Today’s topics

- Natural gradient
- Compatible function approximation
- Natural actor-critic (NAC)
- Biases, stochastic approximation, test experiments
Last time: Policy gradient

Average reward give a (parametric) policy:

\[
\rho_{Q, \pi_\omega} = \sum_{x, a} \mu^{\pi_\omega}(x) \pi_\omega(a|x) Q^{\pi_\omega}(x, a)
\]

In order to realise the policy gradient

\[
\omega_{t+1} = \omega_t + \beta_t \nabla_\omega \rho_\omega
\]

we assume that the dependency of \( \mu \) and \( Q \) on \( \omega \) to be “weak”, i.e. use a simplifying assumption for the dependency of \( \mu \) and \( Q \) on \( \omega \), namely

\[
\nabla_\omega \rho(\omega) = \sum_{x, a} \mu^{\pi}(x) \{ \nabla_\omega \pi_\omega(a|x) \} Q^{\pi}(x, a)
\]

Many versions of the algorithm possible (REINFORCE)
Algorithm (SARSA/$Q$):

- Initialise $x$ and $\omega$, sample $a \sim \pi_\omega (\cdot|x)$
- Iterate:
  - obtain reward $r$, transition to new state $x'$
  - new action $a' \sim \pi_\omega (\cdot|x')$
  - $\delta = r + \gamma Q_\theta (x', a') - Q_\theta (x, a)$
  - $\omega = \omega + \beta \nabla_\omega \log \pi_\omega (a|x) Q_\theta (x, a)$
  - $\theta = \theta + \alpha \delta \frac{\partial Q_\theta}{\partial \theta}$
  - $a \leftarrow a'$, $x \leftarrow x'$

- Until termination criterion
Actor-critic algorithms maintain two sets of parameters \((\theta, \omega)\), one \((\theta)\) for the representation of the value function and one \((\omega)\) for the representation of the policy.

Policy gradient methods are realised via stochastic descent using the current estimate of the value function.

Simultaneously, the estimate of the value function is gradually improved.

It is a suggestive idea to harmonise the two aspects of the optimisation process.
The approximation of the policy gradient introduces bias and variance. We need to be careful with the choice of the function approximation for $Q$.

For compatibility of the representations of value function and policy, aim at

$$\nabla_\theta Q_\theta \sim \nabla_\omega \log \pi_\omega$$

In order to check whether the approximation $\hat{Q}^\pi (x, a; \theta)$ of the true value function $Q^\pi (x, a)$ is good enough for calculating $\rho$ consider the following error measure:

$$\epsilon^\pi (\theta) = \sum_{x,a} \mu^\pi (x) \left( \hat{Q}^\pi (x, a; \theta) - Q^\pi (x, a) \right)^2 \pi_\omega (a|x)$$

We want to show now that using the best (w.r.t. $\theta$) approximation $\hat{Q}^\pi (x, a; \theta)$ leaves the gradient of $\rho$ (w.r.t to $\omega$) unchanged.

Compatible function approximation

Idea: Use score function

\[ \Psi_i(x, a)^\pi = \frac{\partial}{\partial \omega_i} \log \pi_\omega(a|x) \]

as basis functions for the approximation of the state-action value function in terms of \( \Psi \)

\[ \hat{Q}^\pi(x, a; \theta) = \sum_i \theta_i \Psi_i^\pi(x, a) \]

This implies \( \nabla_\theta Q_\theta = \nabla_\omega \log \pi_\omega \).

