Action-Value Approximation - on policy

- CANNOT INTO FULL TABULAR VALUE FCNS FOR TASKS WITH LARGE STATE AND ACTION SPACES, OR CONTINUOUS SPACES.
- IDEA: GENERALIZE FROM KNOWN, AVAILABLE VALUES ⟶ FUNCTION APPROXIMATION, REGRESSION, SUPERVISED LEARNING **ANY SUPERVISED LEARNING IS ✓**
- CAN USE ANNS OR OTHER TECHS ⟷ BACKUPS AREN'T TRIVIAL SHIFTS ANYMORE BUT ARBITRARILY COMPLEX
- IDEA: USE $S \mapsto V$ BACKUP PAIRS AS SUPERVISED TRAINING INSTANCES, OUTPUT IS ESTIMATED VALUE FUNCTION
- SUITABLE SV ARE THOSE CONVENIENTLY ALLOWING ONLINE LEARNING OF NONSTATIONARY STUFF.
- **PERFORMANCE METRIC:** RMSE

$$RMSE(w)=\sqrt{\sum d(s)\left[v_\pi(s)-\hat{v}(s,w)\right]^2}.$$

d { IS DISTRIBUTION SPECIFYING RELATIVE IMPORTANCE OF ERROR IN DIFFERENT STATES
DISTRIBUTION WE DRAW TRAINING EXAMPLES FROM
DISTRIBUTION WE DO BACKUPS OF (OF STATES)
- $d=\pi$ ⟶ ON-POLICY DISTRIBUTION, THAT OF FREQS OF STATES ENCOUNTERED

- UNCLEAR IF MINIMIZING RMSE COINCIDES WITH ULTIMATE GOAL OF FINDING BETTER POLICIES

Gradient Descent

- W WEIGHTS, $\hat{v}(s,w)$, W_T WEIGHTS AT STEP t. ASSUME A NEW EXAMPLE PER STEP $(S, V_\pi(S))$ TRUE VALUE UNDER π
  - $W_{T+1} = W_T + \alpha\left[V_\pi(S_T)-\hat{V}_\pi(S_T,W_T)\right]\nabla\hat{V}(S_T,W_T)$
- **NO TRUE VALUES** BECAUSE NOISE OR OTHER SHIT. USE ESTIMATE | BACKUP V_T INSTEAD, SAME FORM. STILL CONVERGES FOR DECREASING αS
- $V_T=G_T$ MC RETURN ⟶ CONVERGENCE OK
- $V_T=G_T^\lambda$ λ-RETURN IS NOT UNBIASED ESTIMATE, NO CONVERGENCE, STILL EFFECTIVE AND RELEVANT
- $$\begin{cases} W_{T+1}=W_T+\alpha\delta_T e_t \\ \delta_T = R_{T+1}+\gamma\hat{V}(S_T,W_T)-\hat{V}(S_T,W_T) \\ e_T=\lambda\gamma e_{T-1}+\nabla\hat{V}(S_T,W_T)\end{cases}$$ ⟵ **BWD TD(λ)** GRADIENT WITH e ELIGIBILITY TRACES
- USUALLY ANN + BACKPROP, OR LINEAR METHODS

LINEAR METHODS: $\begin{cases}\hat{V}(S,W)=W^TX \\ \nabla\hat{V}(S,W)=X(S)\end{cases}$ • $X(S)$ FEATURE VECTOR, STATE IS IDENTIFIED WITH FEATURES!!! **LINEAR REGRESSION!**
- CONVERGENCE IS GUARANTEED ~~BUT~~ BUT TO W_∞, NOT ACTUAL W^* (MIN POSSIBLE ERROR)
- FEATURE SELECTION IS KEY

ON-POLICY + LINEAR! $RMSE(W_\infty) \le \frac{1-\gamma\lambda}{1-\gamma}RMSE(W^*)$

- **COARSE CODING:** FEATURES WITH 'RECEPTIVE FIELDS' THAT OVERLAP. (UNCOMB?) ~~[illegible]~~ RESOLUTION/GENERALIZATION TRADEOFF ON SIZE OF RECEPTIVE FIELDS AND NO OF FEATURES. ⟶ **BANDWIDTH**
- **TILE CODING:** COARSE. WITH ONLY BINARY FEATURES WITH NON-OVERLAPPING R.F. TILINGS NOT NECESSARILY UNIFORM GRIDS
  COMPUTATION BECOMES VERY EFFICIENT BECAUSE WE CAN COUNT|SUM INSTEAD THAN MULTIPLY. ALSO HASHING.
- **RBF CODING:** GAUSSIAN KERNEL. SMOOTHER, DIFFERENTIABLE APPROXIMATIONS ARE PRODUCED. CAN ALSO LEARN μ, Σ FOR ~~[illegible]~~ GREAT JUSTICE.
- **KANERVA CODING:** DECOUPLE DIMENSIONALITY|COMPLEXITY OF STATE SPACE FROM THAT OF TARGET FUNCTION. FOR RESILIENCE WRT CURSE OF DIMENSIONALITY
  ⟶ USE PROTOTYPES (KERNELS!) AND EXPRESS STATES WRT CLOSENESS TO PROTOTYPES **EXAMPLE:** BINARY SPACE, HAMMING DISTANCE

Control with Function Approximation

- USUAL GPI APPROACH
- **ACTION-VALUE PREDICTION:** $\hat{q} \approx q_\pi$ $(S_T, A_T) \mapsto Q_T$: MC RETURN. SARSA RETURN. ETC... $W_{T+1}=W_T+\alpha\left[Q_T-\hat{q}(S_T,A_T,W_T)\right]\nabla\hat{q}(S_T,A_T,W_T)$
- **POLICY IMPROVEMENT, ACTION SELECTION:**
  - DISCRETE ACTION SET, NOT TOO LARGE ⟶ USUAL TECHNIQUES OK: FOREACH a, COMPUTE $\hat{q}$, AND FIND GREEDY ACTION
  - CONTINUOUS OR LARGE DISCRETE ACTION SPACES ⟶ NO CLEAR SOLUTION | ONGOING RESEARCH

  { **ON-POLICY** PI W/ SOFT APPROX, ε-GREEDY, SELECT WITH SAME POLICY
  OFF-POLICY PI W/ GREEDY POLICY, SELECTION W/ ARBITRARY POLICY
- **TRACES:** CAN USE ANY VARIANT. NO TRACE FOR STATES, BUT FOR WEIGHTS ⟶ TREAT FEATURES AS STATES AND DO TRACES ON THEM. OPTIONAL CLEARING OF TRACES OF NONSELECTED ACTIONS

Bootstrapping

- **BOOTSTRAPPING:** BETTER AT SOLVING PROBLEMS. NOT CLEAR WHY.
- **NONBOOTSTRAPPING:** BETTER AT MINIMIZING RMSE.