence

Policy Improvement

$$q(s, a) = \mu^{\pi_l}(s)\pi_l(a|s)$$

$$\pi_{l+1}(a|s) = \frac{\pi_l(a|s)\exp\left(\frac{1}{\eta}\delta_\theta(s,a)\right)}{\sum_b \pi_l(b|s)\exp\left(\frac{1}{\eta}\delta_\theta(s,b)\right)}$$

References

Deisenroth, M. P., Neumann, G., and Peters, J. (2013).
A survey on policy search for robotics.
Foundations and Trends in Robotics, pages 388–403.

Peters, J., Muelling, K., and Altun, Y. (2010).
Relative entropy policy search.
In *Proceedings of the Twenty-Fourth National Conference on Artificial Intelligence (AAAI), Physically Grounded AI Track*.

The End

Thank you :-)

Discussion

Any questions?

Discussion

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Any questions?

nils.moehrle@stud.tu-darmstadt.de