Using a score function as basis function is often possible, but may not always be a good choice (e.g. Gaussians \( \pi_\omega \) give linear \( \Psi \))
Consequences of the compatible function approximation

For $\hat{Q}^\pi (x, a; \theta) = \theta \Psi^\pi (x, a)$, minimisation of $\epsilon$ implies $\frac{\partial \epsilon}{\partial \theta} = 0$:

$$\frac{\partial}{\partial \theta_i} \epsilon^\pi(\theta) = \sum_{x, a} \mu^\pi(x) \Psi_i(x, a)^\pi \left( \hat{Q}^\pi (x, a; \theta) - Q^\pi (x, a) \right) \pi_\omega (a|x) = 0$$

or equivalently (this is what we wanted to show!)

$$\sum_{x, a} \mu^\pi(x) \Psi_i(x, a)^\pi \hat{Q}^\pi (x, a; \theta) \pi_\omega (a|x) = \sum_{x, a} \mu^\pi(x) \Psi_i(x, a)^\pi Q^\pi (x, a) \pi_\omega (a|x)$$

and in vector form using the basis functions for $\hat{Q}^\pi = \theta^\top \Psi(x, a)^\pi$

$$\sum_{x, a} \mu^\pi(x) \Psi(x, a)^\pi \theta^\top \Psi(x, a)^\pi \pi_\omega (a|x) = \sum_{x, a} \mu^\pi(x) \Psi(x, a)^\pi Q^\pi (x, a) \pi_\omega (a|x)$$
Consequences of the compatible function approximation

\[
\sum_{x,a} \mu(\pi(x)) \Psi(x, a) \pi(x, a)^\top \theta \pi_{\omega}(a|x) = \sum_{x,a} \mu(\pi(x)) \Psi(x, a) \pi Q_{\pi}(x, a) \pi_{\omega}(a|x)
\]

By definition \(\nabla_{\omega} \pi = \pi \Psi_i(x, a)^\pi\) because \(\Psi_i(x, a)^\pi = \frac{\partial}{\partial \omega_i} \log \pi_{\omega}(a|x)\)

\[
\left( \sum_{x,a} \mu(\pi(x)) \Psi(x, a) \pi(x, a)^\top \pi_{\omega}(a|x) \right) \theta = \sum_{x,a} \mu(\pi(x)) Q_{\pi}(x, a) \nabla_{\omega} \pi_{\omega}(a|x)
\]

\[
= \nabla_{\omega} \rho(\omega)
\]

Compare left hand side and the Fish information matrix:

\[
F_{ij}(\omega) = \mathbb{E}_{\mu(\pi(x))} \left[ \mathbb{E}_{\pi_{\omega}(a|x)} \left[ \frac{\partial \log \pi_{\omega}(a|x)}{\partial \omega_i} \frac{\partial \log \pi_{\omega}(a|x)}{\partial \omega_j} \right] \right]
\]

\[
= \sum_{x,a} \mu(\pi(x)) \pi_{\omega}(a|x) \frac{\partial \log \pi_{\omega}(a|x)}{\partial \omega_i} \frac{\partial \log \pi_{\omega}(a|x)}{\partial \omega_j}
\]

\[
F(\omega) = \sum_{x,a} \mu(\pi(x)) \Psi(x, a)^\pi \Psi^\top(x, a)^\pi \pi_{\omega}(a|x) \Rightarrow \theta = F^{-1}(\omega) \nabla_{\omega} \rho(\omega)
\]
1. Given the current policy $\pi$ we determine the score function $\Psi$.

2. Using $\pi$ we also get a sample of rewards which we can use to estimate the value function $\hat{Q}$.

3. At the same time we estimate the probability $\hat{\mu}$ of the agent in the state space.

4. From $\mu, \pi, \Psi, Q$ we can now find the optimal parameters $\theta$ by solving a (linear) equation.

5. $\theta$ is used in order to update the parameters of the policy ($\beta$ learning rate).

$$\omega_{t+1} = \omega_t + \beta_t \theta_t$$

This makes sense because we have seen that $\theta = F^{-1} \nabla_\omega \rho(\omega)$ which is a natural gradient on $\rho$.

What is a natural gradient?
The gradient is orthogonal to the level lines of the cost function. For a circular problem it points towards the optimum, while, for non-circular problem, we might be able to do better.

The natural gradient can be interpreted as a removal of the adverse effects of the particular model: In the above example we could simply “divide by the eigenvalues”, i.e. apply a linear transformation with the inverse eigenvalues and appropriate eigenvectors.

