policy gradients

ARTICLE — Communicated by Steven Nowlan and Erkki Oja

Natural Gradient Works Efficiently in Learning

Shun-ichi Amari
RIKEN Frontier Research Program, Saitama 351-01, Japan

A Natural Policy Gradient

Sham Kakade
Gatsby Computational Neuroscience Unit
17 Queen Square, London, UK WC1N 3AR
http://www.gatsby.ucl.ac.uk
sham@gatsby.ucl.ac.uk

Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator

Maryam Fazel [*1] Rong Ge [*2] Sham M. Kakade [*1] Mehran Mesbahi [*1]

- consider optimization problem $\min_{u} c(u)$
  - this could have come from an MDP / SOCP,
    eg $c(u) = E\left[\sum_{t=0}^{T} \mathcal{L}_t(x_t, u_t) \mid u_t \sim P_u(x),\ x_{t+1} \sim P_x(x_t, u_t) \right]$
    or a trajectory optimization problem,
    eg $c(u) = \sum_{t=0}^{T} \mathcal{L}_t(x_t, u_t)$ s.t. $x_t^+ = F(x_t, u_t)$
  - → but it doesn't really matter! methods below are agnostic...

policy gradient

PG: $u^+ = u - \gamma Dc^T(u)$

let $w = Lu$ so $w^+ = Lu^+ = Lu - \gamma L Dc^T(u)$
$= w - \gamma L Dc^T(L^{-1} w)$

→ solve $\min_{v \in \mathbb{R}^m} \langle Dc(u), v \rangle$ ← find the steepest / most rapid descent direction
s.t. $\|v\|_2 \le \|Dc(u)\|_2$

– recalling that $\langle x, y \rangle = \|x\|_2 \|y\|_2 \cos\theta$,

θ = angle between x & y

$\Rightarrow$ solution is $v^* = -Dc(u)^T \in \mathbb{R}^m$

$\rightarrow$ $\min_v \underbrace{c(x) + Dc(x) \cdot v + \frac{1}{2} v^T D^2 c(x) v}$

$D_v[\cdot] = Dc(x) + v^T D^2 c(x)$

$= 0 \iff v = -\left[D^2 c(x)\right]^{-1} Dc(x)^T$

$\rightarrow$ $\min_v g \cdot v$ s.t. $\|v\|_H \leq \|g\|_H$, $\|x\|_H = \sqrt{\frac{1}{2} x^T H x}$

$\tilde{c} = g \cdot v + \frac{\lambda}{2}\left(v^T H v - g H g^T\right)$

$D_v \tilde{c} = g + \lambda v^T H = 0 \iff v = -\lambda H^{-1} g$

* solve for λ from $D_\lambda \tilde{c} = 0$

"natural" policy gradient

NPG: $u^+ = u - \gamma\left(D^2 \tilde{c}(u)\right)^{-1} D\tilde{c}^T(u)$ ie Newton – Raphson

let $w = Lu$ so $w^+ = w - \gamma L\left(D^2 c(L^{-1} w)\right)^{-1} D\tilde{c}^T(L^{-1} w)$

ex: $c(u) = \frac{1}{2}(u - u^*)^T H (u - u^*) \Rightarrow Dc(u) = (u - u^*)^T H$, $D^2 c = H$

PG: $u^+ = u - \gamma H (u - u^*) \iff w^+ = w - \gamma L H (L^{-1} w - u^*)$

NPG: $u^+ = u - \gamma H^{-1} H (u - u^*) \iff w^+ = w - \gamma L H^{-1} H (L^{-1} w - u^*)$

$$\underbrace{= u - \gamma (u - u^*) \qquad\qquad = w - \gamma (w - u^*)}_{\text{same descent regardless of coordinates } L}$$

ex: LQR $\quad x^+ = Ax + Bu$, $x_0 \sim D$, $u_t = -K x_t$

- cost $c(K) = E\left[\sum_{t=0}^{\infty} x_t^T Q x_t + u_t^T R u_t\right]$

prop: $Dc(K) = 2\left((R + B^T P_K B)K - B^T P_K A\right)\Sigma_K$

where $P_K = Q + K^T R K + (A-BK)^T P_K (A-BK)$ — Riccati equation

$\Sigma_K = E\left[\sum_{t=0}^{\infty} x_t x_t^T\right]$ — state variance

pf: $c(K; x_0) = E\left[x_0^T P_K x_0\right]$

$= E\left[x_0^T (Q + K^T R K) x_0 + x_0^T (A-BK)^T P_K (A-BK) x_0\right]$

$= E\left[x_0^T (Q + K^T R K) x_0 + c(K; (A-BK)x_0)\right]$

$*\ D_M x^T f(M) x = Df(M) x x^T, \quad D_M M^T N M = NM$

$\Rightarrow D_K c(K, x_0) = E\left[\underbrace{2RKx_0x_0^T - 2B^T P_K (A-BK) x_0 x_0^T}\, + D_K c(K; \underbrace{(A-BK)x_0}_{=x_1})\right]$

$D_K c(K; (A-BK)x_0) = (\text{w/ } x_0 \mapsto (A-BK)x_0) + D_K c(K; (A-BK)^2 x_0)$

$\Rightarrow = E\left[2 \cdot \left((R + B^T P_K B)K - B^T P_K A\right) \underbrace{\sum_{t=0}^{\infty} x_t x_t^T}_{=: \Sigma_K}\right]$

- policy gradient $K^+ = K - \gamma Dc(K)$
- natural policy gradient $K^+ = K - \gamma Dc(K) \Sigma_K^{-1}$