RL 11A: Natural Actor-Critic*

Michael Herrmann

University of Edinburgh, School of Informatics

26/02/2015

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:

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

In order to realise the policy gradient

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

we assume that the dependency of μ and $\mathcal{Q}$ on ω to be "weak", i.e. use a simplifying assumption for the dependency of μ and $\mathcal{Q}$ on ω, namely

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

Many versions of the algorithm possible (REINFORCE)

Actor-Critic Algorithm

Algorithm (SARSA/$\mathcal{Q}$):

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

Policy gradient

- Actor-critic algorithms maintain two sets of parameters (θ, ω), one (θ) for the representation of the value function and one (ω) 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

Recipe for Natural Actor-Critic

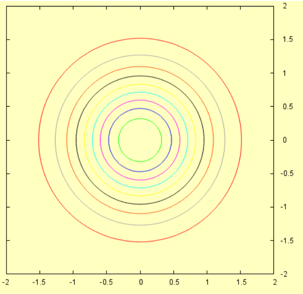
1. Given the current policy π we determine the score function Ψ.
2. Using π we also get a sample of rewards which we can use to estimate the value function $\hat{\mathcal{Q}}$.
3. At the same time we estimate the probability $\hat{\mu}$ of the agent in the state space.
4. From μ, π , Ψ, $\mathcal{Q}$ we can now find the optimal parameters θ by solving a (linear) equation.
5. θ is used in order to update the parameters of the policy (β learning rate).

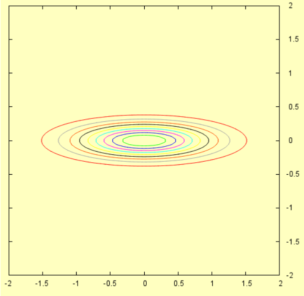
$$\omega_{t+1} = \omega_t + \beta_t \theta_t$$

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

What is a natural gradient?

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.

Gradient descent

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

$$\begin{aligned} W\varphi_{t+1} &= W\varphi_t - \eta W W^\top \nabla_\theta f(\theta_t) \\ \theta'_{t+1} &= \theta_t - \eta W W^\top \nabla_\theta f(\theta_t) \end{aligned}$$

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 $H_{ij}(\theta_0) = \frac{\partial^2 f}{\partial \theta_i \theta_j}(\theta_0)$. In this approximation we can optimise f w.r.t. θ 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 $\nabla_\varphi \to W^\top \nabla_\theta$:

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

$$\theta = \theta_0 - H^{-1}(\theta_0)\nabla_\theta f(\theta) \qquad \text{(after multiplication by } W\text{)}$$

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 η, we find

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

i.e. optimally $(\theta-\theta_0)$ has length η 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

Kullback-Leibler divergence

$$KL\left(\pi_{\omega_1}\left(a|x\right), \pi_{\omega_2}\left(a|x\right)\right) = \sum_{a,x} \pi_{\omega_1}\left(a|x\right) \log \frac{\pi_{\omega_1}\left(a|x\right)}{\pi_{\omega_2}\left(a|x\right)}$$

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

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

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

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 appropriate parameter transformations).

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)

Natural gradient

$\theta = F^{-1}(\omega)\,\nabla_\omega \rho(\omega)$ is a natural gradient on

$$\rho_{\mathcal{Q},\pi,\mu} = \sum_{x,a} \mu^{\pi_\omega}(x)\, \mathcal{Q}^{\pi_\omega}(x,a)\, \pi_\omega(a|x)$$

if we can assume the dependency of μ and $\mathcal{Q}$ on ω to be "weak", i.e.

