POLICY GRADIENT METHODS

STOCHASTIC GRADIENT ASCENT ON SURFACE INDUCED BY (SMOOTH) POLICY CLASS $\Pi = (\pi_w, w \in \mathbb{R}^d)$, STATIONARY, w PARAMETERS

- **GIBBS POLICY:** $\pi_w(a|x) = \dfrac{\exp(w^T \xi(x,a))}{\sum_A \exp(w^T \xi(x,a'))}$ ξ: FEATURE EXTRACTION FCN, x STATE, a ACTION
- **GAUSSIAN (MVN) POLICY:** ...
- **PROBLEM:** $\arg\max_w \rho_w$, ρ PERFORMANCE, IE EXPECTED RETURNS OF π_w

POLICY GRADIENT THEOREM

- ASSUME MARKOV CHAIN OF π_w ERGODIC $\forall w$. HOW TO GRADIENT?
- **SCORE FUNCTION** $\psi_w(x,a) = \dfrac{\partial}{\partial w} \log \pi_w(a|x)$, **EXAMPLE:** FOR GIBBS POLICY $\psi(x,a) = \xi(x,a) - \sum_A \pi(a'|x)\xi(x,a')$
- $G(w) = (Q^{\pi_w}(x,a) - h(x))\,\psi(x,a)$, h IS ANY BOUNDED FCN, Q^{π_w} SAMPLE FROM A-V FCN OF π_w
  (ALSO CALLED e)
  || ALT NOTATION: $w_{t+1} = w_t + \alpha\left[V_\pi(s_t) - \hat{V}_\pi(s_t, w_t)\right] \nabla \hat{v}(s_t, w_t)$

$\longrightarrow$ G IS UNBIASED ESTIMATOR OF GRADIENT $\nabla_w \rho_w = E[G(w)]$

UPDATE RULE: $w_{t+1} = w_t + \beta \hat{G}_t$, • DOES SGA AS LONG AS $E\left[\hat{Q}_t(x_t,a_t)\,\psi_{w_t}(x_t,a_t)\right] = E\left[Q^{\pi_{w_t}}(x,A)\,\psi_{w_t}(x_t,a_t)\right]$

- h IS FOR VARIANCE REDUCTION, SPEEDS UP CONVERGENCE, IE USE $V^{\pi_{w_t}}$, THE STATE VALUE FCN ITSELF
- **DIFFICULT** TO CONSTRUCT GOOD $\hat{Q}_t$ $\longrightarrow$ **REINFORCE** DOES UPDATES AT END OF EPISODES. DIRECT POLICY SEARCH (NO VALUE FCN)
- **NON-EPISODIC TASKS!** $\hat{Q}_t$ ON FASTER TIMESCALE, POLICY PARAMS ON SLOWER
  - **COMPATIBLE FCN APPROXIMATION:** $\hat{Q}_t$ LINEAR IN PARAMS, ξ IS SCORE FCN FOR POLICY CLASS $Q_\theta(x,a) = \theta^T \psi(x,a)$ COMPATIBLE BECAUSE CAN SOLVE FOR θ, $F_w \theta = g_w$
    θ ON FAST SCALE, IE SARSA; w_{t+1} ON SLOWER SCALE
  - **NATURAL ACTOR-CRITIC:** • $w_{t+1} = w_t + \beta_t \theta_t$, ELSE SAME AS COMPATIBLE FCN APPROX, SAME CONVERGENCE | USE LSTD-Q(λ) FOR $\theta^*(w)$
    - **NATURAL GRADIENT:** $\theta_x(w)$ IS NG OF ρ_w: DOES GRAD ASCENT DIRECTLY IN METRIC SPACE UNDERLYING OBJECTS OF INTEREST, SPACE OF STOCHASTIC POLICIES VS DOING GRAD ASCENT IN METRIC SPACE OF PARAMS
    - TRAJECTORIES OF $\dot{w} = \theta_x(w)$ ARE INVARIANT TO SMOOTH EQUIVALENT REPARAMETRIZATIONS OF POLICY CLASS
    - WE THINK NATURAL GRADIENTS ARE NICE AND LEAD TO FASTER CONVERGENCE

