687 2017-11-28

Policy gradient thm: $\frac{\partial J(\theta)}{\partial \theta} = \sum_s d^\pi(s) \sum_a \pi(s,a,\theta) q^\pi(s,a) \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}$

$q_w(s,a)$ is an approximation

let $q_w(s,a) = w^T \underbrace{\frac{\partial \ln \pi(s,a\,\theta)}{\partial \theta}}_{\text{called compatible features}}$: a linear approximator

called compatible features $\longrightarrow$ because using this q preserves equality of $\frac{\partial J(\theta)}{\partial \theta}$

If w^* is a critical point of $\rightarrow$ a place where $\frac{\partial}{\partial w}$ is 0

$$L(w) = \sum_s d^\pi(s) \sum_a \pi(s,a,\theta)\left(q^\pi(s,a) - w^T \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}\right)^2$$

Then $\frac{\partial J(\theta)}{\partial \theta} = \sum_s d^\pi(s) \sum_a \pi(s,a,\theta)\, q_{w^*}(s,a) \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}$

Policy gradient thm with approximation.

Proof: $\frac{\partial L(w)}{\partial w} = 0 = \sum_s d^\pi(s) \sum_a \pi(s,a,\theta)(-2)\left(q^\pi(s,a) - w^T \frac{\partial \ln \pi(s,a\,\theta)}{\partial \theta}\right)\left(\frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}\right)$

$$\Rightarrow \sum_s d^\pi(s) \sum_a \pi(s,a,\theta)\Big(\underbrace{w^T \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}}_{q_w(s,a)}\Big) \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta} = \underbrace{\sum_s d^\pi(s) \sum_a \pi(s,a,\theta) q^\pi(s,a) \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}}_{\text{is } \nabla J(\theta)}$$

by linear algebra

$$= w \sum_s d^\pi(s) \sum_a \pi(s,a,\theta) \underbrace{\frac{\partial \ln \pi(s,a,\theta)}{\partial \theta} \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}^T}$$

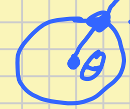
$\rightarrow$ If θ is n-dim, this is $n \times n$.

Natural (Policy) Gradients

- The gradient is not the direction of steepest ascent (from a point of view).
- Natural gradient is the "true" direction of steepest ascent.
    - Often easier to compute.
    - Often a better update direction.
    - It is a covariant learning rate.

Derive gradient (ignoring magnitude):

Direction of steepest ascent: $\nabla f(\theta)$

Imagine:  $\leftarrow \theta + \epsilon\Delta$, $\|\Delta\| = 1$

Let this be the best direction for infinitesimal ϵ.

$$\nabla f(\theta) = \lim_{\epsilon \to 0} \operatorname*{argmax}_{\Delta : \|\Delta\| = 1} f(\theta + \epsilon\Delta)$$

Taylor expansion of $f(\theta + a)$ centered at $f(\theta)$:

$$f(\theta + a) = f(\theta) + a^T \underbrace{\frac{\partial f(\theta)}{\partial \theta}}_{\text{Jacobian}} + \frac{1}{2} a^T \underbrace{\frac{\partial^2 f(\theta)}{\partial \theta^2}}_{\text{Hessian}} + \dots$$

$$\lim_{\epsilon \to 0} \operatorname*{argmax}_{\Delta : \|\Delta\| = 1} f(\theta) + \epsilon \Delta^T \frac{\partial f(\theta)}{\partial \theta} + \epsilon^2 \Delta^T \frac{\partial^2 f(\theta)}{\partial \theta^2} + \dots$$

$$= \lim_{\epsilon \to 0} \operatorname*{argmax}_{\Delta : \|\Delta\| = 1} \epsilon \Delta^T \frac{\partial f(\theta)}{\partial \theta} + \dots$$

- can drop higher order terms
- $f(\theta)$ does not affect optimal Δ

$$= \operatorname*{argmax}_{\Delta : \|\Delta\| = 1} \Delta^T \frac{\partial f(\theta)}{\partial \theta}$$

Lagrange Multipliers:

Optimize $h(x)$ s.t. $g(x) = 0$.

All solutions must be critical pts of

$$\mathcal{L}(x, \lambda) = h(x) - \lambda g(x) \qquad \Delta^T\Delta$$

$$\mathcal{L}(\Delta, \lambda) = \Delta^T \frac{\partial f(\theta)}{\partial \theta} - \lambda(\|\Delta\| - 1)$$

$$\frac{\partial \mathcal{L}(\Delta, \lambda)}{\partial \Delta} = 0 = \frac{\partial f(\theta)}{\partial \theta} - \lambda 2\Delta$$

$$\Delta = \frac{1}{2\lambda} \frac{\partial f(\theta)}{\partial \theta},$$ So direction is $\frac{\partial f(\theta)}{\partial \theta}$ ... the ordinary gradient.

But: we used $\sqrt{\Delta^T \Delta}$ as $\|\Delta\|$. That is, we assumed the inputs (θ) lie in Euclidean space!

Example of why this can be bad:

Goal: Fit a normal distribution to observed points, maximizing likelihood.

$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$ $\quad f(\theta) = L(\theta / D) = \Pr(D \mid \mu, \sigma^2)$

$\Downarrow$

$$\theta \leftarrow \theta + \alpha \frac{\partial L(\theta)}{\partial \theta}$$

$$\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \leftarrow \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} + \alpha \frac{\partial L(\mu, \sigma^2)}{\partial (\mu, \sigma^2)}$$

but $\|\cdot\|$ we used is Euclidean

distance $= \sqrt{\Delta\mu^2 + (\Delta\sigma^2)^2}$

What if I used $\begin{bmatrix} \mu \\ \sigma \end{bmatrix}$? Would get $\sqrt{\Delta\mu^2 + \Delta\sigma^2}$.

Any $\begin{bmatrix} \mu^k \\ \sigma^{k'} \end{bmatrix}$ should work, but $\|\cdot\|$ gives different behavior in terms of movement through the space.

→ Showed evolution of estimate for various k of $\begin{bmatrix} \mu \\ \sigma^k \end{bmatrix}$. Quite varied. can be circuitous & can even not reach the target.

Want to allow choice of distance measure $d(\theta, \theta + \Delta)$.

Will use: $\|\Delta\|_G = \sqrt{\Delta^T G \Delta}$

where G is a $|\theta| \times |\theta|$ (square) positive definite matrix.

$G = I \Rightarrow$ Euclidean distance

G leads to stretched/rotated hypersphere.

Can allow G to depend on θ: $\|\Delta\|_{G(\theta)} = \sqrt{\Delta^T G(\theta) \Delta}$