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

We have seen that

$$F(\omega)\,\theta = \nabla_\omega \rho(\omega) \Leftrightarrow \theta = F(\omega)^{-1} \nabla_\omega \rho(\omega) =: \tilde{\nabla}_\omega \rho(\omega)$$

which defines the natural gradient. This implies the following natural gradient learning rule (Kakade, 2001/2)

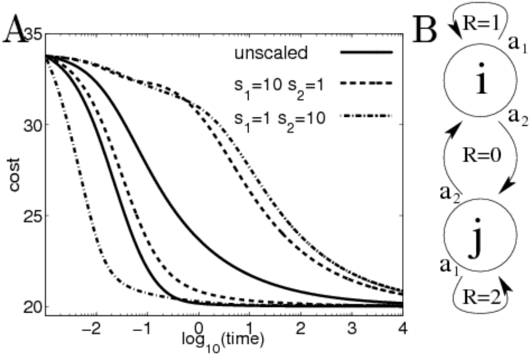
$$\omega_{t+1} = \omega_t + \beta_t \theta_t$$

which is better and simpler than standard policy gradient. Remember this result is obtained at the cost of the calculation of θ!

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 were we get it for free: We were lucky!)
− Requires sampling of high-dimensional probability distributions
+ May still work if some approximation is used, e.g. Gaussian

Kakade's Example

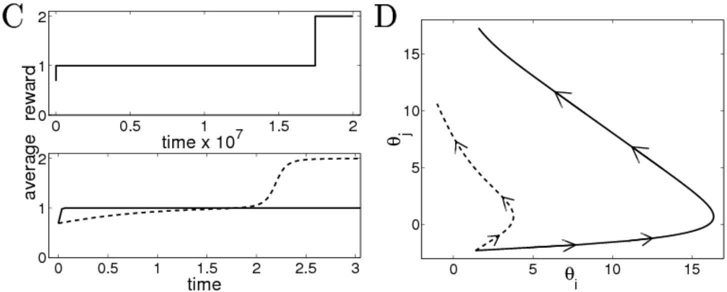


Three right curves: standard gradient, three left curves: natural gradient

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

Starting conditions: $\omega_1 s_1 = \omega_2 s_2 = -0.8$

Kakade's Example

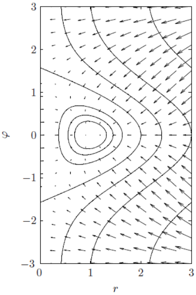


Left: average reward for the policy
$\pi\left(a=1|s;\omega\right) \sim \exp\left(\omega\right)/\left(1+\exp\left(\omega\right)\right)$

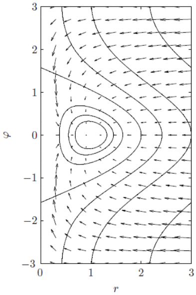
Lower plot represents the beginning of the upper plot (different scales!): dashed: natural gradient, solid: standard gradient.

Right: Movement in the parameter space (axes are actually ω_i!)

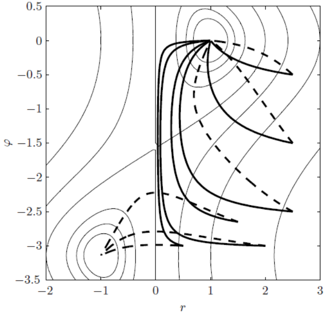
Examples of natural gradients



(a) Standard



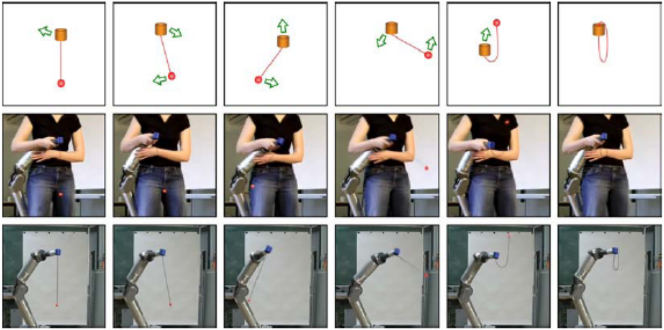
(b) Natural



standard gradient (dashed)
natural gradient (solid)

Grondman et al. (2012) A survey of actor-critic reinforcement learning: Standard and natural policy gradients. IEEE TA Systems, Man, and Cybernetics 42(6), 1291-1307.

More examples



J. Kober & J. R. Peters: Policy search for motor primitives in robotics. NIPS 2009, pp. 849-856.

Summary

- 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

Acknowledgements: See lecture 12.