- **VAPS FORMULATION**

  1ST GENERAL FORMULATION OF POLICY GRADIENTS. MODEL-FREE, ACTION INDEPENDENT + ANCIENT (1999)
  CAN SEARCH FOR VALUE FCNS OR EXPLICIT POLICIES $\Delta w: -\alpha\left[\frac{\partial}{\partial w} e(s_t) + e(s_t) \cdot T_t\right]$ e: ANY FCN OF W, PREVIOUS STATES, ACTIONS, REINFORCEMENTS.
  - STOCHASTIC POLICIES FCN OF W

  T: TRACE $\frac{\partial}{\partial w} \ln(P(u_{t-1}|s_{t-1}))$ EG: e_{SARSA}: $\frac{1}{2}\left[E^2\left[R_{t-1} + \gamma Q(x_t,u_t) - Q(x_{t-1}, u_{t-1})\right]\right]$

- **PG W/ FUNCTION APPROXIMATION**

  IDEA: DIRECTLY APPROXIMATE STOCHASTIC POLICY VIA INDEPENDENT FCN OF W PARAMS; IE A NEURAL NET <u>INPUT</u>: STATE, <u>OUTPUT</u>: ACTION SELECTION <u>PROBS</u> PARAMS: WEIGHTS

  $\Delta\theta = \alpha \dfrac{\partial \rho}{\partial \theta}$, ρ PERFORMANCE (LONG-TERM REWARD → AVERAGE ~~SUM OF DISCOUNTED~~), • SMALL CHANGES IN θ → SMALL CHANGES IN POLICY / VISITATION DISTRIBUTION

  CONDITION: $\sum_s d^\pi(s) \sum_a \pi(s,a)\left[Q^\pi(s,a) - f_w(s,a)\right] \dfrac{\partial f_w(s,a)}{\partial w} = 0$ **AT EQUILIBRIUM/OPTIMUM**, $d^\pi(s)$ STATIONARY DISTRIBUTION OF STATES UNDER π; Q^π A/V FUNCTION SUM OF DISCOUNTED REWARDS GIVEN STATE S ACTION A

  THEN $\rightarrow$ $\dfrac{\partial \rho}{\partial \theta} \approx \sum_s d^\pi(s) \sum_a \dfrac{\partial \pi(s,a)}{\partial \theta} f_w(s,a)$ **EXAMPLE (GIBBS POLICY):** $\pi(s,a) = \dfrac{e^{\theta^T \phi(s,a)}}{\sum_b e^{\theta^T \phi(s,b)}}$, $\dfrac{\partial f_w(s,a)}{\partial \theta} = \phi_{sa} - \sum_b \pi(s,b)\phi_{sb}$

  $f_w(s,a) = w^T\left[\phi_{sa} - \sum \pi(s,b)\phi_{sb}\right]$ → f_w MUST BE LINEAR IN SAME FEATURES AS POLICY

- **REINFORCE** 1ST POLICY GRADIENT EVER, ALREADY NETWORKS

  $\Delta w_{ij} = \alpha_{ij}(R - b_{ij})\, e_{ij}$, $e_{ij} = \dfrac{\partial \ln g_i}{\partial w_{ij}}$, g_i PDF ACTIVATIONS. PER-WEIGHT UPDATES

DETERMINISTIC POLICY GRADIENT

DETERMINISTIC POLICY: $a = \mu_\theta(s)$, GRADIENT ONLY INTEGRATES OVER STATE SPACE, ~~OF~~ USED OFF-POLICY OTHERWISE NO EXPLORATION, ACTOR-CRITIC.

RL PERFORMANCE: $J(\pi_\theta) = \int_S \rho^\pi(s) \int_A \pi_\theta(s,a) R(s,a)\, da\, ds = \underset{s\sim\rho^\pi, a\sim\pi_\theta}{E}[R(s,a)]$. ρ IMPROPER DISCOUNTED STATE DISTRIBUTION

- USUAL PG THEOREM: $\nabla_\theta J(\pi_\theta) = \underset{s\sim\rho^\pi\, a\sim\pi_\theta}{E}[\nabla_\theta \log \pi(a|s) Q^\pi(s,a)] \longrightarrow \nabla_\theta J$ INDEPENDENT OF STATE DISTRIBUTION !!!