27/02/2015 Michael Herrmann RL 12
Gradient decent

\[ \theta_{t+1} = \theta_t - \eta \nabla_\theta f (\theta_t) \]

Assume a affine (linear) transformation \( \varphi = W^{-1} \theta \), so we have

\[ \varphi_{t+1} = \varphi_t - \eta \left( \frac{\partial \theta}{\partial \varphi} \right) \nabla_\theta f (\theta_t) = \varphi_t - \eta W^\top \nabla_\theta f (\theta_t) \]

Multiply by \( W \)

\[ W\varphi_{t+1} = W\varphi_t - \eta WW^\top \nabla_\theta f (\theta_t) \]

\[ \theta'_{t+1} = \theta_t - \eta WW^\top \nabla_\theta f (\theta_t) \]

In general \( \theta'_{t+1} \neq \theta_{t+1} \implies \) Gradient is not affine invariant.

This is nothing to worry about: The gradient works reasonably well with any positive definite matrix in front, but we can do better.
Beyond gradient decent

Gradient decent is based on a first-order Taylor expansion

\[ f(\theta) \approx f(\theta_0) + \nabla_\theta f(\theta_0)^\top (\theta - \theta_0) \]

Consider second-order Taylor expansion

\[ f(\theta) \approx f(\theta_0) + \nabla_\theta f(\theta_0) + \frac{1}{2} (\theta - \theta_0)^\top H(\theta_0) (\theta - \theta_0) \]

where the Hessian is given by \( H_{ij}(\theta_0) = \frac{\partial^2 f}{\partial \theta_i \partial \theta_j}(\theta_0) \). In this approximation we can optimise \( f \) w.r.t. \( \theta \) by

\[ \theta = \theta_0 - H^{-1}(\theta_0) \nabla_\theta f(\theta_0) \]

This is left unchanged by a linear transform \( \varphi = W^{-1} \theta \): \( H(\varphi_0) = W^\top H(\theta_0) W \) and \( \nabla \varphi \rightarrow W^\top \nabla_\theta \):

\[ \varphi = \varphi_0 - \left( W^\top H W \right)^{-1} W^\top \nabla_\theta f(W \varphi_0) = W^{-1} H^{-1}(\theta_0) \nabla_\theta f(W \varphi_0) \]

\[ \theta = \theta_0 - H^{-1}(\theta_0) \nabla_\theta f(\theta) \] (after multiplication by \( W \))

Second order method (Newton) is affine invariant.
Reformulation of gradient descent

Gradient descent improves the current estimate, perfect for a linear cost function in a specific coordinate system. Is it the best we can do (ignoring the 2nd order correction by the Hesse matrix)?

Given step size $\eta$, we find

$$
\theta^* = \arg \max_{\theta: \|\theta - \theta_0\| \leq \eta} f(\theta) \approx \arg \max_{\theta: \|\theta - \theta_0\| \leq \eta} f(\theta_0) + \nabla_\theta f(\theta_0)(\theta - \theta_0) \\
= \arg \max_{\theta: \|\theta - \theta_0\| \leq \eta} \nabla_\theta f(\theta_0)(\theta - \theta_0) \\
= \theta_0 + \eta \frac{\nabla_\theta f(\theta_0)}{\|\nabla_\theta f(\theta_0)\|}
$$

i.e. optimally $(\theta - \theta_0)$ has length $\eta$ and is parallel to the unit vector $\frac{\nabla_\theta f}{\|\nabla_\theta f\|}$, where $\|\cdot\|$ is the Euclidean norm.