  - Q^x (FUNCTION ESTIMATOR) NEEDS TO BE 'COMPATIBLE' WITH Q^π (TRUE AV FCN)
    1- LINEAR IN FEATS OF STOCHASTIC POLICY.
    2- W PARAMS ARE OLS WEIGHTS $Q^x \to Q^\pi$
    └ USUALLY REPLACED WITH T-D UPDATES

- OFF-POLICY GRADIENT $\nabla_\theta J_\beta(\pi_\theta) = \underset{s\sim\rho^\beta\, a\sim\beta}{E}\left[\underbrace{\frac{\pi_\theta(a|s)}{\beta_\theta(a|s)}}_{\text{IMPORTANCE-SAMPLING RATIO}} \nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a)\right]$. β BEHAVIOR POLICY

- DETERMINISTIC POLICY GRADIENTS

$$\theta^{k+1} = \theta^k + \alpha \underset{s\sim\rho^{\mu^k}}{E}\Big[\underbrace{\nabla_\theta \mu_\theta(s)}_{\text{POLICY WRT PARAMS}} \underbrace{\nabla_a Q^{\mu^k}(s,a)}_{\text{ACTION-VALUE WRT ACTIONS}}\Big|_{a=\mu_\theta(s)}\Big]$$

OBJ
$$J(\mu_\theta) = \int \rho^\mu(s) R(s, \mu_\theta(s))\, ds = \underset{s\sim\rho^\mu}{E}[R(s,\mu_\theta(s))].$$

GRAD
$$\nabla_\theta J_\theta(\mu) = \underset{s\sim\rho^\mu}{E}\Big[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)\Big|_{a=\mu(s)}\Big]$$

- IS LIMIT FOR NOISE $\sigma \to 0$ OF STOCHASTIC POLICY GRADIENT

- ON-POLICY A/C DPG ALGO:

SARSA CRITIC
$$\begin{cases} \delta_t = R_t + \gamma Q^w(s_{t+1}, a_{t+1}) - Q^w(s_t, a_t) \\ w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \\ \theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu_\theta(s) \nabla_a Q^w(s_t, a_t)\big|_{a=\mu(s)} \end{cases}$$

- OFF-POLICY A/C DPG ALGO: BEHAVIOR IS STOCHASTIC $\pi(s,a)$ OPDAC
  → LIKE ONPOLICY BUT Q-LEARNING CRITIC $\delta_t = R_t + \gamma Q^w(s_{t+1}, \mu_\theta(s_{t+1})) - Q_w(s_t, a_t)$
  - DPG REMOVES IMPORTANCE SAMPLING ON ACTOR, NO INTEGRAL OVER ACTION
  - Q " " " " CRITIC

COMPATIBLE APPROXIMATIONS $\nabla_a Q(s,a) = \nabla_\theta \mu_\theta(s)^T w$ AND MINIMIZES MSE → SEE IT AS REGRESSION PROBLEM WITH FEATURES

→ COPDAC
$$\begin{cases} w_{t+1} = \alpha_w \delta_t \phi(s_t, a_t) + w_t \\ v_{t+1} = v_t + \alpha_v \delta_t \phi(s_t) \end{cases}$$

- CRITIC IS LINEAR FCN $\phi(s,a) = a^T \nabla_\theta \mu_\theta(s)$
  LINEAR NET

→ COPDAC - GQ USES TD-UPDATES

→ NATURAL GRADIENT ALSO MEANINGFUL FOR DETERMINISTIC POLICIES $\theta_{t+1} = \theta_t + \alpha_\theta w_t$

STOCHASTIC VALUE GRADIENT

VALUE GRADIENT: IS POLICY GRADIENT VIA BACKPROPAGATION, DIFFERENTIABLE MODELS

DETERMINISTIC VG: $V_s = R_s + R_a \pi_s + \gamma V'_{s'} (f_s + f_a \pi_s)$; $a = \pi(s, \theta)$ POLICY ; $s' = f(s, a)$ MODEL

θ NET PARAMS $V_\theta = R_a \pi_\theta + \gamma V' f_a \pi_\theta + \gamma V'_\theta \theta$

- USE REPARAMETRIZATION TRICK ON DET-VG: BELLMAN EQUATION

ALGORITHMS:

- **SVG (∞)** VG VIA BACKWARD RECURSIONS ON FINITE TRAJECTORIES. END-OF-EPISODE TRAIN MODEL $\hat{f}$ AND POLICY π. ON-POLICY.
- **SVG (1)** OFF-POLICY. USES EXPERIENCE REPLAY. DERIVATIVE OF CRITIC WRT STATES IS USED FOR UPDATES, INSTEAD OF SAMPLE GRADIENT
- **SVG (0)** IS STOCHASTIC ANALOGUE OF DPG, ESTIMATES DERIVATIVE AROUND POLICY NOISE

- JOINT TRAINING OF MODEL AND POLICY