Can we use also other norms (or distance functions)?
Kullback-Leibler divergence

\[ KL(\pi_1(a|x), \pi_2(a|x)) = \sum_{a,x} \pi_1(a|x) \log \frac{\pi_1(a|x)}{\pi_2(a|x)} \]

Consider two similar policies \( \pi_\omega(a|x) \) and \( \pi_{\omega+\delta\omega}(a|x) \). Perform a Taylor expansion of \( KL(\pi_\omega(a|x), \pi_{\omega+\delta\omega}(a|x)) \):

Constant term: \( KL(\pi_\omega(a|x), \pi_\omega(a|x)) = 0 \)

Linear term: \( \frac{\partial}{\partial \omega} KL(\pi_\omega(a|x), \pi_\omega(a|x)) = 0 \)

Quadratic term is the Fisher information matrix.
Fisher information

In other words, the Hessian for the Kullback-Leibler divergence is the Fisher information matrix.

\[ F_{ij}(x; \omega) = \mathbb{E}_{\pi_\omega(a|x)} \left[ \frac{\partial \log \pi_\omega(a|x)}{\partial \omega_i} \frac{\partial \log \pi_\omega(a|x)}{\partial \omega_j} \right] \]

Benefits

- As a Hessian the Fisher matrix gives an affine invariant descent.
- The approximation of the down/uphill direction becomes better for non-linear cost functions.
- Fisher information matrix is covariant (means: invariant against transformations of the parameter).

Literature:

- Natural gradient (S. Amari: Natural gradient works efficiently in learning, NC 10, 251-276, 1998)
- Examples by Bagnell and Schneider (2003) and Jan Peters (2003, 2008)
\[ \theta = F^{-1}(\omega) \nabla_\omega \rho(\omega) \]
is a natural gradient on
\[ \rho_{Q,\pi,\mu} = \sum_{x,a} \mu^{\pi \omega}(x) Q^{\pi \omega}(x, a) \pi_{\omega}(a|x) \]
if we can assume the dependency of \( \mu \) and \( Q \) on \( \omega \) to be “weak”, i.e.
\[ \nabla_\omega \rho(\omega) = \sum_{x,a} \mu^{\pi}(x) Q^{\pi}(x, a) \nabla_\omega \pi_{\omega}(a|x) \]
We have seen that
\[ F(\omega) \theta = \nabla_\omega \rho(\omega) \iff \theta = F(\omega)^{-1} \nabla_\omega \rho(\omega) =: \tilde{\nabla}_\omega \rho(\omega) \]
which defines the natural gradient. This implies the following natural gradient learning rule (Kakade, 2001/2)
\[ \omega_{t+1} = \omega_t + \beta_t \theta_t \]
which is better and simpler than standard policy gradient. Remember this result is obtained at the cost of the calculation of \( \theta \)!
Pros and Cons of the Fisher information

+ “Natural” (covariant): uses the geometry of the goal function rather than the geometry of the parameter space (Choice of parameters used to be critical, but isn’t any more so).
+ Related to Kullback-Leibler divergence and to Hessian
+ Describes efficiency in statistical estimation (Cramer-Rao)
+ Many applications in machine learning, statistics and physics
  - Depends usually on parameters and is computationally complex (but not here where we were lucky!)
  - Requires sampling of high-dimensional probability distributions
+ May still work if some approximation is used, e.g. Gaussian
Three right curves: standard gradient, three left curves: natural gradient

Policy $\pi(a|x; \omega) \sim \exp(\omega_1 s_1 x^2 + \omega_2 s_2 x)$

Starting conditions: $\omega_1 s_1 = \omega_2 s_2 = -0.8$
Kakade’s Example

Left: average reward for the policy
$$\pi (a = 1 \mid s; \omega) \sim \exp (\omega) / (1 + \exp (\omega))$$

Lower plot represents the beginning of the upper plot (different scales!): dashed: natural gradient, solid: standard gradient.

Right: Movement in the parameter space (axes are actually $\omega_i$!)

27/02/2015 Michael Herrmann RL 12
Examples of natural gradients

A promising approach for continuous action and state spaces (in discrete time)

Policy gradient as direct maximisation of the averaged state-action value

Natural policy gradient arises from the optimisation of the value function

Model-free reinforcement learning
Some material was adapted from web resources associated with Sutton and Barto’s Reinforcement Learning book.

Today mainly based on C. Szepesvári: *Algorithms for RL*, Ch. 3.4.

See also: David Silber’s Lecture 7: Policy Gradient and Pieter Abbeel’s lecture 20 on Advanced Robotics.

Check papers by Jan Peters and others (see above in the slides).